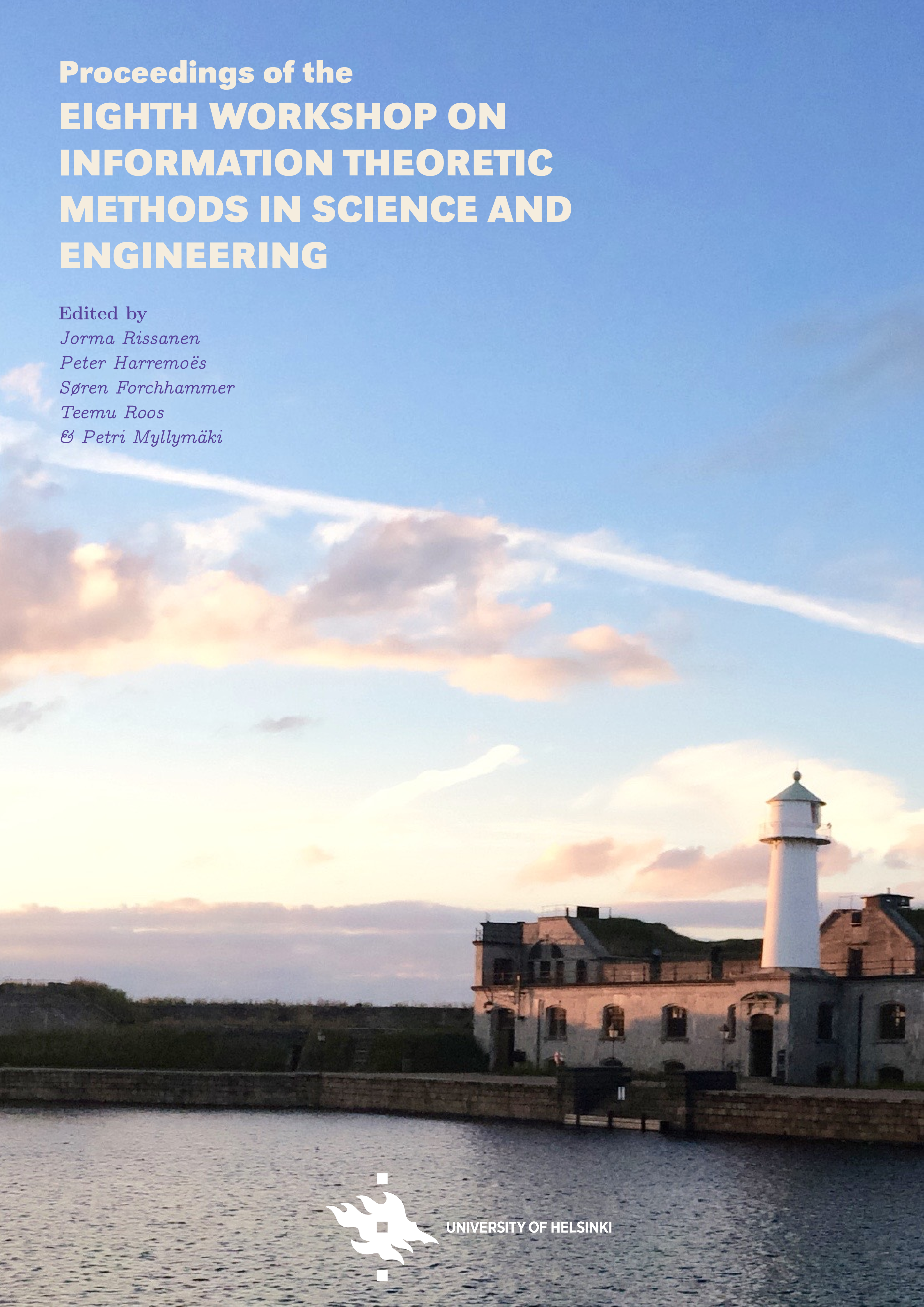
Proceedings of the

EIGHTH WORKSHOP ON INFORMATION THEORETIC METHODS IN SCIENCE AND ENGINEERING

Edited by
Jorma Rissanen
Peter Harremoës
Søren Forchhammer
Teemu Roos
& Petri Myllymäki



UNIVERSITY OF HELSINKI

University of Helsinki
Department of Computer Science
Series of Publications B
Report B-2015-1

PREFACE

The Eighth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2015) took place on June 24–26, 2015, in Copenhagen, Denmark. The workshop was organized jointly by the Department of Computer Science of the University of Helsinki, the Helsinki Institute for Information Technology HIIT, Niels Brock, Copenhagen Business College, and the Danish Technical University.

The WITMSE series started in 2008 and has continued annually at locations in Tampere (2008–2009), Helsinki (2011), Amsterdam (2012), Tokyo (2013), and Honolulu (2014). 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 workshop programme included seventeen invited talks, and two plenary talks that were given by Steffen Lauritzen and Gerhard Kramer. In addition there was a mini-tutorial on the Minimum Description Length principle by Teemu Roos, and an informal recent results session.

Outside the technical sessions the program included a guided tour and a welcoming reception at the historial Rundetårn tower and a banquet dinner at the Trekroner sea fortress. An optional excursion to Roskilde Viking Ship Museum and Cathedral was organized on Saturday, June 27.

We would like to thank all the participants to the workshop. We hope to see many of you again next year.

October 8, 2015
San Jose, Copenhagen, and Helsinki
Workshop Co-Chairs

Jorma Rissanen,
Peter Harremoës,
Søren Forchhammer,
Teemu Roos,
& Petri Myllymäki

Contents

UPPER BOUNDS ON THE CAPACITY OF FIBER CHANNELS

Gerhard Kramer

Technical University of Munich

ABSTRACT

The capacity of optical fiber channels seems difficult to compute or even bound. The best capacity lower bounds are based on numerical simulations using the split-step Fourier method. We review a recent capacity upper bound that applies two basic tools to this method: maximum entropy under a correlation constraint and Shannon's entropy power inequality (EPI). The main insight is that the non-linearity that is commonly used to model optical fiber propagation does not change the differential entropy of a signal. As a result, the spectral efficiency of fiber is at most $\log(1 + \mathrm{SNR})$, where SNR is the receiver signal-to-noise ratio. The results extend to other channels, including multi-mode fiber.

PROPER LOCAL SCORING RULES

Steffen Lauritzen

University of Copenhagen

ABSTRACT

A scoring rule is a loss function measuring the quality of a quoted probability distribution Q for a random variable X, in the light of the realized outcome x of X; it is proper if the expected score, under any distribution P for X, is minimized by quoting $Q = P$. Using the fact that any differentiable proper scoring rule on a finite sample space X is the gradient of a concave homogeneous function, we consider when such a rule can be local in the sense of depending only on the probabilities quoted for points in a nominated neighborhood of x. Under mild conditions, we characterize such a proper local scoring rule in terms of a collection of homogeneous functions on the cliques of an undirected graph on the space X. We also mention proper scoring rules for continuous distributions on the real line. Here we allow further dependence on a finite number m of derivatives of the density at the outcome, and describe a large class of such m-local proper scoring rules.

REGULAR HILBERG PROCESSES: NONEXISTENCE OF UNIVERSAL REDUNDANCY RATIOS

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warszawa, POLAND, ldebowsk@ipipan.waw.pl

ABSTRACT

A regular Hilberg process is a stationary process that satisfies both a power-law growth of topological entropy and a hyperlogarithmic growth of maximal repetition. Such processes may arise in statistical modeling of natural language. A puzzling property of ergodic regular Hilberg processes is that the length of the Lempel-Ziv code is orders of magnitude larger than the block entropy. This is possible since regular Hilberg processes have a vanishing entropy rate. In this paper, we provide some constructive example of regular Hilberg processes, which we call random hierarchical association (RHA) processes. We show that for those RHA processes, the expected length of any uniquely decodable code is orders of magnitude larger than the block entropy of the ergodic component of the RHA process. Our proposition complements the classical result by Shields concerning nonexistence of universal redundancy rates.

1. REGULAR HILBERG PROCESSES

Consider a measurable space of infinite sequences $(\mathbb{A}^\mathbb{N}, \mathcal{A}^\mathbb{N})$ from a finite alphabet $\mathbb{A} \subset \mathbb{N}$. The random symbols will be denoted as $\xi_k : \mathbb{A}^\mathbb{N} \ni (x_i)_{i\in\mathbb{N}} \mapsto x_k \in \mathbb{A}$, whereas blocks of symbols will be denoted as $x_{k:l} = (x_i)_{i=k}^{l}$. We define two functions of an individual sequence $\xi_{1:\infty}$. The first one is the maximal repetition

$$L(\xi_{1:k}) := \max\left\{m : x_{1:m} \text{ is repeated in } \xi_{1:k}\right\} \quad (1)$$

[1, 2, 3, 4, 5], whereas the dual one is the topological entropy

$$H_{top}(m|\xi_{1:\infty}) := \log \operatorname{card}\left\{x_{1:m} : x_{1:m} \sqsubset \xi_{1:\infty}\right\}, \quad (2)$$

where we write $a \sqsubset b$ when a is a subword of b.

The maximal repetition and the topological entropy are linked by the following simple proposition:

Theorem 1 ([6]) *If $H_{top}(m|\xi_{1:\infty}) < \log(k - m + 1)$ then $L(\xi_{1:k}) \ge m$.*

In particular, using the Big O notation, we have

$$H_{top}(m|\xi_{1:\infty}) = O\left(m^\beta\right) \Rightarrow L(\xi_{1:m}) = \Omega\left((\log m)^{1/\beta}\right),$$

$$L(\xi_{1:m}) = O\left((\log m)^{1/\beta}\right) \Rightarrow H_{top}(m|\xi_{1:\infty}) = \Omega\left(m^\beta\right).$$

There is a hypothesis, based on experimental measurements of maximal repetition, that for texts in natural language (such as English, French and German), scaling

$$L(\xi_{1:m}) = \Theta\left((\log m)^{1/\beta}\right), \quad (3)$$

$$H_{top}(m|\xi_{1:\infty}) = \Theta\left(m^\beta\right) \quad (4)$$

holds with $\beta \approx 0.5$ [6, 7]. Moreover, the lower bound for the maximal repetition and the upper bound for the topological entropy seem to be text-independent.

The goal of the present paper is to investigate some abstract stationary processes that satisfy conditions (3) and (4) almost surely. We hope that our examples may inspire some progress in statistical modeling of natural language, as explained in the final Section 3. Throughout this paper we identify stationary processes with their distributions (stationary measures) and we use terms "measure" and "process" interchangeably.

Definition 1 (a variation of a definition in [7]) *A stationary measure μ on the measurable space of infinite sequences $(\mathbb{A}^\mathbb{N}, \mathcal{A}^\mathbb{N})$ is called a* regular Hilberg process with an exponent $\beta \in (0, 1)$ *if it satisfies conditions (3)–(4) μ-almost surely, where the lower bound for the maximal repetition and the upper bound for the topological entropy are uniform in $\xi_{1:\infty}$.*

We call these processes "regular Hilberg processes" to commemorate the research by Hilberg [8], who was the first one to notice the power-law scaling of the entropy of natural language.

It can be seen easily that so defined regular Hilberg processes have a vanishing entropy rate. To demonstrate this result, let us introduce some notation. The expectation with respect to a stationary measure μ is denoted as $\mathbf{E}_\mu$. We also use shorthand $\mu(x_{1:m}) = \mu(\xi_{1:m} = x_{1:m})$. The block entropy of measure μ is

$$H_\mu(m) := \mathbf{E}_\mu\left[-\log \mu(\xi_{1:m})\right], \quad (5)$$

and the entropy rate of μ is the limit

$$h_\mu := \inf_{m\in\mathbb{N}} \frac{H_\mu(m)}{m} = \lim_{m\to\infty} \frac{H_\mu(m)}{m}. \quad (6)$$

Now will show the mentioned result.

Theorem 2 $h_\mu = 0$ *for a regular Hilberg process μ.*

Proof: The argument involves the random ergodic measure $F = \mu(\cdot|\mathcal{I})$, where $\mathcal{I}$ is the shift-invariant algebra [9, 10]. By the ergodic theorem [9], we have μ-almost surely

$$H_F(m) \le H_{top}(m|\xi_{1:\infty}), \tag{7}$$

so $h_F = 0$, whereas as shown in [10] we have

$$h_\mu = \mathbf{E}_\mu \, h_F, \tag{8}$$

from which $h_\mu = 0$ follows. □

The vanishing entropy rate is equivalent to the process being asymptotically deterministic and infinitely compressible in the following sense. Firstly, the process μ will be called asymptotically deterministic when each symbol ξ_i is μ-almost surely a function of the infinite past $\xi_{-\infty:i-1}$, cf. [11]. Secondly, the process μ will be called infinitely compressible when for every universal code C the compression rate $|C(\xi_{1:m})|/m$ tends to zero μ-almost surely for the block length m tending to infinity.

Let us note, however, that processes with a vanishing entropy rate may be practically very difficult to predict or to compress if we do not know their exact distribution. Ergodic regular Hilberg processes fall exactly under this case. In fact, these processes have a notable counterintuitive compression property. Namely, the length of the Lempel-Ziv code for a block $\xi_{1:m}$, which is a universal code [12], is orders of magnitude larger than the block entropy $H_\mu(m)$. Precisely, we have:

Theorem 3 *Let $|C(\xi_{1:m})|$ be the length of the Lempel-Ziv code for a block $\xi_{1:m}$. For an ergodic regular Hilberg process μ with exponent β, μ-almost surely*

$$|C(\xi_{1:m})|/H_\mu(m) = \Omega\left(\frac{m^{1-\beta}}{(\log m)^{1/\beta-1}}\right). \tag{9}$$

Proof: By ergodicity, we have $\mu = F$. Thus, by (7) and (4), we obtain

$$H_\mu(m) = H_F(m) \le H_{top}(m|\xi_{1:\infty}) = O\left(m^\beta\right). \tag{10}$$

On the other hand, the length of the Lempel-Ziv code $|C(\xi_{1:m})|$ for a block $\xi_{1:m}$, by (3), μ-almost surely satisfies

$$\begin{aligned} |C(\xi_{1:m})| &\ge \frac{m}{L(\xi_{1:m})+1} \log \frac{m}{L(\xi_{1:m})+1} \\ &= \Omega\left(\frac{m}{(\log m)^{1/\beta-1}}\right). \end{aligned} \tag{11}$$

The first inequality in (11) stems from a simple observation in [7] that the length of the Lempel-Ziv code is greater than $V \log V$, where V is the number of Lempel-Ziv phrases, whereas the Lempel-Ziv phrases may not be longer than the maximal repetition plus 1. □

In view of Theorem 3, we cannot estimate the block entropy of an ergodic regular Hilberg process by the length of the Lempel-Ziv code! We will show that a similar statement holds in expectation for an arbitrary uniquely decodable code and some specific regular Hilberg processes. These processes, called RHA process, will be introduced in Section 2. As we will see later in Section 3, our impossibility result for the RHA processes strengthens the classical result by Shields concerning nonexistence of universal redundancy rates [13].

2. THE RHA PROCESSES

In this section we will construct some examples of regular Hilberg processes. The processes will be called random hierarchical association (RHA) processes. The RHA processes are parameterized by certain free parameters which we will call perplexities (a name borrowed from computational linguistics). Approximately, perplexity k_n is the number of distinct blocks of length 2^n that appear in the process realization. It occurs that controlling perplexities, we can control the value of block entropy and force the entropy rate to be zero. It occurs as well that we can control the value of the topological entropy and the maximal repetition to obtain regular Hilberg processes.

The RHA processes are formed in two steps. First, we sample recursively random pools of k_n distinct blocks of length 2^n, which are formed by concatenation of randomly selected k_n pairs chosen from k_{n-1} distinct blocks of length 2^{n-1} (the recursion stops at blocks of length 1, which are fixed symbols). Second, we obtain an infinite sequence of random symbols by concatenating blocks of lengths 2^0, 2^1, 2^2, ... randomly chosen from the respective pools. As a result there cannot be more that k_n^2 distinct blocks of length 2^n that appear the final process realization. The selection of these blocks is however random and we do not know them a priori.

Now we write down the construction using symbols.

Step 1: Formally, let perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ be some sequence of strictly positive natural numbers that satisfy

$$k_{n-1} \le k_n \le k_{n-1}^2. \tag{12}$$

Next, for each $n \in \mathbb{N}$, let $(L_{nj}, R_{nj})_{j\in\{1,...,k_n\}}$ be an independent random combination of k_n pairs of numbers from the set $\{1, ..., k_{n-1}\}$ drawn without repetition. That is, we assume that each pair (L_{nj}, R_{nj}) is different, the elements of pairs may be identical ($L_{nj} = R_{nj}$), and the sequence $(L_{nj}, R_{nj})_{j\in\{1,...,k_n\}}$ is sorted lexicographically. Formally, we assume that random variables L_{nj} and R_{nj} are supported on some probability space $(\Omega, \mathcal{J}, P)$ and have the uniform distribution

$$\begin{aligned} P((L_{n1}, R_{n1}, ..., L_{nk_n}, R_{nk_n}) = (l_{n1}, r_{n1}, ..., l_{nk_n}, r_{nk_n})) \\ = \binom{k_{n-1}^2}{k_n}^{-1}. \end{aligned} \tag{13}$$

Subsequently we define random variables

$$Y_j^0 = j, \qquad j \in \{1, ..., k_0\}, \tag{14}$$

$$Y_j^n = Y_{L_{nj}}^{n-1} \times Y_{R_{nj}}^{n-1}, \quad j \in \{1, ..., k_n\}, n \in \mathbb{N}, \tag{15}$$

where $a \times b$ denotes concatenation. Hence Y_j^n are k_n distinct blocks of 2^n natural numbers, selected by some sort of random hierarchical concatenation.

Step 2: Variables Y_j^n will be the building blocks of yet another process. Let $(C_n)_{n\in\{0\}\cup\mathbb{N}}$ be independent random variables, independent from $(L_{nj}, R_{nj})_{n\in\mathbb{N}, j\in\{1,...,k_n\}}$, with uniform distribution

$$P(C_n = j) = 1/k_n, \qquad j \in \{1, ..., k_n\}\,. \tag{16}$$

Definition 2 *The* random hierarchical association (RHA) process $\mathcal{X}$ with perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ *is defined as*

$$\mathcal{X} = Y_{C_0}^0 \times Y_{C_1}^1 \times Y_{C_2}^2 \times \dots\,. \tag{17}$$

This completes the construction of the RHA processes but it is not the end of our story yet.

It is convenient to define a few more random variables for the RHA process. First, sequence $\mathcal{X}$ will be parsed into a sequence of numbers X_j, where

$$\mathcal{X} = X_1 \times X_2 \times X_3 \times ..., \tag{18}$$

and, second, we denote blocks starting at any position as

$$X_{k:l} = X_k \times X_{k+1} \times ... \times X_l. \tag{19}$$

The RHA processes defined in Definition 2 are not stationary but they possess a stationary mean, which is a condition related to asymptotic mean stationarity. Let us introduce shift operation $T : \mathbb{A}^{\mathbb{N}} \ni (x_i)_{i\in\mathbb{N}} \mapsto (x_{i+1})_{i\in\mathbb{N}} \in \mathbb{A}^{\mathbb{N}}$. We recall this definition:

Definition 3 *A measure ν on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ is called* asymptotically mean stationary (AMS) *if limits*

$$\mu(A) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nu(T^{-i}A) \tag{20}$$

exist for every event $A \in \mathcal{A}^{\mathbb{N}}$ [14].

For an AMS measure ν, function μ is a stationary measure on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$, called the stationary mean of ν. Moreover, measures μ and ν are equal on the shift invariant algebra $\mathcal{I} = \left\{A \in \mathcal{A}^{\mathbb{N}} : T^{-1}A = A\right\}$, i.e., $\mu(A) = \nu(A)$ for all $A \in \mathcal{I}$.

Now, let $\mathbb{A}^+ = \bigcup_{n\in\mathbb{N}} \mathbb{A}^n$. There is a related relaxed condition of asymptotic mean stationarity:

Definition 4 *A measure ν on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ is called* asymptotically mean stationary with respect to blocks (AMSB) *if limits*

$$\mu(x_{1:m}) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nu(\xi_{i:i+m-1} = x_{1:m}) \tag{21}$$

exist for every block $x_{1:m} \in \mathbb{A}^+$.

For an AMSB measure ν over a finite alphabet $\mathbb{A}$, function μ, extended via $\mu(\xi_{1:m} = x_{1:m}) := \mu(x_{1:m})$, is also a stationary measure on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$. We shall continue to call this μ a stationary mean of ν. However, an AMSB measure need not be AMS, cf. [15, Example 6.3]. In particular, for an AMSB measure ν we need not have $\mu(A) = \nu(A)$ for shift invariant events $A \in \mathcal{I}$.

It turns out that the RHA processes are AMSB.

Theorem 4 *The RHA processes are AMSB. In particular, for $m \le 2^n$ and $k \in \mathbb{N}$, the stationary mean is*

$$\mu(x_{1:m}) = \frac{1}{2^n} \sum_{j=0}^{2^n-1} P(X_{k2^n+j:k2^n+j+m-1} = x_{1:m}). \tag{22}$$

We suppose that the RHA processes are also AMS but we could not prove it so far. However, we have been able to show that certain RHA processes are regular Hilberg processes:

Theorem 5 *For perplexities*

$$k_n = \left\lfloor \exp\left(2^{\beta n}\right)\right\rfloor, \tag{23}$$

where $\beta \in (0, 1)$, the stationary mean μ of the RHA process constitutes a regular Hilberg process with the exponent β, whereas its block entropy is sandwiched by

$$\frac{C_1 m}{(\log m)^\alpha} \le H_\mu(m) \le C_2 m \left(\frac{\log\log m}{\log m}\right)^\alpha, \tag{24}$$

where $\alpha = 1/\beta - 1$.

The measure μ for perplexities (23) is nonergodic and the entropy of the shift invariant algebra $H_\mu(\mathcal{I})$, as defined in [11], is infinite. If we need an ergodic process, however, we may consider the random ergodic measure $F = \mu(\cdot|\mathcal{I})$. The measure F is μ-almost surely an ergodic regular Hilberg process with the exponent β.

3. CONCLUDING REMARKS

Having Theorem 5, we may return to the question of nonexistence of universal redundancy rates. Shields [13] showed that for any uniquely decodable code C and any sublinear function $\rho(m) = o(m)$ there exists such an ergodic source F that

$$\limsup_{m\to\infty} [\mathbf{E}_F \left|C(\xi_{1:m})\right| - H_F(m) - \rho(m)] > 0. \tag{25}$$

Shields' result concerns nonexistence of a universal sublinear bound for the difference $\mathbf{E}_F \left|C(\xi_{1:m})\right| - H_F(m)$. Some way of strengthening this result is to investigate ratio $\mathbf{E}_F \left|C(\xi_{1:m})\right| / H_F(m)$. Although this ratio is asymptotically equal to 1 for universal codes and processes with a positive entropy rate $h_F > 0$, Shields' result does not predict how the ratio behaves for processes with a vanishing entropy rate $h_F = 0$.

Now we will show that there may be no universal sublinear bound for the ratio $\mathbf{E}_F \left|C(\xi_{1:m})\right| / H_F(m)$, either. Precisely, we obtain a weaker result in expectation:

Theorem 6 *Let $|C(\xi_{1:m})|$ be the length of an arbitrary uniquely decodable code for a block $\xi_{1:m}$. For the stationary mean μ of the RHA process with perplexities (23) and its random ergodic measure $F = \mu(\cdot|\mathcal{I})$, we have*

$$\mathbf{E}_\mu \frac{\mathbf{E}_F |C(\xi_{1:m})|}{H_F(m)} = \Omega\left(\frac{m^{1-\beta}}{(\log m)^{1/\beta-1}}\right), \qquad (26)$$

Ratio (26) can be larger than any function $o(m^{1-\epsilon})$.

Proof: The claim follows by (7), (4), (24), and the source coding inequality

$$\mathbf{E}_\mu \mathbf{E}_F |C(\xi_{1:m})| = \mathbf{E}_\mu |C(\xi_{1:m})| \geq H_\mu(m). \qquad (27)$$

□

We hope that our example of the RHA processes may also stimulate some progress in statistical modeling of natural language. Let us recall that Hilberg [8] replotted Shannon's seminal estimates of block entropy of printed English [16] in a doubly logarithmic scale and observed relationship

$$H_\mu(m) = \Theta(m^\beta), \qquad (28)$$

which implies vanishing entropy rate $h_\mu = 0$. So far we have not been aware of any explicit construction of a stationary process with a similar asymptotics of block entropy. In [17, 18], some stationary processes were constructed which satisfy a relaxed condition

$$2H_\mu(m) - H_\mu(2m) = \Theta\left(m^\beta\right) \qquad (29)$$

with an entropy rate $h_\mu > 0$. In contrast, in this paper, we have introduced the class of RHA processes which satisfy the regular Hilberg conditions (3)–(4) and therefore they obey $h_\mu = 0$. Possibly, the ergodic components of these processes satisfy also condition (28).

Seen from a larger perspective, we have shown that processes satisfying regular Hilberg conditions (3) and (4) arise in a quite simple setting of random sampling of texts from a restricted random hierarchical pool. Such scheme of sampling may arise in the course of human cultural evolution, since humans tend to copy existing texts, phrases, or words at least as much as to create new instances of them. The question remains how much randomness there is in the process of cultural evolution. Is it more or less than in the RHA processes? The point of view suggested by the mainstream information theory is that the entropy rate of natural language is strictly positive. In contrast, the analysis of empirical data by [8, 6] suggests that natural language may be a regular Hilberg process—with a vanishing entropy rate. Further research is required to determine which of these two hypotheses is true.

4. REFERENCES

[1] A. de Luca, "On the combinatorics of finite words," *Theor. Comput. Sci.*, vol. 218, pp. 13–39, 1999.

[2] P. C. Shields, "String matching: The ergodic case," *Ann. Probab.*, vol. 20, pp. 1199–1203, 1992.

[3] ——, "String matching bounds via coding," *Ann. Probab.*, vol. 25, pp. 329–336, 1997.

[4] R. Kolpakov and G. Kucherov, "Finding maximal repetitions in a word in linear time," in *40th Annual Symposium on Foundations of Computer Science, 1999*, 1999, pp. 596–604.

[5] ——, "On maximal repetitions in words," *J. Discr. Algor.*, vol. 1, pp. 159–186, 1999.

[6] Ł. Dębowski, "Maximal repetitions in written texts: Finite energy hypothesis vs. strong Hilberg conjecture," 2014, http://www.ipipan.waw.pl/ ldebowsk/.

[7] ——, "Hilberg's conjecture — a challenge for machine learning," 2015, http://www.ipipan.waw.pl/ ldebowsk/.

[8] W. Hilberg, "Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?" *Frequenz*, vol. 44, pp. 243–248, 1990.

[9] O. Kallenberg, *Foundations of Modern Probability*. Springer, 1997.

[10] R. M. Gray and L. D. Davisson, "The ergodic decomposition of stationary discrete random processses," *IEEE Trans. Inform. Theory*, vol. 20, pp. 625–636, 1974.

[11] Ł. Dębowski, "A general definition of conditional information and its application to ergodic decomposition," *Statist. Probab. Lett.*, vol. 79, pp. 1260–1268, 2009.

[12] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. 23, pp. 337–343, 1977.

[13] P. C. Shields, "Universal redundancy rates don't exist," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 520–524, 1993.

[14] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, pp. 962–973, 1980.

[15] Ł. Dębowski, "Variable-length coding of two-sided asymptotically mean stationary measures," *J. Theor. Probab.*, vol. 23, pp. 237–256, 2010.

[16] C. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1951.

[17] Ł. Dębowski, "On the vocabulary of grammar-based codes and the logical consistency of texts," *IEEE Trans. Inform. Theory*, vol. 57, pp. 4589–4599, 2011.

[18] ——, "Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks," *IEEE Trans. Inform. Theory*, vol. 58, pp. 3392–3401, 2012.

DISTRIBUTED SOURCE CODING OF VIDEO

Søren Forchhammer[1] *and Huynh Van Luong*[2]

[1]DTU Fotonik, Technical University of Denmark
2800 Kgs. Lyngby, sofo@fotonik.dtu.dk
[2]Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg
91058 Erlangen, huynh.luong@fau.de

ABSTRACT

A foundation for distributed source coding was established in the classic papers of Slepian-Wolf (SW) [1] and Wyner-Ziv (WZ) [2]. This has provided a starting point for work on Distributed Video Coding (DVC), which exploits the source statistics at the decoder side offering shifting processing steps, conventionally performed at the video encoder side, to the decoder side. Emerging applications such as wireless visual sensor networks and wireless video surveillance all require lightweight video encoding with high coding efficiency and error-resilience. The video data of DVC schemes differ from the assumptions of SW and WZ distributed coding, e.g. by being correlated in time and non-stationary. Improving the efficiency of DVC coding is challenging. This paper presents some selected techniques to address the DVC challenges. Focus is put on pin-pointing how the decoder steps are modified to provide adaptive decoding in distributed coding.

1. INTRODUCTION

Conventional video coding employs temporal prediction of frames to be coded. The apparent motion is represented by displacement vectors of blocks from previously coded data. This provides efficient coding, but also puts a heavy processing load on the encoder. In DVC an important issue is to use distributed techniques to encode the video frames individually, but utilize the temporal correlation on the decoder side for efficient video coding.

The Slepian-Wolf and Wyner-Ziv theorems addresses distributed coding in a set-up with two sequences, X and Y, each independent and identically distributed (iid), but jointly statistically dependent. The Slepian-Wolf theorem states that X can be independently encoded but decoded given the side-information (SI) Y at the same rate, $H(X|Y)$, as an optimal encoder having access to Y, under certain conditions. The Wyner-Ziv theorem extends this to the lossy case in a rate-distortion setting again under certain conditions.

We shall take this mind set but investigate it for real data in DVC were the assumptions of iid sequences do not hold. We shall use the term Side Information Generation to the processing of decoded data at a given point to provide estimates of the data, X, to be decoded. A prominent approach to DVC is Transform domain Wyner-Ziv (TDWZ) video coding [3], where a feedback channel is employed to let the decoder control the rate by requests. In the basic setting (called GOP2) every other frame (called Key Frames) is coded using intra-coding and the frames in between are coded using distributed techniques and decoded using the two surrounding frames as side information and called WZ frames. The feedback introduced serves to adapt the bit-rate as the required number of bits is varying and not known.

The TDWZ DVC coding architecture employs a DCT like transform on 4 x 4 blocks. While providing some decorrelation, there is still significant correlation in the transformed data. The coding efficiency has been improved considerably by a number of techniques.

In Sec. 2, we present a basic TDWZ DVC architecture as in [3] and improved in the DISCOVER codec [4]. In Sec. 3, improvements by making the decoder adaptive based on reestimations are presented. First to capture crossband correlations [5] and extended in the side information and noise learning (SING) codec [6] introducing an optical flow technique for motion estimation to compensate the weaknesses of the block based SI generation and in the motion and reconstruction reestimation (MORE) [7] codec, where the updated information is used to iteratively reestimate the motion and reconstruction. Finally, an adaptive mode decision (AMD) is investigated to take advantage of skip and intra mode in DVC by deciding the coding modes based on the quality of key frames and rate of WZ frames. Benchmark results of the resulting MORE-AMD [8] and the other techniques are briefly presented in Sec. 4. In Sec. 5, the SW coding based on rate-adaptive error-correcting techniques is revisited [9].

2. DISTRIBUTED VIDEO CODING

The architecture of a TDWZ video codec [4] is depicted in Fig. 1. In this codec, the sequence of frames is split into key frames and Wyner-Ziv (WZ) frames. Key frames are intra coded using conventional video coding techniques such as H.264/AVC intra coding. The Wyner-Ziv frames are transformed (4×4 DCT), quantized and decomposed into bitplanes.

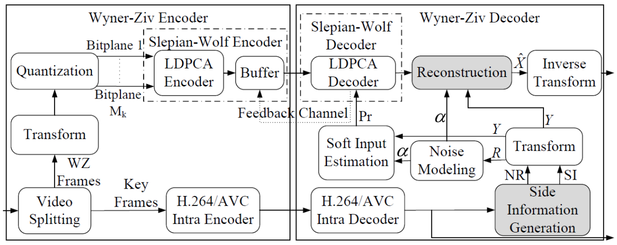


Figure 1. The architecture of TDWZ video codec.

Each bitplane is fed to a rate-compatible LDPC Accumulate (LDPCA) encoder [10] from most significant bitplane to least significant bitplane. The corresponding error correcting information is stored in a buffer and requested by the decoder through a feedback channel. The WZ frame is predicted at the decoder side by using decoded frames as references. The predicted frame, called the SI frame, is an estimate of the original WZ frame. Given the available SI, soft-input information (conditional probabilities *Pr* for each bit) within each bitplane is estimated using a noise model. Thereafter the LDPCA decoder starts to decode the bitplanes selected by the quantizer, ordered from most to least significant bitplane, to correct bit errors. The decoder requests bits from the buffer until the bitplane is decoded. Thereafter CRC check bits are sent for confirmation. After all bitplanes are successfully decoded, the WZ frame can be decoded through combined de-quantization and reconstruction followed by an inverse transform.

3. ADAPTIVE DVC USING REESTIMATIONS

Adaptive coding in distributed source coding is enabled by the feedback based request of parity bits for rate-adaptation. After a successful decoding of a code block, the decoder can update the side information and thereby the soft-information for decoding the next block. Thus the side-information used in the decoding may generally be seen as a mapping of the causal data and the frame level approach presented in Sec. 2 may be extended to (sub)band level and bit-plane level, where a code block in this section is given by the information required to decode a bit-plane of one coefficient subband. This adaptation using decoded blocks may also be applied in the motion estimation step using partially decoded frames. We shall first focus on how the decoder may introduce adaptive coding, while the encoder remains the same.

3.1. Crossband correlations in DVC

The Crossband DVC scheme [5] enhances the DISCOVER architecture using previously decoded subbands in the noise modeling for the next subbband. Specifically after decoding a subband, a classification is performed. When modeling a new subband one or two previous subbands are used to predict the classification. This subband level adaptive processing is combined with a bit-plane level updating of estimates for each coefficient.

3.2. Multi-hypotheses decoding using optical flow

In the SING codec [6] multi-hypothesis decoding was used integrate a number of decoder-side adaptive techniques.

3.2.1. Dense motion fields using optical flow

The motion field in the side information generation is backward adaptive in the DVC scheme, thus the motion (vectors) are not coded. This allows using a dense motion field. In the SING codec [6], global optical flow (OF) was used to calculate dense motion fields to supplement a more conventional overlapping block motion compensation (OBMC).

3.2.2. Multi-hypotheses decoding

The rate-adaptive LDPCA decoder may be fed with multiple sets of soft-information, and terminate and ‘selecting’ the set first to decode (subject to a CRC). This provides a generic approach to decoder side adaptation in distributed coding. SI based on both optical flow and block-based OBMC can e.g. be combined to provide multiple hypotheses [6].

3.2.3. Adaptive noise modeling

Different noise modeling may also be adaptively selected using the multi-hypothesis approach. In SING, techniques based on clustering of DCT blocks, calculating feature vectors and updating and refining these was applied. Distributions of the residuals from previous frames were also used and the number of clusters adapted.

3.3. Re-estimation of motion

A challenge in DVC, incl. the scheme presented so far is the prediction of the motion at the decoder side for the WZ frame, which is not available as opposed to conventional video coding, especially for sequences with high or complex motion. As the WZ frame is being decoded, also the motion may be reestimated. Two instances of this were introduced in the MORE codec [7]. The partially reconstructed frames were used to reestimate motion for both the optical-flow (after each band) and the block-based (OBMC after each bit-plane) techniques. This can improve the prediction of the values. To also improve the estimate of the distribution, the residue may also be motion compensated using an updated estimate of the residue of the previous WZ frame and the current motion estimation to calculate a motion compensated residue. These techniques were integrated in the SING [6] codec (Figure 2).

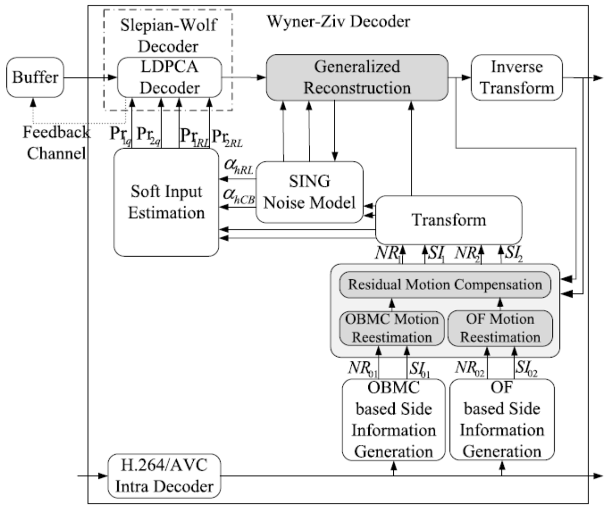


Figure 2. TDWZ decoder with the motion and reconstruction reestimation (MORE).

The combination of initial side information and reestimated side information based on reestimated motion in MORE is based on an adaptive selection process, trying to estimate influence on rate and distortion. The rate is estimated by the *ideal code length* (ICL), which after decoding a bit-plane may be calculated by summing minus log of the conditional probability assigned by the soft-input to each bit. A Lagrangian based rate-distortion cost function is used to adaptively selecting one block-based and one OF based estimate, SI and residual, NR, for the further processing to form the multiple inputs to the LDPCA decoder (Figure 2).

3.4. Decoder side driven adaptive mode decision

In video coding, skip coding and intra coding are used as additional modes. When advantageous in an operational rate-distortion sense [11] applying these modes improves performance. Introducing this adaptive mode decision (AMD) in DVC does require a change of the encoder to switch between modes. The decision can be encoder based introducing extra encoder processing steps and/or fed back from the decoder. Initial experimental results of integrating AMD in the MORE codec were reported in [8].

4. NUMERICAL RESULTS

The methods presented in Secs. 2 and 3 were tested on the four standard test sequences: *Foreman*, *Hall*, *Soccer* and *Coast* for a number of different bit-rates. Operational rate-distortion performance were calculated, expressing the quality by PSNR values [7,8]. The weighted average improvements (measured by Bjöntegaard differences [12]) over DISCOVER [4] are given in Tables 1 and 2. The resulting MORE [7] codec achieved an average improvement in PSNR is 2.5 dB on the WZ frames (for GOP2) and gained 1.2 dB measured over all frames. The performance of Crossband, SING, and MORE(AMD) are also given for comparison.

Table 1. Bjøntegaard PSNR improvement (dB) over DISCOVER for WZ frames

Codec	Crossband	SING	MORE	MORE(AMD)
Foreman	0.65	1.52	**3.00**	2.93
Hall	0.39	0.99	1.42	**1.95**
Soccer	1.33	2.70	**4.19**	4.182
Coast	0.36	0.41	0.65	**0.85**
Average	0.64	1.49	2.47	**2.58**

Table 2. Bjøntegaard PSNR improvement (dB) over DISCOVER for all frames

Codec	Crossband	SING	MORE	MORE(AMD)
Foreman	0.33	0.75	**1.43**	1.41
Hall	0.19	0.40	0.58	**0.61**
Soccer	0.73	1.51	**2.26**	2.23
Coast	0.19	0.22	0.27	**0.34**
Average	0.33	0.76	1.22	**1.22**

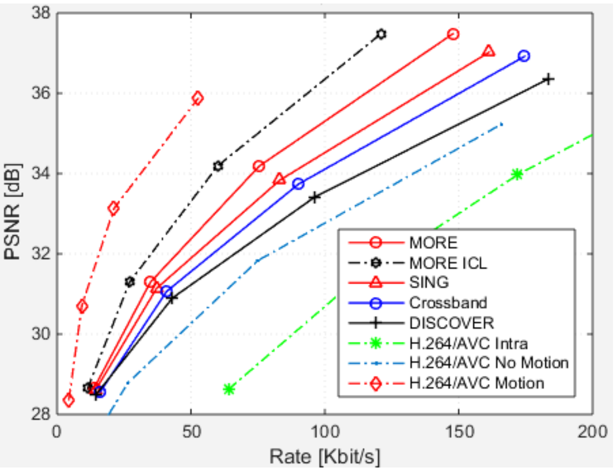


Figure 3. PSNR vs. rate for selected codecs for WZ frames (QCIF, 15Hz) for *Coast*.

The RD performance of the MORE, SING, and Crossband as well as H.264/AVC coding is also depicted in Figure 3 for the *Coast* sequence for WZ frames. In addition, the MORE ICL, which is obtained by replacing LDPCA coding with a calculation of the ideal code length (ICL) over all the decoded bitplanes, is also given. We calculate the ideal code length, ICL [6][7], at the decoder side based on the soft-input values used when decoding the information bits.

5. SW CODING REVISITED

Comparing the ideal code lengths (ICL) and the actual code lengths in DVC provides an evaluation of the loss in distributing the coding applying error-correcting coding, instead of e.g. arithmetic coding based on the conditional probabilities, see Figure 2. Investigations of the results obtained using the LDPCA code, widely used in DVC, show that especially for low conditional probabilities, there is a relatively high loss, which may be an issue in DSC in general. As an alternative rate-adaptive BCH [9] for coding with feed-back was studied. The feedback provides the capability to adapt to unknown statistics and also to reduce the coding loss FEC codes endures when small code blocks are used. Linear block codes with extensible parity matrix, H, may readily be used for rate-adaptive coding, extending the matrix and sending new syndromes when more information is requested.

In [9], a rate-adaptive BCH (RA-BCH) code was introduced and analyzed for the case of bounded distance decoding and assuming iid error probability with known error probability between the side information, Y, and the information data, X. The scheme also involved using syndromes for checking and making the number of syndromes used to confirm a decoding adaptive to the number of syndromes received thus far. For error probability, p = 0.01, $H(X|Y) \sim 0.08$ (which also gives the average ICL). Based on simulations with this set-up at a bit-error-rate of 10^{-5}, the average code lengths where ~ 0.10 for RA-BCH for length 1023 and ~ 0.144 for LDPCA of length 1584. For fixed rate coding, a bound of ~ 0.146 was calculated and for both fixed rate LDPCA and BCH of the lengths considered the rate would be above 0.2, thus showing the clear benefit of using feed-back for these short code block lengths. In these comparisons BCH was clearly better that LDPCA. The challenge towards using RA-BCH in DVC is to generalize to soft-input decoding.

6. CONCLUDING REMARKS

We have given a brief overview of elements of a state-of-the-art DVC scheme with focus on aspects which may be of general interest when applying DSC to real data, especially video data. This included ways to make a DSC scheme with feed-back adaptive on the decoder side. Also it was pointed out that as DVC and DSC schemes improve performance, the loss in current error-correcting techniques applied become an issue towards achieving distributed coding without out performance loss as suggested by the classic Slepian-Wolf and Wyner-Ziv papers.

7. REFERENCES

[1] D. Slepian and J. K. Wolf "Noiseless coding of correlated information sources", *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471-480, 1973.

[2] A. D. Wyner and J. Ziv "The rate-distortion function for source coding with side information at the decoder", *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1-10, 1976.

[3] B. Girod, A. M. Aaron, S. Rane and D. Rebollo-Monedero "Distributed video coding", *Proc. IEEE*, vol. 93, no. 1, pp. 71-83, 2005.

[4] (2007, Dec.). Discover Project [Online]. Available: http://www.discoverdvc.org/

[5] X. Huang and S. Forchhammer "Cross-band noise model refinement for transform domain Wyner-Ziv video coding", *Image Commun.*, vol. 27, no. 1, pp. 16-30, 2012.

[6] H. V. Luong, L. L. Rakêt, X. Huang, and S. Forchhammer, “Side Information and Noise Learning for Distributed Video Coding using Optical Flow and Clustering,” *IEEE Trans. Image Proc.*, vol. 21, no. 12, pp. 4782-4796, 2012.

[7] H. V. Luong, L. L. Rakêt, and S. Forchhammer, "Re-estimation of Motion and Reconstruction for Distributed Video Coding," IEEE *Trans. Image Proc.*, vol. 23, no. 7, pp. 2804-2819, 2014.

[8] H. V. Luong, S. Forchhammer, J. Slowack, J. De Cock, and R. Van de Walle, "Adaptive Mode Decision with Residual Motion Compensation for Distributed Video Coding," *APSIPA Transactions on Signal and Information Processing*, vol. 4, no. e1, pp. 1-11, 2015. (DOI: http://dx.doi.org/10.1017/ATSIP.2014.21).

[9] M. Salmistraro, K. Larsen and S. Forchhammer, “Rate-adaptive BCH codes for distributed source coding,” *EURASIP Signal Process. J.*, 2013.

[10] D. Varodayan, A. Aaron, and B. Girod, “Rate-adaptive codes for distributed source coding,” *EURASIP Signal Process. J.*, vol. 86, no. 11, pp. 3123–3130, 2006.

[11] J. Slowack, S. Mys, J. Skorupa, N. Deligiannis, P. Lambert, A. Munteanu, and R. Van de Walle, “Rate-distortion driven decoder-side bitplane mode decision for distributed video coding,” *Signal Process.: Image Commun.*, 25 (9) (2010), 660–673.

[12] G. Bjøntegaard, “Calculation of average PSNR differences between RD curves,” in *Proc. VCEG Contrib.*, Apr. 2001.

PROPER SCORING AND SUFFICIENCY

Peter Harremoës

Niels Brock, Copenhagen Business College,
Copenhagen, DENMARK, harremoes@ieee.org

ABSTRACT

Logarithmic score and information divergence appear in both information theory, statistics, statistical mechanics, and portfolio theory. We demonstrate that all these topics involve some kind of optimization that leads directly to the use of Bregman divergences. If a sufficiency condition is also fulfilled the Bregman divergence must be proportional to information divergence. The sufficiency condition has quite different consequences in the different areas of application, and often it is not fulfilled. Therefore the sufficiency condition can be used to explain when results from one area can be transferred directly from one area to another and when one will experience differences.

1. INTRODUCTION

The use of scoring rules has a long history in statistics. An early contribution was the idea of minimizing the sum of square deviations that dates back to Gauss and works perfectly for Gaussian distributions. In the 1920's Ramsay and de Finetti proved versions of the Dutch book theorem where determination of probability distributions were considered as dual problems to maximizing a payoff function. Later it was proved that any consistent inference corresponds to optimizing with respect to some payoff function. A more systematic study of scoring rules was given by McCarthy [1] and has recently been studied by Dawid, Lauritzen and Parry [2] where the notion of a local scoring rule has been extended. The basic result is that the only strictly local proper scoring rule is logarithmic score.

Thermodynamics is the study of concepts like heat, temperature and energy. A major objective is to extract as much energy from a system as possible. Concepts like entropy and free energy play a significant role. The idea in statistical mechanics is to view the macroscopic behavior of a thermodynamic system as a statistical consequence of the interaction between a lot of microscopic components where the interacting between the components are governed by very simple laws. Here the central limit theorem and large deviation theory play a major role. One of the main achievements is the formula for entropy as a logarithm of a probability.

One of the main purposes of information theory is to compress data so that data can be recovered exactly or approximately. One of the most important quantities was called entropy because it is calculated according to a formula that mimics the calculation of entropy in statistical mechanics. Another key concept in information theory is information divergence (KL-divergence) that was introduced by Kullback and Leibler in 1951 in a paper entitled information and sufficiency. The link from information theory back to statistical physics was developed by E.T. Jaynes via the maximum entropy principle. The link back to statistics is now well established [3, 4, 5].

The relation between information theory and gambling was established by Kelly[6]. Logarithmic terms appear because we are interested in the exponent in an exponential growth rate of of our wealth. Later Kelly's approach has been generalized to training of stocks although the relation to information theory is weaker [7].

Related quantities appear in statistics, statistical mechanics, information theory and finance, annd we are interested in a theory that describes when these relations are exact and when they just work by analogy. First we introduce some general results about optimization on convex sets. This part applies exactly to all the topics under consideration and lead to Bregman divergences. Secondly, we introduce a notion of sufficiency and show that this leads to information divergence and logarithmic score. This second step is not always applicable which explains when the different topics are really different.

Proofs of the theorems in this short paper can be found in an appendix that is part of the arXiv version of the paper.

2. STATE SPACE

The present notion of a state space is based on [8], and is mainly relevant for quantum systems.

Before we do anything we prepare our system. Let $\mathcal{P}$ denote the set of preparations. Let p_0 and p_1 denote two preparations. For $t \in [0,1]$ we define $(1-t) \cdot p_0 + t \cdot p_1$ as the preparation obtained by preparing p_0 with probability $1-t$ and t with probability t. A measurement m is defined as an affine mapping of the set of preparations into a set of probability measures on some measurable space. Let $\mathcal{M}$ denote a set of feasible measurements. The state space $\mathcal{S}$ is defined as the set of preparations modulo measurements. Thus, if p_1 and p_2 are preparations then they represent the same state if $m(p_1) = m(p_2)$ for any $m \in \mathcal{M}$.

In statistics the state space equals the set of preparations and has the shape of a simplex. The symmetry group of a simplex is simply the group of permutations of the extreme points. In quantum theory the state space has the shape of the density matrices on a complex Hilbert space and the state space has a lot of symmetries that a simplex does not have. For simplicity we will assume that the state space is a finite dimensional convex compact space.

3. OPTIMIZATION

Let $\mathcal{A}$ denote a subset of the feasible measurements $\mathcal{M}$ such that $a \in A$ maps S into a distribution on the real numbers i.e. a random variable. The elements of $\mathcal{A}$ may represent actions like the score of a statistical decision, the energy extracted by a certain interaction with the system, (minus) the length of a codeword of the next encoded input letter using a specific code book, or the revenue of using a certain portfolio. For each $s \in \mathcal{S}$ we define $F(s) = \sup_{a \in \mathcal{A}} E[a(s)]$. We note that F is convex but F need not be strictly convex. We say that a sequence of actions $(a_n)_n$ is *asymptotically optimal* for the state s if $E[a_n(s)] \to F(s)$ for $n \to \infty$.

If the state is s_1 but one acts as if the state were s_2 one suffers a *regret* that equals the difference between what one achieves and what could have been achieved.

Definition 1. If $F(s_1)$ is finite *the regret* is defined by

$$D_F(s_1, s_2) = F(s_1) - \sup_{(a_n)_n} \limsup_{n \to \infty} E[a_n(s_1)] \quad (1)$$

where the supremum is taken over all sequences $(a_n)_n$ that are asymptotically optimal over s_2.

Proposition 2. *The regret D_F has the following properties:*

- $D_F(s_1, s_2) \geq 0$ with equality if $s_1 = s_2$.
- $\sum t_i \cdot D_F(s_i, \tilde{s}) \geq \sum t_i \cdot D_F(s_i, \hat{s}) + D_F(\hat{s}, \tilde{s})$ where $(t_1, t_2, \ldots, t_\ell)$ is a probability vector and $\hat{s} = \sum t_i \cdot s_i$.
- $\sum t_i \cdot D_F(s_i, \tilde{s})$ is minimal when $\hat{s} = \sum t_i \cdot s_i$.

If the state space is finite dimensional and there exists a unique action a_2 such that $F(s_2) = E[a(s_2)]$ then $D_F(s_1, s_2) = E[a_1(s_1)] - E[a_2(s_1)]$. If unique optimal actions exists for any state then F is differentiable which implies that the regret can be written as a *Bregman divergence* in the following form

$$D_F(s_1, s_2) = F(s_1) - (F(s_2) + \langle s_1 - s_2, \nabla F(s_2) \rangle). \quad (2)$$

In the context of forecasting and statistical scoring rules the use of Bregman divergences dates back to [9].

We note that $D_{F_1}(s_1, s_2) = D_{F_2}(s_1, s_2)$ if and only if $F_1(s) - F_2(s)$ is an affine function of s. If the state s_2 has the unique optimal action a_2 then

$$F(s_1) = D_F(s_1, s_2) + E[a_2(s_1)] \quad (3)$$

so the function F can be reconstructed from D_F except for an affine function of s_1. The closure of the convex hull of the set of functions $s \to E[a(s)]$ is uniquely determined by the convex function F.

4. SUFFICIENCY

Let $(s_\theta)_\theta$ denote a family of states and let Φ denote *a completely positive* transformation $\mathcal{S} \to \mathcal{T}$ where $\mathcal{S}$ and $\mathcal{T}$ denote state spaces. Then Φ is said to be *sufficient* for $(s_\theta)_\theta$ if there exists a completely positive transformation $\Psi : \mathcal{T} \to \mathcal{S}$ such that $\Psi(\Phi(s_\theta)) = s_\theta$.

We say that the regret D_F on the state space S satisfies the *sufficiency property* if $D_F(\Phi(s_1), \Phi(s_2)) = D_F(s_1, s_2)$ for any completely positive transformation $\mathcal{S} \to \mathcal{S}$ that is sufficient for (s_1, s_2). The notion of sufficiency as a property of divergences was introduced in [10]. The crucial idea of restricting the attention to transformations of the state space into itself was introduced in [11].

Theorem 3. *Assume that S is a state space. If the divergence D_F satisfies the sufficiency property then for any state s and any completely positive transformation $\Phi : S \to S$ one has $F(\Phi(s)) = F(s)$.*

If the alphabet size is two the above condition on F is sufficient to conclude that

$$D_F(\Phi(s_1), \Phi(s_2)) = D_F(s_1, s_2). \quad (4)$$

Theorem 4. *Assume that the state space S is a classical or quantum state space on three or more letters. If the regret D_F satisfies the sufficiency property, then F is proportional to the entropy function and D_F is proportional to information divergence (relative entropy).*

This theorem can be proved via a numer of partial results as explained in the next section.

5. APPLICATIONS

5.1. Statistics

Consider an experiment with $\mathcal{X} = \{1, 2, \ldots, \ell\}$ as sample space. A *scoring rule* f is defined as a function with domain $\mathcal{X} \times M_1^+(\mathcal{X}) \to R$ such that the score is $f(x, Q)$ when the prediction was given by Q and $x \in \mathcal{X}$ has been observed. A scoring rule is *proper* if for any probability measure $P \in M_1^+(\mathcal{X})$ the score $\sum_{x \in \mathcal{X}} P(x) \cdot f(x, Q)$ is minimal when $Q = P$.

Theorem 5. *The scoring rule f is proper is and only if there exists a smooth function F such that $f(x, Q) = D_F(\delta_x, Q) + \tilde{f}(x)$.*

Definition 6. A *strictly local scoring rule* is a scoring rule of the form $f(x, Q) = g(Q(x))$.

Lemma 7. *On a finite space a Bregman divergence that satisfies the sufficiency condition gives a strictly local scoring rule.*

The following theorem was given in [11] with a much longer proof.

Theorem 8. *On a finite alphabet with at least three letters a Bregman divergence that satisfies the sufficiency condition is proportional to information divergence.*

Proof. Since any strictly local proper scoring rule corresponds to separable divergence a divergence that is Bregman and satisfies sufficiency must also be separable. If the alphabet size is at least three the only separable divergences that are Bregman divergences are the ones proportional to information divergence [10]. □

5.2. Information theory

Let $b_1, b_2, \ldots, b_n$ denote the letters of an alphabet and let $\ell(\kappa(b_i))$ denote the length of the codeword $\kappa(b_i)$ according to some code book κ. If the code is uniquely decodable then $\sum 2^{-\ell(\kappa(b_i))} \leq 1$. Note that $\ell(\kappa(b_i))$ is an integer. If only integer values of ℓ are allowed then h is piecewise linear and sufficiency is not fulfilled. If arbitrary real numbers are allowed then it obvious we get a proper local scoring rule.

5.3. Statistical mechanics

Statistical mechanics can be stated based on classical mechanics or quantum mechanics. For our purpose this makes no difference because Theorem 4 can be applied for both classical systems and quantum systems.

Proof of Theorem 4. If we restrict to any commutative sub-algebra the divergence is proportional to information divergence as stated in Theorem 8 so that F is proportional to the entropy function H restricted to the sub-algebra. Any state generates a commutative sub-algebra so the function F is proportional to H on all states and the divergence is proportional to information divergence. □

Assume that a heat bath of temperature T is given and that all the states are close to the state of the heat bath. An action $a \in A$ is some interaction with the thermodynamic system that extracts some energy from the system. In thermodynamics the quantity $F(s) = \sup_{a \in A} E[a(s)]$ is normally called the free energy. If the temperature is kept fixed under all interactions F is called *Helmholtz free energy*. Any sufficient transformation Φ for s_1 and s_2 is quasi-static and can be approximately realized by a physical process Ψ that is reversible in the thermodynamic sense of the word.

$$D_F(\Phi(s_1), \Phi(s_2)) = a_{\Phi(s_1)}(\Phi(s_1)) - a_{\Phi(s_2)}(\Phi(s_1)). \tag{5}$$

Now

$$\begin{aligned} a_{\Phi(s_2)}(\Phi(s_2)) &= \left(a_{\Phi(s_2)} \circ \Phi\right)(s_2) \\ &\leq a_2(s_2) = a_2(\Psi(\Phi(s_2))) \\ &= (a_2 \circ \Psi)(\Phi(s_2)) \leq a_{\Phi(s_2)}(\Phi(s_2)). \end{aligned} \tag{6}$$

Hence $a_{\Phi(s_2)} = a_2 \circ \Psi$ so that

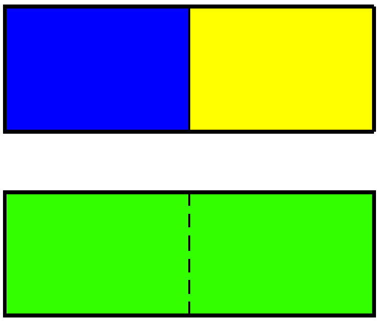
$$\begin{aligned} &D_F(\Phi(s_1), \Phi(s_2)) \\ &\quad = (a_1 \circ \Psi)(\Phi(s_1)) - (a_2 \circ \Psi)(\Phi(s_1)) \\ &\qquad = a_1(s_1) - a_2(s_1) = D_F(s_1, s_2). \end{aligned} \tag{7}$$

The amount of extractable energy Ex is proportional to information divergence. The quotient between extractable energy and information divergence depends on the temperature and one may even define the absolute temperature via the formula

$$Ex = kT \cdot D(s_1 \| s_2) \tag{8}$$

where $k = 1.381 \cdot 10^{-23} \mathrm{J}/\mathrm{K}$ is Boltzmann's constant. Equation (8) was derived already in [12] by a similar argument.

According to Equation (8) any bit of information can be converted into an amount of energy! One may ask how this is related to the mixing paradox (a special case of Gibbs' paradox). Consider a container divided by a wall with a blue and a yellow gas on each side of the wall. The question is how much energy can be extracted by mixing the gasses?



We loose one bit of information about each molecule by mixing the gasses, but if the color is the *only difference* no energy can be extracted. This seems to be in conflict with Equation (8), but in this case different states cannot be converted into each other by reversible processes. For instance one cannot convert the blue gas into the yellow gas. To get around this problem one can restrict the set of preparations and one can restrict the set of measurements. For instance one may simply ignore measurements of the color of the gas. What should be taken into account and what should be ignored, can only be answered by an experienced physicist. Formally this solves the mixing paradox but from a practical point of view nothing has been solved. If for instance the molecules in one of the gasses are much larger than the molecules in the other gas then a semi-permeable membrane can be used to create an osmotic pressure that can be used to extract some energy. It is still an open question which differences in properties of the two gasses that can be used to extract energy.

5.4. Portfolio theory

Let $X_1, X_2, \ldots, X_k$ denote *price relatives* for a list of stocks. For instance $X_5 = 1.04$ means that stock no. 5 increases its value by 4 %. A *portfolio* is a probability vector $\vec{b} = (b_1, b_2, \ldots, b_k)$ where for instance $b_5 = 0.3$ means that 30 % of your money is invested in stock no. 5. The total price relative is $X_1 \cdot b_1 + X_2 \cdot b_2 + \cdots + X_k \cdot b_k = \vec{X} \cdot \vec{b}$. We now consider a situation where the stocks are traded once every day. For a sequence of price relative vectors $\vec{X}_1, \vec{X}_2, \ldots \vec{X}_n$ and *a constant re-balancing portfolio* $\vec{b}$ the wealth after n days is

$$S_n = \prod_{i=1}^{n} \left\langle \vec{X}_i, \vec{b} \right\rangle \tag{9}$$

According to law of large numbers

$$\frac{1}{n} \log (S_n) \to E\left[\log \left\langle \vec{X}, \vec{b} \right\rangle\right] \tag{10}$$

Here $E\left[\log \left\langle \vec{X}, \vec{b} \right\rangle\right]$ is proportional to the *doubling rate* and is denoted $W\left(\vec{b}, P\right)$ where P indicates the probability distribution of $\vec{X}$. Our goal to maximize $W\left(\vec{b}, P\right)$ by choosing an appropriate portfolio $\vec{b}$.

Let $\vec{b}_P$ denote the portfolio that is optimal for P. As proved in [7]

$$W\left(\vec{b}_P, P\right) - W\left(\vec{b}_Q, P\right) \leq D\left(P \| Q\right). \tag{11}$$

Theorem 9. *The Bregman divergence*

$$W\left(\vec{b}_P, P\right) - W\left(\vec{b}_Q, P\right) \tag{12}$$

satisfies the equation

$$W\left(\vec{b}_P, P\right) - W\left(\vec{b}_Q, P\right) = D\left(P \| Q\right). \tag{13}$$

if and only if the measure P on k distinct vectors of the form $(a_1, 0, 0, \ldots 0)$*,* $(0, a_2, 0, \ldots 0)$*, until* $(0, 0, \ldots a_k)$*.*

6. CONCLUSION

On the level of optimization the theory works out in exactly the same way in statistics, information theory, statistical mechanics, and portfolio theory. The sufficiency condition is more complicated to apply. It requires that we restrict to a certain class of mappings of the state space into itself. In the case where the state space can be identified with a set of density matrices one should restrict to completely positive maps. In case the state space has a different structure it is not obvious which mappings one should restrict to. The basic problem is that we have to introduce a notion of tensor product for convex sets and it is not obvious how to do this, but this will be the topic of further investigations and results on this topic may have some impact on our general understanding of quantum theory.

The original paper of Kullback and Leibler [13] was called “On Information and Sufficiency”. In the present paper we have made the relation between information divergence and the notion of sufficiency more explicit. The idea of sufficiency has different consequences in different applications but in all cases information divergence prove to be the quantity that convert the general notion of sufficiency into a number. For specific applications one cannot identify the sufficient variables without studying the specific application in detail. For problems like the the mixing paradox there is still no simple answer to the question about what the sufficient variables are, but if the sufficient variables have been specified we have the mathematical framework to develop the rest of the theory in a consistent manner.

7. REFERENCES

[1] J. McCarthy, “Measures of the value of information,” *Proc. Nat. Acad. Sci.*, vol. 42, pp. 654–655, 1956.

[2] A. P. Dawid, S. Lauritzen, and M. Parry, “Proper local scoring rules on discrete sample spaces,” *The Annals of Statistics*, vol. 40, no. 1, pp. 593–608, 2012.

[3] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.

[4] A. R. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998, commemorative issue.

[5] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory. Now Publishers Inc., 2004.

[6] J. L. Kelly, “A new interpretation of information rate,” *Bell System Technical Journal*, vol. 35, pp. 917–926, 1956.

[7] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[8] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*, ser. North-Holland Series in Statistics and Probability, P. R. Krishnaiah, C. R. Rao, M. Rosenblatt, and Y. A. Rozanov, Eds. Amsterdam: North-Holland, 1982, vol. 1.

[9] A. D. Hendrickson and R. J. Buehler, “Proper scores for probability forecasters,” *Ann. Math. Statist.*, vol. 42, pp. 1916–1921, 1971.

[10] P. Harremoës and N. Tishby, “The information bottleneck revisited or how to choose a good distortion measure,” in *Proceedings ISIT 2007, Nice*. IEEE Information Theory Society, June 2007, pp. 566–571. [Online]. Available: www.harremoes.dk/Peter/flaske2.pdf

[11] J. Jiao, T. C. amd Albert No, K. Venkat, and T. Weissman, “Information measures: the curious case of the binary alphabet,” *Trans. Inform. Theory*, vol. 60, no. 12, pp. 7616–7626, Dec. 2014.

[12] P. Harremoës, *Time and Conditional Independence*, ser. IMFUFA-tekst. IMFUFA Roskilde University, 1993, vol. 255, original in Danish entitled Tid og Betinget Uafhængighed. English translation partially available.

[13] S. Kullback and R. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

ITERATIVE DECODING OF PRODUCT CODES

Jørn Justesen[1]

jorn@justesen.info

ABSTRACT

Product codes provide excellent performance in high rate optical communication when decoded by iterating the decoding of the component codes. We analyze the decoding of error patterns where any subset of j rows and columns contains less than *jd/*2 errors. We prove that all such error patterns can be decoded when the component decoding algorithm is applied to rows and columns in the right sequence, and we discuss how this performance can be achieved by passing messages between component decoders.

1. INTRODUCTION

Product codes are important 2d codes offering constructions of long codes from relatively short component codes. These codes have a long tradition in coding theory, but most published results are rather weak. The potential of iterated decoding has certainly been noticed by many researchers, but it is only recently than an analysis of this approach has been presented [1]. Decoders based on iterative decoding are used in optical transmission systems even at very high data rates [2]. It is known that for component (n,k,d) codes, $t < d/2$, the decoding is possible with high probability when the number of errors is less than $n(t+\sqrt{t\log t})$. However, several questions about the performance remain unclear, including the error probability and the effects of component decoding errors.

2. DECODABLE ERROR PATTERNS

We consider only standard product codes here, although many of the conclusions extend to related structures where each code symbols is part of two different component codes (braided codes, graph codes, etc.). For simplicity we assume that the two component codes are the same (n,k,d) code (usually a binary BCH code or a RS code). The decoding is based on a hard decision decoding algorithm for the component code decoding $t = \lfloor (d-1)/2 \rfloor$ errors. It is well known that the minimum distance is $D = d^2$, but to get a useful decoder we must correct error patterns of weight much greater than $D/2$.

It follows from a result about random graphs that iterative decoding using a t error decoding algorithm for the component code succeeds with high probability as long as the average number of errors in each row or column is less than $t+\sqrt{t\log t}$ [1]. For this result to apply, we must keep t fixed and let n be large, but the conclusion is a good approximation for the parameters typically used in optical communication, n=1023, t=3. It is also assumed that there are no decoding errors in the component codes, and this condition is at best a rough approximation to the real situation.

We noted in [3] that a product code can always correct an errors pattern is any subset of m rows and columns contains less than $jd/2$ errors. *Proof*: If the difference between two errors patterns with the same syndrome is nonzero on j rows or columns, it is a codeword and as such has weight at least *jd.* Thus there is a unique error pattern satisfying the weight condition. In particular this restriction gives that the total number of errors is less than *nd/2*, less than required for iterative decoding. However, even though this limit is below the threshold for iterative decoding, it is close to what can usually be decoded, since the number of iterations in most real decoders is quite low.

If the rows and columns are decoded in the right order, any error pattern satisfying the weight restriction can be decoded by iterative decoding. *Proof*: It follows from the condition that at least one row and column contains at most t errors, and the result follows by induction.

3. TESTING FOR THE WEIGHT CONDITION

It is not obvious that it can be decided effectively whether a given error pattern satisfies the weight condition. Such problems are usually treated in graph theory terminology. In this case the symbols of the product code are associated with the branches of a complete bipartite graph, the nodes representing row and column component constraints. Thus the error pattern is a subset of this graph where only error branches are preserved. The question we want to answer is if there is a subset, *S*', of the error graph with 2*j* nodes and *E'=e(S')* edges, such that the density of the subgraph satisfies *E'/j* ≥ *d/*2. We can express the number of edges in S' as

$$\left[\sum \deg(v) - e(S', S")\right]/2$$

subtracting the edges connecting *S*' to the complement *S*" from the total number of edges connecting to nodes in *S*'. Since the sum of the degrees of all nodes is 2*E*, we can write the condition as

$$\sum\nolimits_{S'+S"} \deg(v) - \left\{\sum\nolimits_{S'} \deg(v) + e(S', S") + jd/2\right\} \geq 0$$

We associate a capacity of 1 (undirected) with each error branch. Furthermore we add a source node and a sink node such that each node, *v*, in the error graph is connected to the source with a branch of capacity *deg(v),* and each node is connected to the sink with a branch of capacity *d*/4. The sum in {} above is exactly the cost of a cut separating *S*"+source from *S*'+sink. Thus there is a set satisfying the condition if the cost of the min cut satisfies it, and we may answer the question by applying one of the max flow / min cut algorithms. The computation is simplified by scaling the capacities to small integers [4].

4. ERROR PROBABILITY

If random errors occur with probability *p*, the number of errors in each row and column follows a Poisson distribution with mean *np* (for large *n*). Thus *np* is the density of the array as a whole. However if we take some initial decoding steps correcting rows and columns with at most *np*/2 errors, we get a somewhat smaller array with a higher density. When more than *np*/2 errors are corrected in the following stages, the density must decrease.

Thus on the average, there is a large array with maximal density, and to get a sufficiently low bound on the probability of decoding, we choose *d* large enough that the array has density below *d*/2 with sufficiently high probability. The average size of the critical array and the expected density can be calculated using the random graph analysis in [1]. Thus we can correctly decode the critical array if the algorithm decodes up to the weight constraint. If no error pattern satisfies the constraint, the algorithm fails, but a decoding error with a large support is highly unlikely

Actual decoding errors are almost always associated with low weight codewords. We can find a union type upper bound on this probability as

$$\sum_{j\geq d} \binom{n}{j}^2 2^{jd/2} p^{jd/2}$$

This probability is very small for *p<d/(2n)* and the largest term occurs for *j=d*. This term could be further reduced by using the actual number of weight *d* words in the component code.

5. DECODING UP TO THE WEIGHT LIMIT

Most practical decoders alternate between decoding all rows and all columns. However, for typical parameters (small *t* and large *n*), the performance is degraded by decoding errors, and the effects are difficult to analyze. We shall consider algorithms that decode only a subset of the rows/columns in each step, and we do not allow a component decoder to change a symbol after a decision has been made. Thus in case of decoding errors, the algorithm would need to back-track to an earlier stage and make a different choice.

We assume that the only type of component code decoding to be used is hard decision *t* error correction. Thus the syndrome is computed from the received values with the possible changes in symbols that have been decided, and the component decoded determines at most *t* error locations and values, or it indicates that the word cannot be decoded. The error values are passed as messages to the row/column in question, but the symbols and the syndromes are not changed yet.

Thus after an initial round of component code decoding, each row and column is marked as either decodable or not decodable. Later some rows and columns are marked as decided. If a component decoder finds an error in a position that is already decided, the code is considered not decodable.

In each step one or more decodable rows (columns) are selected and a decision is made. Thus the symbols are fixed and the status is changed to 'decided'. In the positions where errors have been corrected, the column syndrome changes unless the message coincided with the decision. Thus the status of each such column is updated, and if it is decodable, new messages are computed. In positions that were not changed, there may be messages. In those cases the syndrome does not change, but the status changes to 'not decodable' and all messages from the column are erased.

The process continues until no more rows and columns can be decoded. If all rows and columns are decided, a codeword in the product code has been reached, and the error pattern is tested against the weight constraint. If it passes, a correct decision has been reached, otherwise a new attempt is necessary. If one or more rows or and columns are not decided, we go back and make a new attempt.

The algorithm makes a tree search through the possible decisions until the correct codeword is obtained (or it has been decided that there is no such word).

As mentioned in the previous section, it is usually possible to reduce the number of rows and column by decoding error patterns of low weight, and the probability of decoding error is low in these cases. However, the important part of the algorithm is the decoding of the critical subset, where most of the rows and column have close to *t* errors, and the error graph is connected.

If a component decoder correctly locates the error positions in a row, the corresponding messages from the columns may be missing, since a significant fractions of the columns have more than *t* errors and may be decoded in error. However, a similar consideration shows that there are few messages in positions that are not corrected, since each column can only contribute *t* false messages. Thus the rows should be selected first to agree with columns already decoded, and next to have few messages is positions that are not corrected. The choice could be further refined by considering the new syndromes that are generated in columns with errors, and it is preferable to decode rows where an error has already been decided (to stay within a connected part of the error graph).

In principle it is not possible to avoid more than one attempt. An error pattern of weight *jd* on *j* rows and columns could be split into two error patterns of almost the same weight, each with *t* of t+1 errors in each row and column. Thus a decision cannot be reached until the weight of the entire pattern is known. On the other hand the number of iterations is quite small for the density considered here. If correct decisions are made, the errors are corrected in two or three row/column iterations.

6. 2-D APPLICATIONS

When a page is protected by an error correcting code or the page contains an area which uses such a code (like a QD code), particular attention must be paid to alignment errors and likely forms of degradation. RS codes are often used to correct small patches of errors. There is a long tradition for using some form of product codes on 2d storage media, and such codes allow the correction of errors that typically affect only a small number of errors in each row or column. Such errors could be caused by scratches in the surface or by scanning errors due to improper alignment.

7. REFERENCES

[1] J. Justesen and S. Forchhammer, *Two-Dimensional Information Theory and Coding*. Cambridge, UK: Cambridge University Press, 2010.

[2] J. Justesen," Performance of product codes and related Structures with iterated decoding" *IEEE Trans Communications*, vol. 59, pp. 407-415, Feb. 2011.

[3] J. Justesen, K.J. Larsen, and L.A. Pedersen, "Error correcting coding for OTN," *IEEE Communications Magazine*, vol. 48, pp. 70–75, Sep. 2010.

[4] E. Lawler, *Combinatorial Optimization - Networks and Matroids.* New R\York: Holt, Rinehart and Winston, 1976.

CAUSALITY AND DIRECTED INFORMATION ESTIMATION AS A HYPOTHESIS TEST

Ioannis Kontoyiannis[1] *and Maria Skoularidou*[2]

[1]Department of Informatics, Athens University of Economics and Business,
Patission 76, Athens 10434, Greece. yiannis@aueb.gr
[2]Department of Informatics, Athens University of Economics and Business,
Patission 76, Athens 10434, Greece. m.skoularidou@gmail.com

ABSTRACT

We consider the problem of estimating the directed information rate between two Markov chains with memory length $k \geq 1$, using the plug-in estimator. We show that the estimator is asymptotically Gaussian and conclude that it converges at a rate $O(1/\sqrt{n})$, which, we argue, is best possible. We also draw a connection between this estimation problem and that of performing a hypothesis test for the presence of causal influence between the two processes. Under the null hypothesis, which corresponds to the absence of causality, we show that the plug-in estimator has an asymptotic χ^2 distribution. Also, we establish that this estimator can be precisely expressed in terms of the classical likelihood ratio. Combining these two results facilitates the design of a Neyman-Pearson likelihood ratio test for the presence of causal influence.

1. INTRODUCTION

Throughout the sciences and engineering, the χ^2 test for independence is one of the most commonly used statistical techniques. Given a sample of independent and identically distributed data pairs (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$, suppose we wish to test whether the X and Y variables are independent or not. Assuming both sets of variables take on finitely many values, we can compare the joint empirical distribution $\hat{P}_{XY,n}(a, b)$ with the product of the empirical marginals $\hat{P}_{X,n}(a)\hat{P}_{Y,n}(b)$; as usual, $\hat{P}_{XY,n}(a, b)$ denotes the proportion of times the pair (a, b) appears in the whole sample, and similarly for the marginals. Following classical methodology according to the χ^2 test, we compute the normalized χ^2 distance between these two distributions,

$$\bar{\chi}^2_n = n \sum_{a,b} \frac{\big[\hat{P}_{XY,n}(a, b) - \hat{P}_{X,n}(a)\hat{P}_{Y,n}(b)\big]^2}{\hat{P}_{X,n}(a)\hat{P}_{Y,n}(b)}. \qquad (1)$$

Under the null hypothesis – assuming, that is, that the data are independent – the distribution of the statistic $\bar{\chi}^2_n$ for large n is approximately χ^2 with $(m-1)(\ell-1)$ degrees of freedom, where m, ℓ are the sizes of the alphabets of X and Y, respectively. Therefore, computing the probability of the statistic $\bar{\chi}^2_n$ under this distribution, we can decide whether or not to reject the independence hypothesis.

A different approach, closer in spirit to information-theoretic ideas, is offered by the the likelihood ratio test, which is based on the statistic,

$$\Delta_n = 2 \log \left(\frac{\prod_{i=1}^n \hat{P}_{XY,n}(X_i, Y_i)}{\prod_{i=1}^n \hat{P}_{X,n}(X_i)\hat{P}_{Y,n}(Y_i)} \right).$$

Asymptotically, Δ_n has the exact same distribution as $\bar{\chi}^2_n$, so that an analogous test can be performed. An important observation for our purposes is that this likelihood ratio test statistic can exactly be expressed as a mutual information,

$$\Delta_n = 2nI(\hat{X}; \hat{Y}) = 2nD\big(\hat{P}_{XY,n} \| \hat{P}_{X,n}\hat{P}_{Y,n}\big), \qquad (2)$$

where $\hat{X}, \hat{Y}$ are distributed according to the empirical distribution $\hat{P}_{XY,n}$. One way to look at the difference between $\bar{\chi}^2_n$ and Δ_n is that, instead of the χ^2 distance used in (1), the likelihood ratio test statistic (2) examines the (normalized) relative entropy distance between $\hat{P}_{XY,n}$ and $\hat{P}_{X,n}\hat{P}_{Y,n}$. And yet another way to interpret Δ_n is as the "plug-in" estimate of the mutual information $I(X_1; Y_1)$ of the data, using their empirical distribution.

The asymptotic distribution of Δ_n has been re-derived several times historically. In its general form it goes back to the classical result of Wilks [1], see also the text [2]; and more recently it has also appeared in an information-theoretic context, see, e.g., [3, 4].

In this work we examine the problem of estimating a different information-theoretic functional: If $\boldsymbol{X} = \{X_n\}$ and $\boldsymbol{Y} = \{Y_n\}$ are two finite-valued random process, then the *directed information* $I(X_1^n \to Y_1^n)$ between $X_1^n = (X_1, X_2, \ldots, X_n)$ and $Y_1^n = (Y_1, Y_2, \ldots, Y_n)$ is defined as,

$$I(X_1^n \to Y_1^n) = H(Y_1^n) - \sum_{i=1}^n H(Y_i | Y_1^{i-1}, X_1^i), \qquad (3)$$

and the *directed information rate* between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is,

$$I(\boldsymbol{X} \to \boldsymbol{Y}) = \lim_{n\to\infty} \frac{1}{n} I(X_1^n \to Y_1^n), \qquad (4)$$

whenever the limit exits. Directed information was introduced by Massey [5] and Kramer [6], building on earlier work by Marko [7], in order to provide capacity characterizations for channels with causal feedback. Subsequent work in this direction and in other applications is surveyed in [8, 9].

The approach we take here is to consider the problem of estimating the directed information rate, by tracing the path described above in connection with the mutual information in the reverse direction. Our main results are stated in the following section; their proofs and more general results can be found in the longer manuscript [9].

Before closing this introduction, some bibliographical remarks are in order. The problem of testing for causality has a long history. Perhaps the most prominent example is the Granger causality test, which frames the problem of detecting causal influence in terms of conditional independence, a setting we will also follow. Granger [10] uses an autoregressive model (later extended in several directions, most notably to generalized linear models), within which the causality hypothesis is tested. The connection between this test and directed information has been explored in several directions; see [11] for a comprehensive review. Also, several different approaches to the problem of directed information estimation have appeared in the literature in recent years, see, e.g., [12] and [13], where applications in genetics and neuroscience are considered.

In terms of the present development, the most interesting work is [14], where several new estimators for the directed information rate are introduced and they are shown to be consistent under very general conditions. For some of these estimators, particularly those based on the celebrated context tree weighting algorithm, detailed convergence bounds are also obtained. Compared to the estimators of [14], the plug-in suffers two well-known drawbacks. It is computationally ineffective for large alphabet sizes and long memory processes, and its use is restricted to Markovian data. On the other hand, using the plug-in facilitates the connection with hypothesis testing developed here, and also makes it possible to obtain much more accurate, exact asymptotics, instead of convergence bounds. In fact, the converse result in [14, Proposition 3] suggests that the $O(1/\sqrt{n})$ convergence rate of the plug-in estimator established in Section 2 is optimal. Moreover, our convergence results are obtained under conditions at least as general as those for the bounds [14], and the resulting rates are slightly sharper.

2. DIRECTED INFORMATION

2.1. Preliminaries

Suppose X is a discrete random variable with values in a finite set A, and with a distribution described by its probability mass function, $P_X(x) = \Pr\{X = x\}$, for $x \in A$. The entropy of X is defined by, $H(X) = H(P_X) = -\sum_{x\in A} P(x)\log P(x)$, where, throughout the paper, log denotes the natural logarithm to base e. Viewed as a single random element, the joint entropy of any finite collection of random variables $X_1^n = (X_1, X_2, \dots, X_n)$ is defined analogously, and the mutual information between two random variables X and Y is $I(X;Y) = H(X) + H(Y) - H(X,Y)$. As above, we generally write $X_i^j = (X_i, X_{i+1}, \dots, X_j)$, $i \leq j$, for vectors of random variables and similarly $a_i^j = (a_i, a_{i+1}, \dots, a_j) \in A^{j-i+1}$, $i \leq j$, for strings of individual symbols from a finite set.

The joint distribution of an arbitrary number of discrete random variables is described by their joint probability mass function. For example, the joint distribution of (X, Y, Z) is denoted, $P_{XYZ}(x, y, z)$. We write the induced marginal distributions in the obvious way, e.g., $P_{XY}(x, y)$ and $P_Z(z)$, and the induced conditionals are similarly denoted, e.g., $P_{XY|Z}(x, y|z)$.

2.2. The directed information rate of Markov chains

Let $\boldsymbol{X} = \{X_n\ ;\ n \geq 0\}$ and $\boldsymbol{Y} = \{Y_n\ ;\ n \geq 0\}$ be two discrete processes with values in the finite alphabets A and B, respectively. For each $n \geq 1$, recall the definition of the directed information $I(X_1^n \to Y_1^n)$ in (3). This is zero exactly when Y_i is conditionally independent of X_1^i, given its past Y_1^{i-1}, for each $i = 1, 2, \dots, n$. The natural interpretation of this equivalence is to say that the directed information is zero if and only if $\boldsymbol{X}$ has no *causal* influence on $\boldsymbol{Y}$. We are interested in the problem of estimating the directed information rate, $I(\boldsymbol{X} \to \boldsymbol{Y})$, defined in (4).

From now on we assume that the pair process,

$$\{(X_n, Y_n)\ ;\ n \geq -k+1\},$$

is an ergodic (namely, irreducible and aperiodic) Markov chain on the alphabet $A \times B$, of memory length $k \geq 1$, and with an arbitrary initial distribution for (X_{-k+1}^0, Y_{-k+1}^0). We write $\{(\bar{X}_n, \bar{Y}_n)\}$ for the stationary version of the original chain, namely, with (X_{-k+1}^0, Y_{-k+1}^0) distributed according to the unique invariant measure of the bivariate chain.

The following proposition shows that, under appropriate conditions, the directed information rate can be expressed as a functional of only the $(k+1)$-dimensional distribution of $\{(X_n, Y_n)\}$, so that it can easily be estimated and a detailed analysis of the corresponding estimates can be given; see Section 2.3. Although the results of Proposition 2.1 have appeared, at least implicitly, before, we state them here for ease of reference.

Proposition 2.1 *If the Markov chain $\{(X_n, Y_n)\}$ is ergodic, it has memory no larger than k, and an arbitrary initial distribution, then:*

(i) *The entropy rate $H(\boldsymbol{Y})$ of the univariate process $\boldsymbol{Y} = \{Y_n\}$ exists and,*

$$H(\boldsymbol{Y}) = \lim_{n\to\infty} \frac{1}{n} H(Y_1^n) = \lim_{n\to\infty} \frac{1}{n} H(\bar{Y}_1^n).$$

(ii) *The directed information rate $I(\boldsymbol{X} \to \boldsymbol{Y})$ exists and it equals,*

$$H(\boldsymbol{Y}) - H(\bar{Y}_0|\bar{X}_{-k}^0, \bar{Y}_{-k}^{-1}),$$

where $H(Y|X) = H(X,Y) - H(X)$ denotes the conditional entropy.

(iii) *If $\boldsymbol{Y} = \{Y_n\}$ is also a Markov chain of order no larger than k, then,*

$$I(\boldsymbol{X} \to \boldsymbol{Y}) = I(\bar{Y}_0; \bar{X}^0_{-k}|\bar{Y}^{-1}_{-k}),$$

where $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$ denotes the conditional mutual information.

Remarks.

1. Suppose $\{(X_n, Y_n)\})$ is a Markov chain, not necessarily stationary, with memory no larger than some fixed k. For the sake of convenience we assume throughout the remainder of this section that $\{(X_n, Y_n)\}$ has a strictly positive transition matrix Q,

$$Q(a_k, b_k|a_0^{k-1}, b_0^{k-1}) > 0,$$

for all $a_0^k \in A^{k+1}$, $b_0^k \in B^{k+1}$. As discussed below, this assumption can be significantly relaxed.

2. The directed information rate $I(\boldsymbol{X} \to \boldsymbol{Y})$ admits important operational interpretations. For example, in the case of a stationary kth order Markov chain $\{(X_n, Y_n)\}$ such that $\{Y_n\}$ is also a kth order chain, we can use the data processing property of mutual information in the result of part (iii) of the proposition to see that $I(\boldsymbol{X} \to \boldsymbol{Y})$ equals,

$$I(Y_0; X^0_{-k}|Y^{-1}_{-k}){=}I(Y_0; X^0_{-\infty}|Y^{-1}_{-\infty}).$$

This is zero if and only if each Y_i, given its past $Y^{i-1}_{-\infty}$, is conditionally independent of $X^i_{-\infty}$, confirming our original intuition that the directed information is only zero in the absence of causal dependence.

3. In the case of a general stationary chain $\{(X_n, Y_n)\}$ without assuming anything else about the process $\{Y_n\}$, we still have that,

$$\begin{aligned} I(Y_0; X^0_{-k}|Y^{-1}_{-k}) &= I(Y_0; X^0_{-\infty}|Y^{-1}_{-k}) \\ &\geq I(Y_0; X^0_{-\infty}|Y^{-1}_{-\infty}), \end{aligned}$$

by data processing; this is zero if and only if Y_0, conditional only on its k-past Y^{-1}_{-k}, is independent of $X^0_{-\infty}$. In this case the quantity $I(Y_0; X^0_{-k}|Y^{-1}_{-k})$ is not enough to entirely characterize the absence of causal influence from $\boldsymbol{X}$ to $\boldsymbol{Y}$, but knowing its value nevertheless offers some evidence for such an influence. In particular, knowing that it is zero (or sufficiently close to zero), would still imply that $\boldsymbol{X}$ has no (or little) causal influence on $\boldsymbol{Y}$.

4. In view of the above remarks we conclude that, even if $\{Y_n\}$ is not necessarily Markovian, it is always of significant interest to estimate $I(\bar{Y}_0; \bar{X}^0_{-k}|\bar{Y}^{-1}_{-k})$. Indeed, as we explain in detail in Section 2.4, this estimation problem is intimately related to a classical Neyman-Pearson hypothesis test for the presence or absence of causality.

2.3. The plug-in estimator of $I(\boldsymbol{X} \to \boldsymbol{Y})$

Given a sample (X^n_{-k+1}, Y^n_{-k+1}) from the joint process $\{(X_n, Y_n)\}$, we define the $(k+1)$-dimensional, bivariate empirical distribution induced on $A^{k+1} \times B^{k+1}$, as,

$$\hat{P}_{X^0_{-k}Y^0_{-k},n}(a_0^k, b_0^k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X^i_{i-k}=a_0^k, Y^i_{i-k}=b_0^k\}}.$$

Motivated by the discussion in the above remarks, we now define the plug-in estimator for $I(\boldsymbol{X} \to \boldsymbol{Y})$ as

$$\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y}) = I(\hat{Y}_0; \hat{X}^0_{-k}|\hat{Y}^{-1}_{-k}),$$

where $(\hat{X}^0_{-k}, \hat{Y}^0_{-k}) \sim \hat{P}_{X^0_{-k}Y^0_{-k},n}$.

Since all the transition probabilities of the bivariate chain $\{(X_n, Y_n)\}$ are nonzero, it is easy to see that the plug-in estimator $\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y})$ converges almost surely to the desired value, $I(\boldsymbol{X} \to \boldsymbol{Y})$. The following result describes its finer asymptotic behavior.

Theorem 2.2 *Let $\{(X_n, Y_n)\}$ be a Markov chain of memory length $k \geq 1$, with an all positive transition matrix Q on the finite alphabet $A \times B = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, \ell\}$, and with an arbitrary initial distribution. Assume that the univariate process $\{Y_n\}$ is also a Markov chain with memory length k.*

(i) *If the random variables $\{X_n\}$* do *have a causal influence on the $\{Y_n\}$, equivalently, if $I(\boldsymbol{X} \to \boldsymbol{Y}) > 0$ then, as $n \to \infty$,*

$$\sqrt{n}\Big[\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y}) - I(\boldsymbol{X} \to \boldsymbol{Y})\Big] \xrightarrow{\mathcal{D}} N(0, \sigma^2),$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, the normal distribution with mean zero and variance σ^2 is denoted $N(0, \sigma^2)$, and with the variance σ^2 given by the following limit, which exists and is finite,

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left\{\log\left[\prod_{i=1}^{n} \left(\frac{P_{\bar{X}^0_{-k}\bar{Y}_0|\bar{Y}^{-1}_{-k}}(X^i_{i-k}, Y_i|Y^{i-1}_{i-k})}{P_{\bar{Y}_0|\bar{Y}^{-1}_{-k}}(Y_i|Y^{i-1}_{i-k})P_{\bar{X}^0_{-k}|\bar{Y}^{-1}_{-k}}(X^i_{i-k}|Y^{i-1}_{i-k})}\right)\right]\right\}.$$

(ii) *If the $\{X_n\}$* do not *have a causal influence on the $\{Y_n\}$, equivalently, if $I(\boldsymbol{X} \to \boldsymbol{Y}) = 0$ then, as $n \to \infty$,*

$$n\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y}) \xrightarrow{\mathcal{D}} \chi^2\Big(\ell^k(m^{k+1} - 1)(\ell - 1)\Big).$$

Theorem 2.2 is an immediate consequence of a more general result established in [9]. From the proof there, it is evident that the restriction of all-positive transition probabilities $Q(a_k, b_k|a_0^{k-1}, b_0^{k-1})$ for the chain $\{(X_n, Y_n)\}$ is unnecessary: The result of part (i) remains valid with this restriction replaced with the minimal assumption that the pair process $\{(X_n, Y_n)\}$ is irreducible and aperiodic. And for part (ii) the positivity assumption can also be significantly relaxed, in accordance with the discussion around Theorem 5.2 of [15].

An important consequence of Theorem 2.2 is the clear dichotomy between the presence and absence of causal influence: If the $\{X_n\}$ have no causal influence on the $\{Y_n\}$, then $I(\boldsymbol{X} \to \boldsymbol{Y}) = 0$ and the plug-in estimator converges at a rate $O(1/n)$. On the other hand, if such a causal influence does exist, then the directed information rate $I(\boldsymbol{X} \to \boldsymbol{Y})$ is strictly positive, and the plug-in estimator converges at the slower rate $O(1/\sqrt{n})$.

Finally, the proof of the χ^2 convergence part of the theorem exploits an interesting connection of this problem with a classical hypothesis test for causality; cf. [15].

2.4. A hypothesis test for causality

Suppose we wish to test whether or not the samples $\{X_n\}$ have a causal influence on the $\{Y_n\}$. In this context, as discussed above, this translates to testing the null hypothesis that each random variable Y_i is conditionally independent of X_{i-k}^i given Y_{i-k}^{i-1}, within the larger hypothesis that the pair process $\{(X_n, Y_n)\}$ is a kth order Markov chain on $A \times B$ with all positive transitions.

As we describe in detail in [9], all relevant transition matrices $Q = Q_\theta$ can be parametrized by a vector θ taking values in an $m^k \ell^k (m\ell - 1)$-dimensional open set Θ. Informally, the null hypothesis corresponding to each random variable Y_i being conditionally independent of X_{i-k}^i given Y_{i-k}^{i-1}, is described by transition matrices Q_θ which can be decomposed as,

$$Q_\theta(a_0, b_0|a_{-k}^{-1}, b_{-k}^{-1}) = Q_\theta^x(a_0|a_{-k}^{-1}, b_{-k}^{-1}) Q_\theta^y(b_0|b_{-k}^{-1}).$$

Formally, this can be described by a lower-dimensional parameter set Φ, which will be embedded in Θ via a map $h : \Phi \to \Theta$, such that all induced transition matrices $Q_{h(\phi)}$ correspond to Markov chains that satisfy the required conditional independence property.

In order to test the null hypothesis Φ within the general model Θ, we employ a likelihood ratio test. Specifically, we define the log-likelihood $L_n(X_{-k+1}^n, Y_{-k+1}^n; \theta)$ of the sample (X_{-k+1}^n, Y_{-k+1}^n) under the distribution corresponding to θ as,

$$\log\left[\Pr_\theta(X_1^n, Y_1^n | X_{-k+1}^0, Y_{-k+1}^0)\right],$$

so that the likelihood ratio test statistic is simply,

$$\Delta_n = 2\left\{\max_{\theta\in\Theta} L_n(X_{-k+1}^n, Y_{-k+1}^n; \theta) - \max_{\phi\in\Phi} L_n(X_{-k+1}^n, Y_{-k+1}^n; h(\phi))\right\}.$$

The key observation here is that:

$$\Delta_n = 2n\hat{I}_n^{(k)}(\boldsymbol{X} \to \boldsymbol{Y}).$$

The asymptotic properties of our plug-in estimator follow from the corresponding results about the likelihood ratio. And conversely, under the null hypothesis, part (ii) of Theorem 2.2 tells us that the distribution of Δ_n is approximately χ^2 with $\ell^k(m^{k+1}-1)(\ell-1)$ degrees of freedom. Therefore, we can decide whether or not the data offer strong enough evidence to reject the null hypothesis by examining the value of Δ_n and computing an appropriate p-value based on its asymptotic distribution.

3. ACKNOWLEDGMENTS

This work was supported in part by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) Research Funding Program "Thales - Investing in knowledge society through the European Social Fund."

4. REFERENCES

[1] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, no. 1, pp. 60–62, March 1938.

[2] E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics. Springer, 2005.

[3] J. Hagenauer and J.C. Mueller, "Genomic analysis using methods from information theory," in *Proc. of the Inform. Theory Workshop*, San Antonio, TX, October 2004, pp. 55–59.

[4] H.M. Aktulga, I. Kontoyiannis, L.A. Lyznik, L. Szpankowski, A.Y. Grama, and W. Szpankowski, "Identifying statistical dependence in genomic sequences via mutual information estimates," *EURASIP J. Bioinformatics Syst. Biol.*, vol. 2007, pp. 3:1–3:11, July 2007.

[5] J. Massey, "Rate-distortion in near-linear time," in *International Symposium on Information Theory and its Applications*, November 1990, pp. 303–305.

[6] G. Kramer, *Directed Information for Channels With Feedback*, Ph.D. thesis, Swiss Fed. Inst. Technol. (ETH), Zurich, Switzerland, 1998.

[7] H Marko, "The bidirectional communication theory: A generalization of information theory," *IEEE Trans. Comm.*, vol. 21, no. 1, pp. 1335–1351, 1973.

[8] H.H. Permuter, Y-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3248–3259, June 2011.

[9] I. Kontoyiannis and M. Skoularidou, "Estimating the directed information and testing for causality," *Preprint*, June 2015.

[10] C.W.J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[11] P-O. Amblard and O.J.J. Michel, "The relation between Granger causality and directed information theory: A review," *Entropy*, vol. 15, no. 1, pp. 113, 2012.

[12] A. Rao, A.O. Hero, D.J. States, and J.D. Engel, "Using directed information to build biologically relevant influence networks," *Journal of Bioinformatics and Computational Biology*, vol. 06, no. 03, pp. 493–519, 2008.

[13] C.J. Quinn, T.P. Coleman, N. Kiyavash, and N.G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[14] J. Jiao, H.H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6220–6242, Oct 2013.

[15] P. Billingsley, *Statistical Inference for Markov Processes*, Statistical Research Monographs. University of Chicago Press, 1961.

THE RUZSA DIVERGENCE ON GROUPS

Mokshay Madiman^1 *and Ioannis Kontoyiannis*^{2}

^{1}Department of Mathematical Sciences, University of Delaware,
501 Ewing Hall, Newark DE 19716, USA, madiman@udel.edu
^{2} Department of Informatics, Athens University of Economics and Business,
Patission 76, Athens 10434, GREECE, yiannis@aueb.gr

ABSTRACT

We introduce the notion of the Ruzsa divergence between two probability densities with respect to the Haar measure on a locally compact, Polish, abelian group, and develop its properties. Among other things, this leads to very general inequalities relating the entropies of sums and differences of indepedent random variables taking values in such a group.

1. INTRODUCTION

The entropy of sums of random variables is ubiquitous in information theory, appearing routinely when studying communication as well as compression. Therefore it is a perfectly natural question to ask: What is the most general setting in which studying these makes sense? In order to talk about a "sum", one needs at the very least a binary operation on the state space of our random variables, and in most applications of interest, one would expect the binary operation to have additional properties such as commutativity and associativity, and the existence of an identity element and inverses (so that one can talk not just about sums but also differences). In other words, perhaps the most general setting that is still natural for applications to information theory is when the state space of our signals (random variables of interest) is an *abelian group*, and it is of interest to explore what can be said about the entropies of sums and differences of such random variables.

There are numerous more concrete reasons why we should we care about such investigations. Indeed, within information theory, our work has already played a key role in recent advances in the understanding of the interference channel [1, 2], and carries much promise for other problems. In probability, our work is related to basic questions such as the rate of convergence in the (entropic) Central Limit Theorem (e.g., [3, 4, 5, 6, 7, 8, 9, 10]), even when the group is plain old $\mathbb{R}^n$. In additive combinatorics, sumset inequalities (inequalities for cardinalities of sums of sets) play a key role in this fast-developing area of mathematics, and entropy allows one to adopt a more general probabilistic approach to additive combinatorics (e.g., [11, 12, 13, 14, 15, 16, 17, 18]). And finally, in convex geometry, our study is related to the "geometrization of probability" program popularized by V. Milman (and pioneered by C. Borell and K. Ball); see, e.g., [19, 20, 21, 22, 23, 24].

The differential entropy of a random vector X with density $f(x)$ on $\mathbb{R}^n$ is

$$h(X) = h(f) := -\int_{\mathbb{R}^n} f(x) \log f(x) dx$$

where dx represents Lebesgue measure on $\mathbb{R}^n$. Key properties of differential entropy include translation-invariance, namely the fact that $h(X + b) = h(X)$ for any constant $b \in \mathbb{R}^n$, and $GL(n, \mathbb{R})$-contravariance, i.e., the fact that $h(AX) = h(X) + \log \det(A)$ for any $n \times n$ matrix A of real entries. Key properties that hold for discrete entropy but fail for differential entropy are non-negativity ($h(X)$ can lie anywhere in $[-\infty, \infty]$), and invariance with respect to bijections (as already observed, even simply scaling alters the differential entropy by an additive term).

In order to retain the translation-invariance of differential entropy, which is one reason it is such a useful functional on the space of probability measures on $\mathbb{R}^n$, we need a measure on our ambient abelian group. This is where the seemingly technical topological assumptions come in– by assuming that the group is a locally compact topological group, we are guaranteed by well known results in analysis that there exists a translation-invariant measure (namely the "Haar measure") with respect to which we can define entropy, and by assuming the group is Polish, we are guaranteed by well known results in probability that conditional distributions exist when looking at sufficiently nice random variables jointly distributed on the group. This is why the most general setting we treat is that of *locally compact, Polish, abelian groups*.

In this setting, we discuss a variety of inequalities that hold between entropies of various sums and differences of group-valued random variables. After developing some required terminology in the next section, we describe our main results. More details on all the results in this note can be found in [25].

2. DEVELOPING THE LANGUAGE

To make our discussion more precise, recall that an abelian group is a set G together with a binary operation $+$ such that $x + y = y + x$ (commutativity), $(x + y) + z = x + (y + z)$ (associativity), G has an "identity element"

0 such that $x + 0 = x$ for all x in G, and every element x has an inverse $-x$, i.e., $\forall x \in G, \exists y \in G$ such that $x + y = 0$. Under the appropriate topological assumptions of being Polish and locally compact (which we will not expand on here), an abelian group G admits a measure λ that is translation-invariant, i.e., such that

$$\lambda(A + x) = \lambda(A) \qquad \forall\, A \subset G, \forall x \in G$$

where $A+x = \{a+x : a \in A\}$. Such a measure is called a *Haar measure*, and is unique up to scaling by a positive constant.

We may now define entropy in the general setting. Let G be a Polish, locally compact, abelian group, and λ be a Haar measure on G. If μ is a probability measure on G that is absolutely continuous with respect to λ, then there exists a nonnegative function $f : G \to \mathbb{R}$ with total integral of 1 such that

$$P(X \in A) = \int_A f(x)\lambda(dx), \qquad A \in \mathcal{G},$$

which we call the *density* (or probability density function) of X or μ. The entropy of $X \sim \mu$ is defined by

$$h(X) = -\int_G f(x) \log f(x)\, \lambda(dx).$$

As is usual, we abuse notation to write $h(X)$ though h depends only on f. In general, $h(X)$ may or may not exist; if it does, it takes values in the extended real line $[-\infty, +\infty]$. In the special case of compact G, the Haar measure λ is finite, and so we can normalize it to get the "uniform" probability measure on G. Then, for every G-valued random variable X,

$$h(X) = -D(\mu\|\lambda) \leq 0.$$

The classical examples of entropy on groups are:

- G is a discrete group, λ is the counting measure, and h is the discrete entropy;
- $G = \mathbb{R}^n$, λ is Lebesgue measure, and h is differential entropy.

Just for illustration, here are 2 non-classical examples:

- Let $G = \mathbb{T}^n$, the torus with Lebesgue measure. Then h is the differential entropy on the torus.
- Let $G = (0, \infty)$ with the Haar measure $\lambda(dx) = x^{-1}dx$. if f is the density (with respect to Lebesgue measure) of a positive random variable X, then

  $$P(X \in A) = \int_A f(x)dx = \int_A xf(x)\frac{dx}{x},$$

  so

  $$\begin{aligned} h(X) &= -\int_0^\infty [xf(x)] \log[xf(x)]\lambda(dx) \\ &= -\int_0^\infty f(x)[\log x + \log f(x)]dx \\ &= h_{\mathbb{R}}(X) - E[\log X], \end{aligned}$$

  where $h_{\mathbb{R}}$ is the entropy if we were to think of X as an $\mathbb{R}$-valued as opposed to G-valued random variable (i.e., the usual differential entropy).

We cannot even talk about things like linear transformations on general groups because they do not have a linear structure. Yet one has two key properties of entropy on groups that carry over from $\mathbb{R}^n$.

Lemma 1. *(Translation-invariance) Let X be a random variable taking values in G. If $b \in G$, then*

$$h(X + b) = h(X).$$

Lemma 2. *($SL(n, \mathbb{Z})$-invariance) [26] Let X be a random variable taking values in G^n, and denote by $SL_n(\mathbb{Z})$ the set of $n \times n$ matrices A with integer entries and determinant* 1*. If $A \in SL_n(\mathbb{Z})$, then*

$$h(AX) = h(X).$$

Note that integer linear combinations of group elements always makes sense in an abelian group, e.g., $2x - 3y$ represents $x + x + (-y) + (-y) + (-y)$.

Having defined entropy, we can define related quantities such as conditional entropy and mutual information in the natural way. The conditional entropy of X given Y is

$$h(X|Y) = \int h(X|Y = y)P_Y(dy)$$

where $h(X|Y = y)$ is the entropy of the (regular) conditional distribution $P_X(\cdot|Y = y)$. Then one has two useful facts: Shannon's Chain Rule says that

$$h(X, Y) = h(Y) + h(X|Y),$$

and Jensen's inequality implies that conditioning reduces entropy (or equivalently, the *mutual information* $I(X; Y)$ is non-negative):

$$h(X) - h(X|Y) = D(p_{X,Y}\|p_X \times p_Y) := I(X; Y) \geq 0.$$

We may now define the central object of study in this note. Suppose X and Y are G-valued random variables with finite entropy. The quantity

$$d_R(X\|Y) := h(X - Y') - h(X),$$

where X and Y' are taken to be independent random vectors with Y' having the same distribution as Y, will be called the *Ruzsa divergence* between X and Y. By using translation-invariance of entropy, it is easy to check that if X and Y are independent random variables, then

$$d_R(X\|Y) = I(X - Y; Y). \tag{1}$$

In particular, $d_R(X, Y) \geq 0$ (although for some groups like $\mathbb{R}^n$, it is never 0 in non-degenerate situations).

3. MAIN RESULTS

Let us state some key properties of the Ruzsa divergence, all of which can be proved by combining the basic tools mentioned in the previous section in various ways (some rather straightforward and others a bit tricky) with the data processing inequality for mutual information.

Theorem 1. *If X_i are independent, then*

$$d_R(X_1\|X_3) \leq d_R(X_1\|X_2) + d_R(X_2\|X_3).$$

Theorem 2. *If X and Y_i are all mutually independent, then*

$$d_R\Big(X\Big\|\sum_{i=1}^{k} Y_i\Big) \leq \sum_{i=1}^{k} d_R(X\|Y_i).$$

Theorem 1 is the analog of what is called Ruzsa's triangle inequality for sumsets in additive combinatorics, and was developed for discrete groups independently by Ruzsa [12] and Tao [14]. On the other hand, Theorem 2 is the analog of what is called the Plünnecke-Ruzsa inequality for sumsets, and is equivalent to the following Submodularity Property for independent G-valued random variables:

$$h(X_1 + X_2 + X_3) + h(X_2) \leq h(X_1 + X_2) + h(X_3 + X_2).$$

For discrete groups, this Submodularity Lemma is implicit in [27] but was rediscovered and significantly generalized by [13] en route to proving some conjectures of Ruzsa [12]. Note that discrete entropy is, trivially, subadditive:

$$H(X_1 + X_2) \leq H(X_1, X_2) \leq H(X_1) + H(X_2).$$

This corresponds to putting $X_2 = 0$ in the discrete form of the Submodularity Lemma. On the other hand, entropy is not subadditive in continuous settings; it is easy to construct examples (using scaling, for instance) on $\mathbb{R}$ with

$$h(X_1 + X_2) > h(X_1) + h(X_2).$$

Note that putting $X_2 = 0$ in the Lemma is no help since $h(\text{const.}) = -\infty$. We extend both theorems to the general setting.

We also define a *conditional Ruzsa divergence*: We say that $X \leftrightarrow Z \leftrightarrow Y$ form a Markov chain if X, Z, Y are defined on a common probability space and the conditional distribution of X given (Z, Y) is the same as that of X given Z alone; equivalently $I(X; Y|Z) = 0$. If $X_1 \longleftrightarrow Y \longleftrightarrow X_2$ forms a Markov chain,

$$d_R(X_1\|X_2|Y) := h(X_1 - X_2|Y) - h(X_1|Y),$$

is the conditional Ruzsa divergence from X_1 to X_2 given Y.

If $X_1 \longleftrightarrow Y \longleftrightarrow X_2$ form a Markov chain, then

$$d_R(X_1\|X_2|Y) = I(X_1 - X_2; X_2|Y).$$

Observe that $d_R(X_1\|X_2|Y) \neq d_R(X_2\|X_1|Y)$ in general, but both are non-negative under the Markov condition Conditioning turns out to reduce Ruzsa divergence: If X_1 is independent of (Y, X_2),

$$d_R(X_1\|X_2|Y) \leq d_R(X_1\|X_2).$$

Our proof of Ruzsa triangle inequality in fact proceeds by proving a refined triangle inequality: If X_i are independent, then

$$d_R(X_1\|X_3) \leq d_R(X_1\|X_2|X_2 - X_3) + d_R(X_2\|X_3).$$

We recover the Ruzsa triangle inequality by using the fact that conditioning reduces Ruzsa divergence.

We now ask a basic question.

Question: If Y and Y' are i.i.d. random variables taking values in G, how different can $h(Y + Y')$ and $h(Y - Y')$ be?

One answer that follows from our two theorems above is that the entropies of the sum and difference of two i.i.d. random variables *are not too different*. More precisely, for any two i.i.d. random variables Y, Y' with finite entropy,

$$\frac{1}{2} \leq \frac{h(Y + Y') - h(Y)}{h(Y - Y') - h(Y)} \leq 2.$$

To prove this, observe that if Y, Y', Z are independent random variables, then the Submodularity Lemma says

$$h(Y + Y' + Z) + h(Z) \leq h(Y + Z) + h(Y' + Z).$$

Since $h(Y + Y') \leq h(Y + Y' + Z)$,

$$h(Y + Y') + h(Z) \leq h(Y + Z) + h(Y' + Z). \quad (2)$$

Also the Ruzsa triangle inequality can be rewritten:

$$h(Y - Y') + h(Z) \leq h(Y - Z) + h(Y' - Z). \quad (3)$$

Taking now Y, Y' to be i.i.d. and Z to be an independent copy of $-Y$ in the inequalities (2) and (3), we get

$$h(Y + Y') + h(Y) \leq 2h(Y - Y'),$$
$$h(Y - Y') + h(Y) \leq 2h(Y + Y'),$$

which are the desired bounds.

Interestingly, for $G = \mathbb{Z}$ or $G = \mathbb{R}$, the entropies of the sum and difference of two i.i.d. random variables *can differ by an arbitrarily large amount*. More precisely, if $G = \mathbb{Z}$ or $G = \mathbb{R}$, given any $M > 0$, it was shown in [28] that there exist i.i.d. random variables Y, Y' of finite entropy such that

$$h(Y - Y') - h(Y + Y') > M,$$

and in [18] that there exist i.i.d. random variables U, U' of finite entropy such that

$$h(Y + Y') - h(Y - Y') > M.$$

These two answers together suggest that the natural quantities to consider are the differences

$$\Delta_+ = h(Y + Y') - h(Y),$$
$$\Delta_- = h(Y - Y') - h(Y).$$

Then our results assert that the *ratio* Δ_+/Δ_- must always lie between $\frac{1}{2}$ and 2, while those of [28, 18] state that the *differences* $\Delta_+ - \Delta_-$ and $\Delta_- - \Delta_+$ can be arbitrarily large. Note that the only way that the differences can be large is if $h(Y)$ itself is large.

There are a number of additional results of interest that can be obtained by developing the properties of the Ruzsa divergence and its conditional cousin. Such a development leads to relatively transparent proofs of general inequalities such as the following:

1. The general sum-difference inequality states that

   $$d_R(X\| - Y) \le 2d_R(X\|Y) + d_R(Y\|X).$$

   In the case where X and Y are i.i.d., we get

   $$d_R(X\| - X) \le 3d_R(X\|X),$$

   while taking X and $-Y$ to have the same distribution gives

   $$d_R(X\|X) \le 3d_R(X\| - X).$$

2. Analogs of the Balog-Szemeredi-Gowers inequality in additive combinatorics can be developed, generalizing that developed for discrete groups by Tao [14] and for $\mathbb{R}$ by the authors [17]. Suppose $X_1 \longleftrightarrow Y \longleftrightarrow X_2$ form a Markov chain. Then

   $$\begin{aligned} d_R(X_1\|X_2|Y) \le\ & 2I(X_1;Y) + I(X_2;Y) \\ & + \tilde{d}_R(X_1\|Y) + \tilde{d}_R(Y\|X_2), \end{aligned}$$

   where $\tilde{d}_R(X\|Y) := h(X - Y) - h(X) = I(X - Y;Y) - I(X;Y)$.

4. REMARKS

While we do not have space to describe the applications of the inequalities developed, several such applications have already been developed. For example, in [25], we develop an entropic analog of the Rogers-Shephard inequality for the difference body of a convex body (cf., [29]), as well as connections to the central limit theorem and stability phenomena for the entropy power inequality.

Some steps have been taken towards an entropy theory of sums of random variables that take values in general abelian groups. For discrete groups, the theory has close connections to and implications for additive combinatorics, while for $\mathbb{R}^n$, the theory has close connections to and implications for probability, convex geometry, and geometric functional analysis. We believe that these results should also have further useful consequences in information and communication theory are waiting to be explored.

5. ACKNOWLEDGMENTS

M. Madiman was supported by NSF grants DMS-1409504 (CAREER) and CCF-1346564. I. Kontoyiannis was supported by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) Research Funding Program "Thales - Investing in knowledge society through the European Social Fund."

6. REFERENCES

[1] Y. Wu, S. Shamai, and S. Verdú, "Information dimension and the degrees of freedom of the interference channel," *IEEE Trans. Inform. Theory*, vol. 61, no. 1, pp. 256–279, 2015.

[2] D. Stotz and H. Bölcskei, "Characterizing degrees of freedom through additive combinatorics," *Preprint,* `arXiv:1506:01866`, 2015.

[3] A.J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, 1959.

[4] A.R. Barron, "The strong ergodic theorem for densities: Generalized Shannon-Mcmillan-Breiman theorem," *Ann. Probab.*, vol. 13, pp. 1292–1303, 1985.

[5] O. Johnson and A.R. Barron, "Fisher information inequalities and the central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 391–409, 2004.

[6] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "On the rate of convergence in the entropic central limit theorem," *Probab. Theory Related Fields*, vol. 129, no. 3, pp. 381–390, 2004.

[7] M. Madiman and A.R. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inform. Theory*, vol. 53, no. 7, pp. 2317–2329, July 2007.

[8] E. A. Carlen and A. Soffer, "Propogation of localization, optimal entropy production, and convergence rates for the central limit theorem," *Preprint,* `arXiv:1106.2256`, 2011.

[9] K. Ball and V. H. Nguyen, "Entropy jumps for isotropic log-concave random vectors and spectral gap," *Studia Math.*, vol. 213, no. 1, pp. 81–96, 2012.

[10] S. G. Bobkov, G. P. Chistyakov, and F. Götze, "Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem," *Ann. Probab.*, vol. 41, no. 4, pp. 2479–2512, 2013.

[11] M. Madiman, "On the entropy of sums," in *Proc. IEEE Inform. Theory Workshop*, pp. 303–307. Porto, Portugal, 2008.

[12] I. Z. Ruzsa, "Entropy and sumsets," *Random Struct. Alg.*, vol. 34, pp. 1–10, 2009.

[13] M. Madiman, A. Marcus, and P. Tetali, "Entropy and set cardinality inequalities for partition-determined functions," *Random Struct. Alg.*, vol. 40, pp. 399–424, 2012.

[14] T. Tao, "Sumset and inverse sumset theory for Shannon entropy," *Combin. Probab. Comput.*, vol. 19, no. 4, pp. 603–639, 2010.

[15] V. Jog and V. Anantharam, "The entropy power inequality and Mrs. Gerber's lemma for groups of order 2^n," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 3773–3786, 2014.

[16] L. Wang, J. O. Woo, and M. Madiman, "A lower bound on the Rényi entropy of convolutions in the integers," in *Proc. IEEE Intl. Symp. Inform. Theory*, pp. 2829–2833. Honolulu, Hawaii, July 2014.

[17] I. Kontoyiannis and M. Madiman, "Sumset and inverse sumset inequalities for differential entropy and mutual information," *IEEE Trans. Inform. Theory*, vol. 60, no. 8, pp. 4503–4514, August 2014.

[18] E. Abbe, J. Li, and M. Madiman, "Entropies of weighted sums in cyclic groups and applications to polar codes," *Preprint*, 2015.

[19] C. Borell, "Convex measures on locally convex spaces," *Ark. Mat.*, vol. 12, pp. 239–252, 1974.

[20] K. Ball, "Logarithmically concave functions and sections of convex sets in $\mathbf{R}^n$," *Studia Math.*, vol. 88, no. 1, pp. 69–84, 1988.

[21] E. Lutwak, D. Yang, and G. Zhang, "Moment-entropy inequalities," *Ann. Probab.*, vol. 32, no. 1B, pp. 757–774, 2004.

[22] B. Klartag and V. D. Milman, "Geometry of log-concave functions and measures," *Geom. Dedicata*, vol. 112, pp. 169–182, 2005.

[23] S. Bobkov and M. Madiman, "Dimensional behaviour of entropy and information," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 349, pp. 201–204, Février 2011.

[24] M. Fradelizi, M. Madiman, and L. Wang, "Optimal concentration of information content for log-concave densities," *Preprint,* `arXiv:1508.04093`, 2015.

[25] M. Madiman and I. Kontoyiannis, "Entropy bounds on abelian groups and the Ruzsa divergence," *Preprint,* `arXiv:1508.04089`, 2015.

[26] M. Madiman and P. Singla, "A note on $GL_n(\mathbb{Z})$-actions on locally compact abelian groups," *Preprint*, 2015.

[27] V. A. Kaĭmanovich and A. M. Vershik, "Random walks on discrete groups: boundary and entropy," *Ann. Probab.*, vol. 11, no. 3, pp. 457–490, 1983.

[28] A. Lapidoth and G. Pete, "On the entropy of the sum and of the difference of two independent random variables," *Proc. IEEEI 2008, Eilat, Israel*, 2008.

[29] S. G. Bobkov and M. M. Madiman, "On the problem of reversibility of the entropy power inequality," in *Limit Theorems in Probability, Statistics, and Number Theory (in honor of Friedrich Götze)*, P. Eichelsbacher et al., Ed., vol. 42 of *Springer Proceedings in Mathematics and Statistics*. Springer-Verlag, 2013, Available online at `http://arxiv.org/abs/1111.6807`.

SCOT MODELING, TRAINING AND STATISTICAL INFERENCE

Mikhail Malyutov[1], *Paul Grosu*[2] *and Tong Zhang*[3]

Math. Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115
[1] m.malioutov@neu.edu, [2] pgrosu@gmail.com, [3] zhang.tong@husky.neu.edu

ABSTRACT

Stochastic COntext Tree (abbreviated as SCOT) is m-Markov Chain (m-MC) with every state of a string independent of the symbols in its more remote past than the **context** of **length** determined by the preceding symbols of this state. We model and apply SCOT for statistical inference about financial, literary and seismological stationary strings in 'Information processes, vol 13, No 4, Vol 14, No. 3 and volume 15, No.1, available online. SCOT construction has been earlier used for compression under various names VLMC, VOMC, PST, CTW. We analyze several models viewed as simplified approaches to financial modeling: evaluate their stationary distribution, entropy rate and convergence to the Brownian motion.

1. Introduction

Modeling random processes as full m-Markov Chains (m-MC) can be inadequate for small m, and over-parametrized for large m. For example, if the cardinality of the base state space is four, $m = 10$, then the number of parameters is around 3,15 millions. The popular Box–Jenkins ARIMA and Engel's GARCH in quality control and finance are not adequate in applications to linguistics, genomics and proteomics, security, etc, where comparatively long *non-isotropic contexts* are relevant that would require huge memory size of the full m-MC. In [10], compressor VLMC was constructed based on consistent statistical estimate of the Stochastic Context Tree (SCOT) of the training string which is then used for compression. SCOT is an m-Markov Chain (MC), where every state is independent of the states which are more remote than the *contexts* of a certain length depending on the preceding m-gram. In most applications, an estimated SCOT turns out to be sparse in agreement with the Occam principle. Instead of compression, we use SCOT for generating the likelihood function of strings, and apply the latter for statistical inference. A substantial part of this paper is devoted to the innovative modeling of SCOT - governed time series. We present theoretical results on SCOT models and its online training algorithm. These results are applied for statistical SCOT - based inference on discrimination between quiet and volatile regions of financial time series, seismological time series, as well as in similar type sequences used in literary research in [5]. Apparently, the first SCOT *Statistical Likelihood* comparison application [1] to *non-stationary Bioinformatics data* is inadequate.

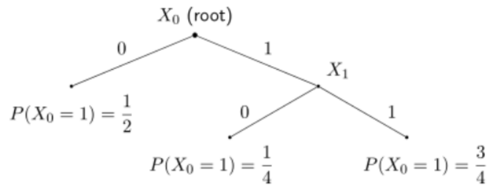

X_0 (root)
0
1
X_1
$P(X_0 = 1) = \frac{1}{2}$
0
1
$P(X_0 = 1) = \frac{1}{4}$
$P(X_0 = 1) = \frac{3}{4}$

Fig. 1. The simplest stochastic context tree (Model 1).

For the simplest SCOT of Fig. 1 contexts are $\{\mathbf{0}\}, \{\mathbf{01}\}, \{\mathbf{11}\}$ written from bottom to top, transition probabilities $P(x_0 = 1)$ given preceding contexts are respectively $1/2, 1/4, 3/4$, as displayed there.

Binary Context Tree K_n repeats n times byfication of the right hand side of Tree K_2 on Fig.1. It has contexts $(0, (01)), ...,$ $(01^{n-1}), (1^n)$ and admits reduction to 1-MC for every n. Let us assign all transition probabilities between consequent contexts as 1/2. This SCOT is ergodic, the **stationary distribution** for this 1-MC is

$(1/2, 1/4, \ldots, (1/2)^{n-1}, (1/2)^n, (1/2)^n)$,

and entropy rate (ER) for SCOT reduced to 1-MC between contexts is by the well-known formula for 1-MC: ER=-$\sum_{i \in A} \sum_{j \in A} q_i p_{ij} \log p_{ij}$ which is $\log 2$ for all n.

- We find: SCOT stationary distribution and ER in several more advanced models,
- show that SCOT ER is much lower than the **maximum** $|A|^n \log(|A|)$ for n-MC.
- prove invariance, asymptotic normality and exponential tails of additive functions of SCOT trajectories.

2. Reduction to 1-MC

An m-MC $\{x_n\}$ with a finite state space (alphabet) A can be regarded as 1-MC

$$\{Y_n = (x_n, x_{n+1}, \ldots, x_{n+m-1})\}$$

with alphabet as the space of *m-grams* A^m. Namely:

$P(Y_{n+1}|Y_n) = P(x_{n+m}|Y_n)$, if $x_{n+1},\dots,x_{n+m-1}$ coincide in both sides, and 0 otherwise.

Sparse SCOT over some alphabet $A = \{a_1,\dots,a_d\}$ is a very special case of m-MC, where m is the **maximal length of contexts**. Given stochastic string $x_{-m},\dots,x_{-1},x_0$, the *context to a current state x_0 given preceding m-gram* is

$$C(x_0) = x_{-1},\dots,x_{-k}, k \le m := x_{-1}^{-k} : \qquad (1)$$

the top part of the preceding m-gram of **minimal length** such that the conditional probability

$$P(x_0|x_{-1}^{-r}) \equiv P(x_0|x_{-1}^{-k}), \forall r > k; k = |C(x_0)| \qquad (2)$$

is called the *length of context* $C(x_0)$, a context tree T is assumed complete, see [6], T^* denotes the set of contexts in T, T_i denotes the subtree T_i of T whose root is a_i, thus $T_i^* := \{u|\ \overline{ua_i} \in T^*\}$; For all pairs $a_i \in A, a_j \in A$, $T_{j,i} := T_i(T_j)$, thus we have $T_{j,i}^* := \{u|\overline{ua_ia_j} \in T^*\}$.

A complete context tree T is called ”tailclosed” if ∀ $c \in T^*, i \in \{1,...,d\}$, ∃ $u \in T^*$, s.t. $\overline{ca_i} = \overline{wu}$, where w is a string.

Theorem (T. Zhang) : Let T be a complete context tree, then the following statements are equivalent.

(a) T is tailclosed.

$(b)\forall\ 1 \le i,j \le d,\ T_{j,i} \subseteq T_i$

$(c)\forall\ 1 \le i,j \le d,\ c \in T^*_{j,i},\ \exists\ c' \in T^*_i,\ s.t.\ c' = \overline{uc}$, where u is a string.

$(d)\forall\ 1 \le i,j \le d,\ c' \in T^*_i,\ \exists\ c \in T^*_{j,i},\ s.t.\ c' = \overline{uc}$, where u is a string.

For tailclosed context tree, then we accept the following:

Definition. *The transition probability from C(x_i) to C(x_{i+1}) is the transition probability from C(x_i) to x_{i+1}.*

Thus, all $|A|$ realizations of x_0 determine the next context ending up with x_0 for predicting x_1 and induce the *transition probability between consecutive contexts* defining a $|\mathscr{C}(n)|\times|\mathscr{C}(n)|$ matrix $\mathbf{P}$ of 1-MC transition probabilities between contexts,

In our first example, this definition gives the following transition probability matrix $\mathbf{P}$ between contexts:

Table 1: Transition probability matrix **P** between contexts (0), (01), (11) in the previous slide

0.5	0.5	0
0.75	0	0.25
0.25	0	0.75

Distribution $(6.1)^{-1}(3, 1.5, 1.6)$ rapidly converges to the stationary one: (1/2, 1/4, 1/4) after iterative multiplications by P.

SCOT reduction to 1-MC is useful in approximating the stationary SCOT distribution: Multiplying empirical estimate of the stationary distribution from the right by powers of matrix $\mathbf{P}$, we approximate the theoretical stationary distribution better.

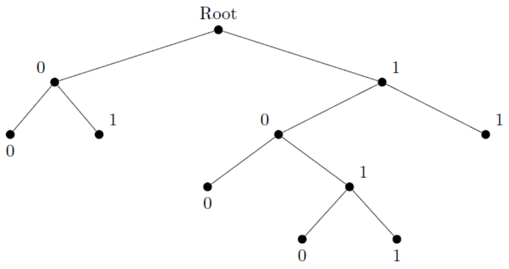


Fig. 2. Counterexample

Namely, the Euclidean norm of the approximation error shrinks exponentially with the power of $\mathbf{P}$. If a context probability shrinks, we remove it.

We define the steady state (stationary) distribution $Q(C, C \in \mathscr{C})$ over all L contexts as the solution to the equation:

$$Q^*\mathbf{P} = Q^*. \qquad (3)$$

If the induced 1-MC is ergodic, then **Ergodic theorem** holds: *The solution to (1) exists, is unique, and iterations $Q_m = Q_0\boldsymbol{P}^m$ converge to Q^* exponentially fast.*

3. Spike model

- Let us: assign randomly the increments of the Random walk to *regular ones* with Probability $1-2/N$, and (with $2/N$ probability) to spikes;
- specify a standard increment distribution for regular increments ±1 and SCOT model with increments of magnitude 0, or $\pm\sqrt{N}$ to spikes;
- convergence of the Spike model to a martingale - *mixture of the Brownian motion and a symmetric pair of ± Poisson processes*.

The family $X_n^N = \sum_{i=1}^n r_i$ of Spike models is a reflected Random Walk on large interval $[-l,l], l > N^{3/2}$. Regular part of X_n^N has increments ±1 and reflects one step from the boundary next time after hitting it. Very rare (probability $2/N$) random interruptions by spikes at random Spike time moments n have magnitudes 0 or $\pm\sqrt{N}$ depending on whether $X_n^N = X_{n-1}^N$, or $X_n^N > X_{n-1}^N$, or the opposite inequality holds.

$X_n^N = \sum_{i=1}^n r_i, r_i$ *in a regular part* is an equally likely sequence of independent identically distributed (IID) $\pm 1, i = 1,\dots,$ inside $(-l,l)$, while *irregular part* is a SCOT model specified above.

4. Continuous time limit

Let the increments of time/space be respectively $1/N, 1/\sqrt{N}$ instead of 1. Introduce $w^N(t) = N^{-1/2}X^N_{\lfloor Nt\rfloor}$ (summation is until the integer part $\lfloor Nt\rfloor$ of Nt). We study the weak convergence of $w^N(t)$ as $N \to \infty$.

Inside $(-l,l)$ *conditionally on no spike at time* $k+1$

$$E(X_{k+1} - X_k) = 0,$$

$$\mathrm{Var}(X_{k+1}^N - X_k^N)) = (1-2/N)/N.$$

Let τ_k be the k-th spike time. Obviously, $\tau_k - \tau_{k-1}$ are IID, independent of σ-algebra spanned by $(x_j, j < k-2)$ converging to the exponential distribution with mean 2.

Theorem. *In the limit we get a weak convergence of $w^N(t)$ to the Wiener process $w(t)$ in between independent of $w(t)$ compound Poisson spikes process of equally likely magnitudes ± 1:*

$$P(\tau > t) = \exp(-t/2),$$

τ and $\{x_t\}, t < \tau$, are independent.

5. 'Thorny' $TH_{a,b}$ SCOT model

Our next model is similar to the Spike model, only rare random time moments of spikes $\pm aN^b$ with similar dependence of spikes magnitude on the past take place with probability $N^{-2b}, 0 < b < 1/8$. In the same limiting situation of time intervals $1/N$ and steps $1/\sqrt{N}$, the KUM criterion is valid with similar parameters, thus trajectories of the limiting $TH_{a,b}$ model are continuous.

Let the martingale sequence $w^N(t)$ be as above. Then $Er_i = 0, Var[w^N(t)] = Nt[(a^2N^{2b-1})N^{-2b} + (1-N^{-2b})] \to a^2+1$. The equality of summands preceding a spike can be neglected. The covariance of $w^N(t)$ converges to that of $\sqrt{(a^2+1)}w(t)$ in a similar way. Thus this model gives larger volatility without noticeable drift in the limit to continuous t. The weak FDD convergence to that of $\sqrt{a^2+1}w(t)$ is valid since the Martingale version of the Lindeberg condition holds. Thus we proved the following statement.

Proposition. *$w^N(t)$ converges weakly to $\sqrt{(a^2+1)}w(t)$.*

6. AN for additive functions of m-MC trajectories

- Given an m-MC $\{x_i\}$ with alphabet A, denote induced 1-MC on m-grams (see our Introduction) as $\{Y_i\}$. If n-MC X_N is ergodic with finite alphabet and the sequence $\{Y_N\}$ is an ergodic 1-MC, then.
- this 1-MC $\{Y_N\}$ is a Harris invariant ([9], chapter 17), with respect to a probability distribution. Let g be a Borel function on $\mathbf{R}$.
- Define $f(Y_i) := f(x_i, x_{i-1}, \dots, x_{i-m+1}) = \sum_{k=0}^{m-1} g(x_{i-k})$.
- Define $\bar{f}_N := (1/N)\sum_{i=1}^N f(Y_i), \bar{g}_N := (1/N)\sum_{i=1}^N g(x_i)$.
- If π is the stationary distribution and $E_\pi|f^2| < \infty$, then the ergodic theorem ([9], 17.3) guarantees that $\bar{f}_N \to E_\pi f$ with probability 1 as $N \to \infty$ and the central limit theorem holds for $\bar{f}_N$ ([9], 17.4):

$\sqrt{N}(\bar{f}_N - E_\pi f) \Rightarrow N(0, f_\pi^2)$ weakly, where $\sigma f_\pi^2 < \infty$ is the variance of f with respect to π.

- $\sqrt{N}(1/N\sigma_{i=1}^N f(\{Y_i\}) - E_\pi f) \Rightarrow N(0, \sigma_\pi f^2)$ weakly,
- $\sqrt{N}(1/N\sigma_{i=1}^N \sigma_{k=0}^{m-1} g(x_{i-k}) - E_\pi f) \Rightarrow N(0, \sigma_\pi f^2)$ weakly,
- $\sqrt{N}(m\bar{g}_N(x) - E_\pi f) \Rightarrow N(0, \sigma_\pi f^2)$ weakly.
- Results on the MC AN convergence rate suggest its increase with lowering the MC alphabet size.
- Thus SCOT AN convergence rate is generally much higher than for the full n-MC.
- Above results justify t-distribution of our homogeneity test statistic based on studentized averages of SCOT log-likelihoods introduced further.

7. Asymptotic expansion for additive functions

[4] proves for finite ergodic MC :

$P(N^{-1/2}(\sum f(x_i) \le x)) = \Phi_\sigma(x) + \Phi_\sigma(x)q(x)N^{-1/2} + O(N^{-1})$ and finds explicit expression for the Hermite polynomial $q(x)$. Here ϕ and Φ are pdf and CDF of the central Normal RV with StD σ.

This result can be generalized for n-MC by the method displayed above.

We believe that the coefficient $q(x)$ for sparse SCOT is substantially less that for general n-MC.

8. Exponential tails for additive functions

Introduce diversion (cross entropy) $D(P_1||P_0) = \mathbf{E}_1 \log(P_1/P_0)$ and consider goodness of fit tests of P_0 vs. P_1 for IID sample of size N.

9. 'Stein' lemma for LRT between two known SCOT distributions [11]

If $D(P_1||P_0) \geq \lambda$ and any $0 < \varepsilon < 1$, then the error probabilities of Likelihood Ratio Test (LRT) satisfy simultaneously

$$P_0(L_0 - L_1 > N\lambda) \leq 2^{-N\lambda}$$

and

$$\lim P_1(L_0 - L_1 > n\lambda)) \geq 1 - \varepsilon > 0.$$

No other test has both error probabilities less in order of magnitude.

10. Nonparametric version of the 'Stein' lemma

Generate an artificial (n)-sequence $\mathbf{z}^N$ independent of $\mathbf{y}^N$, $\mathbf{z}^N$ distributed as P_0 and denote by L_0 its log-likelihood given the SCOT model of the training string.

L is the query log-likelihood given the SCOT model of the training string.

Also assume that the joint distribution of S slices of size n converge to their product distribution in Probability.

Theorem. *Suppose P_1, P_0 are SCOT, $D(P_1||P_0) > \lambda$ and we reject homogeneity, if the 'conditional version of the Likelihood Ratio' test $\mathcal{T} = \overline{L} - \overline{L}_0 > N\lambda$. Then the same error probability asymptotics as for LRT in the 'Stein' lemma is valid for this test.*

11. SCOT training

We develop an parallel SCOT training which removes severe restriction the SCOT alphabet size not to exceed 27 in [3] for applying it in statistical inference such as prediction and testing homogeneity.

11.1. Determining ESI for possible context

The Empirical Shannon Information (ESI) is an approximation to the log-likelihood ratio statistic for testing the consistency of the context of a given source.

Given a context s and let $N(s)$ be the count of s in the source. Define a function $ESI(s)$ as follows:

$$ESI(s) = \sum_{i \in A} \sum_{j \in A} N(i.s.j) \cdot log_2 \left(\frac{N(i.s.j) \cdot N(s)}{N(i.s) \cdot N(s.j)} \right)$$

In our implementation, these values are collected in the following matrix:

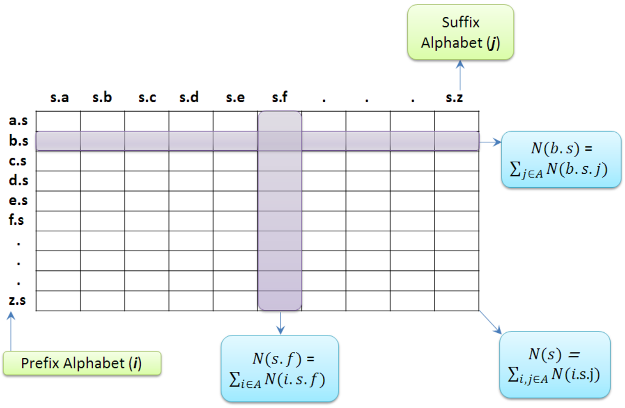

Suffix Alphabet (*j*)
s.a
s.b
s.c
s.d
s.e
s.f
.
.
.
s.z
a.s
b.s
c.s
d.s
e.s
f.s
.
.
.
z.s
$N(b.s) = \sum_{j \in A} N(b.s.j)$
Prefix Alphabet (*i*)
$N(s.f) = \sum_{i \in A} N(i.s.f)$
$N(s) = \sum_{i,j \in A} N(i.s.j)$

Fig. 3. SCOT training

11.2. Deciding about contexts

Using a fixed maximum context length h and a threshold $\varepsilon > 0$, we define a *context* over source as follows:

For any message $'x_1 x_2 \ldots x_t'$, where $t \leq h$. It is decided to be a *context*, if and only if:

(a) For any i, $i = 2, \ldots, t$ such that $ESI('x_1 x_2 \ldots x_t') > \varepsilon$

(b) $0 < ESI('x_1 x_2 \ldots x_t') \leq \varepsilon$

11.3. Building SCOT

Using our criterion on *contexts*, we check all the messages coming from a source and build a SCOT such that each *context* is a path starting from a leaf and ending at a son of the root. A SCOT is built in a step-wise manner starting with depth 1 and ending at the desired *context* length. Below is an example of the SCOT for the message *shannon*:

The stochastic component of SCOT - prediction distribution of symbols in the root - is to be specified at every leaf. This is performed by the following equation, where s is a leaf:

$$P(i|s) = N(s.i)/N(s), \text{where } i \in A.$$

As an example, below are the leaf probabilities generated for the context *sha*: $P(a|sha) = 0, P(h|sha) = 0,$
$P(n|sha) = 1, P(o|sha) = 0, P(s|sha) = 0.$

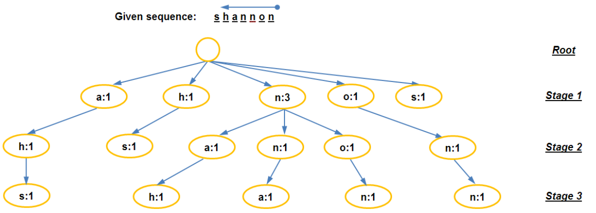


Fig. 3. SCOT training

12. CONCLUSION

Our theoretical study of stationary distributions, limit theorems and asymptotic normality of additive functions for SCOT models prepares a solid base for statistical applications of SCOT models such as described in [5] .

13. Acknowledgements

National Science Foundation, NSF Award OCI 12-29059, “MRI Consortium: Acquisition of a heterogeneous, shared, computing instrument to enable science and computing research by the Mass. Green High performance Computing Consortium”

A. REFERENCES

[1] G. Bejerano, *Automata learning and stochastic modeling for biosequence analysis*. PhD dissertation, Hebrew University, Jerusalem, 2003.

[2] A. Galves and E.. Loecherbach, Stochastic chains with memory of variable length, In: *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, Tampere, TICSP series No. 38, Tampere Tech. Uni., 117–134, 2008.

[3] M. Mächler and P. Bühlmann, Variable Length Markov Chains: methodology, computing, and software, *Journal of Computational and Graphical Statistics*, 2004, Vol. 13, No. 2, 435–455.

[4] V. K. Malinovsky, On limit theorems for Harris Markov chains. I. *Theory Probab. Appl., (English)* 1987, Vol 31, 269–285.

[5] M. Malyutov, T. Zhang, Y. Li , and X. Li, Time series homogeneity tests via VLMC training. *Information Processes*, 2013, Vol. 13, No. 4, 401–414.

[6] M. Malyutov, T. Zhang, and P. Grosu, SCOT stationary distribution evaluation for some examples. *Information Processes*, 2014, Vol. 14, No. 3, 275–283.

[7] M. Malyutov, T. Zhang, Limit theorems for additive functions of SCOT trajectories. *Information Processes*, 2015, Vol. 15, No. 1, 89–96.

[8] M. Malyutov, P. Grosu and H.Sadaka, Separate testing of inputs versus linear programming relaxation, *Information Processes*, 2015, Vol. 15, No. 3, to appear.

[9] S.P. Meyn and R.L. Tweedy, *Markov chains and stochastic stability*. Springer, 1993.

[10] J. Rissanen, A universal data compression system, *IEEE Trans. Inform. Theory*, **29:5**,1983, 656–664.

[11] J. Ziv On classification and universal data compression, *IEEE Trans. on Inform. Th.*, **34:2**, 278-286, 1988.

RELEVANCE SAMPLING

Mads Nielsen[1], *Bo Markussen*[2] *and Marco Loog*[3]

[1]Department of Computer Science, University of Copenhagen,
Sigurdsgade 41, 2200 Copenhagen N, DENMARK, madsn@di.ku.dk
[2] Department of Mathematical Sciences, University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen OE, DENMARK, bomar@math.ku.dk
[3] Pattern Recognition Laboratory, Delft University of Technology,
Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

In this paper we suggest an instance of the *Information Bottleneck Method (IBM)* as an information theoretic alternative of the *Theory of Visual Attention (TVA)*. The proposed method is called *Relevance Sampling (RS)* since it can be interpreted in the spirit of *Importance Sampling (IS)*. The aim of RS is optimally to sample a distribution of a stochastic variable X in order to learn as fast as possible about a related variable Y.

1. INTRODUCTION

Our motivation for writing this paper is to understand biological information processing theories like TVA [1] in terms of information theory. TVA describes the human visual system via two steps, namely *filtering* and *pigeonholing*. The purpose of filtering is to throw away visual input that is considered irrelevant for the current task that the visual system attends to.

We study *filtering* by combining IS [2, 3] and IBM [4, 5]. IS creates optimal strategies of how to sample a distribution to minimize the variance on an integral estimator (such as e.g. the mean or any other moment). The IBM provides a strategy of how to quantize a stochastic variable X into $\tilde{X}$ so as to preserve as much information about a task variable Y as possible. This paper provides a combination of the two creating a sampling strategy of X revealing as much information on Y as possible.

2. NOTATION AND PROBLEM STATEMENT

Let S, X, Y be random variables defined on a probability space $(\Omega_S \times \Omega_X \times \Omega_Y, F_S \otimes F_X \otimes F_Y, P)$. The random variable S represents the *state of the world*, and X is a random variable from which we obtain evidence about the state of the world S. Our task is to determine the state of Y, denoted the *task* variable, based on the observation of X. We assume to have the following Markov property:

$$X \to S \to Y.$$

Now let $s \in \Omega_S$ be fixed. We may sample as many independent measurements X_i from the conditional distribution of X given $S = s$ as we wish, and the aim is to use *filtering* to obtain information on Y as quickly as possible. Here *filtering* is interpreted in the following way; if an observation x_i is considered less relevant for deciding the state of the task variable, then we may choose not to use this observation.

3. RELEVANCE SAMPLING

We introduce an additional state $\nabla \notin \Omega_X$ to model the situation that an observation is thrown away. The probability of retaining the observation $x \in \Omega_X$ is denoted by $M(x) \in [0, 1]$. If $\tilde{X}$ denotes the outcome of this procedure, then we have the Markov diagram

$$\tilde{X} \to X \to Y$$

with $\Omega_{\tilde{X}} = \Omega_X \cup \{\nabla\}$ and transition probabilities from X to $\tilde{X}$ given by

$$p(\tilde{x}|x) = \begin{cases} M(x) & \text{for } \tilde{x} = x, \\ 1 - M(x) & \text{for } \tilde{x} = \nabla, \\ 0 & \text{otherwise.} \end{cases}$$

The probability of *not seeing anything* is given by $p(\nabla) = 1 - \int_{\Omega_X} M(x)p(x)\,\mathrm{d}x$, and the sampling density on the original sample space is given by

$$\tilde{p}(x) = P(\tilde{X} = x|\tilde{X} \neq \nabla) = \frac{M(x)p(x)}{1 - p(\nabla)}.$$

This may be interpreted as importance sampling [2, 3] with sampling measure

$$L(x) = \frac{M(x)}{1 - p(\nabla)}.$$

The conditional density of Y given $\tilde{X} = x$ equals $p(y|x)$, i.e. the same as the conditional density of Y given $X = x$, and the conditional density of Y given $\tilde{X} = \nabla$ is given by

$$\begin{aligned} p(y|\nabla) &= \int_{\Omega_X} p(y|x)p(x|\nabla)\,\mathrm{d}x \\ &= \frac{p(y)}{p(\nabla)} - \frac{\int_{\Omega_X} p(y|x)M(x)p(x)\,\mathrm{d}x}{p(\nabla)}. \end{aligned}$$

If observations are sampled from the density $\tilde{p}(x)$ and the transition probabilities to the task variable is given by $p(y|x)$, then the sampling density of tasks equals

$$\int_{\Omega_X} p(y|x)\frac{M(x)p(x)}{1-p(\nabla)}\,\mathrm{d}x = \frac{1-p(\nabla)\frac{p(y|\nabla)}{p(y)}}{1-p(\nabla)}p(y).$$

This density equals $p(y)$, i.e. provides unbiased inference for the task variable, if and only if $p(y) = p(y|\nabla)$ for every $y \in \Omega_Y$.

In order to choose the importance measure $M(x)$ we use the IBM [4, 5]. Thus, given a parameter $\beta > 0$ quantifying the trade-off between throwing away the most irrelevant samples (*filtering*) while retaining information (*pigeonholing*) the optimal importance measure is given by

$$M^*(x) = \underset{M(x)\in[0,1]}{\arg\min}\; I(\tilde{X}, X) - \beta I(\tilde{X}, Y). \qquad (1)$$

The solution of the optimization problem Eq. (1) is described in the following theorem, where the information gain $R_{Y|X}(x)$ is defined by

$$\begin{aligned} R_{Y|X}(x) &= H(Y) - H(Y|X=x) \\ &= \int_{\Omega_Y} p(y|x)\log\frac{p(y|x)}{p(y)}\,\mathrm{d}y. \end{aligned}$$

We think of the information gain $R_{Y|X}(x)$ as a measure of the *relevance* of the observation $X = x$ for deciding the task Y.

Theorem 1. *The optimal importance measure $M(x)$ and the associated probability $p(\nabla)$ satisfy*

$$\begin{aligned} M(x) &= 1 - \min\left\{1, \frac{p(\nabla)}{p(x)}e^{-\beta R_{Y|X}(x)}\right\}, \\ p(\nabla) &= \int_{\Omega_X} \min\left\{p(x), p(\nabla)e^{-\beta R_{Y|X}(x)}\right\} dx. \end{aligned}$$

The self consistency equation for the probability $p(\nabla)$ can be found by the iteration $p(\nabla) = \lim_{n\to\infty} p_n(\nabla)$ with

$$p_{n+1}(\nabla) = \int_{\Omega_X} \min\left\{p(x), p_n(\nabla)e^{-\beta R_{Y|X}(x)}\right\} dx$$

and $p_0(\nabla) = 1$. Especially, for $\beta \leq \inf_{x\in\Omega_X} \frac{-\log p(x)}{R_{Y|X}(x)}$ we have $p(\nabla) = 1$ and every observation is ignored.

Proof. Let the energy functional $\mathcal{L}$ be defined by

$$\mathcal{L} = I(\tilde{X}, X) - \beta I(\tilde{X}, Y).$$

Some simple algebraic manipulations give that $\mathcal{L}$ equals

$$\begin{aligned} &H(X) - \beta I(X,Y) - p(\nabla)\log p(\nabla) \\ &+ \int_{\Omega_X} p(\nabla, x)\big(\log p(\nabla, x) + \beta R_{Y|X}(x)\big)\,\mathrm{d}x \end{aligned}$$

with $p(\nabla, x) = p(x) - M(x)p(x)$. Using $\frac{\delta p(\nabla)}{\delta p(\nabla,x)} = 1$ we find

$$\frac{\delta\mathcal{L}}{\delta p(\nabla, x)} = \log p(\nabla, x) - \log p(\nabla) + \beta R_{Y|X}(x).$$

The functional $\mathcal{L}$ is convex in distributions of $(\tilde{X}, X)$ with fixed marginal distribution of X, and hence also convex in $p(\nabla, x)$. By itself the stationarity condition $\frac{\delta\mathcal{L}}{\delta p(\nabla,x)} = 0$ implies $p(\nabla, x)$ to be given by $p(\nabla)\mathrm{e}^{-\beta R_{Y|X}(x)}$. But the constraints $0 \leq p(\nabla, x) \leq p(x)$ should also be incorporated. For $p(\nabla, x) \to 0$ we have $\frac{\delta\mathcal{L}}{\delta p(\nabla,x)} \to -\infty$, i.e. the lower bound on $p(\nabla, x)$ poses no constraint. The upper bound on $p(\nabla, x)$ is enforced by the minimum operation in the formula for $M(x)$. The equation for $p(\nabla)$ follows by

$$\begin{aligned} p(\nabla) &= \int_{\Omega_X} p(\nabla, x)\,\mathrm{d}x = \int_{\Omega_X} \big(p(x) - M(x)p(x)\big)\,\mathrm{d}y \\ &= 1 - \int_{\Omega_X} p(x)\left(1 - \min\left\{1, \frac{p(\nabla)}{p(x)}\mathrm{e}^{-\beta R_{Y|X}(x)}\right\}\right)\mathrm{d}x \\ &= \int_{\Omega_X} \min\left\{p(x), p(\nabla)\mathrm{e}^{-\beta R_{Y|X}(x)}\right\}\mathrm{d}x \overset{\text{def}}{=} F\big(p(\nabla)\big). \end{aligned}$$

If the function $F(q)$ is defined by the latter display, then $p(\nabla) \leq F(q) < q$ for $p(\nabla) < q$. It follows that $p(\nabla)$ can be found by the stated iteration. □

3.1. Relevance sampling and features

In many situations it may be intractable to estimate the relevance of the entire observation X. Thus, assume that the relevance only may be based on some feature of the observation, i.e. that there exists another random variable Q defined on (Ω_Q, F_Q, P) such that the following Markov property holds:

$$Q \to X \to Y.$$

If the bivariate variable (Q, X) is considered as the observation, then we are back to the situation studied above. But now we assume that the probability of retaining the observation $(Q, X) = (q, x)$ only may depend on the feature q, i.e. we assume

$$M(q, x) = M_Q(q), \quad p\big(\nabla, (q, x)\big) = p(\nabla, q)\,p(x|q).$$

The proof of the following theorem is similar to that of Theorem 1.

Theorem 2. *The optimal importance measure $M_Q(q)$ is given by*

$$1 - \min\left\{1, \frac{p_Q(\nabla)}{p(q)}e^{-\beta R_{Y|Q}(q) - \beta I(X,Y|Q=q) + H(X|Q=q)}\right\}. \qquad (2)$$

Here the probability $p_Q(\nabla)$ can be found by the iteration $p_Q(\nabla) = \lim_{n\to\infty} p_n(\nabla)$, where $p_0(\nabla) = 1$ and $p_{n+1}(\nabla)$ is given by

$$\int_{\Omega_Q} \min\left\{p(q), p_n(\nabla)e^{-\beta R_{Y|Q}(q) - \beta I(X,Y|Q=q) + H(X|Q=q)}\right\} dq.$$

The interpretation of Eq. (2) is that an observation is more likely to be retained if the mutual information between X and Y given the feature $Q = q$ is large, and less likely to be retained if the conditional entropy $H(X|Q = q)$ of the full observation given the feature is large.

4. APPLICATION TO DECISION THEORY

RS can be applied to the problem of deciding a task Y given i.id. measurements $X_1, \dots, X_n$. The log likelihood ratio test statistic for $Y = y_0$ against $Y \neq y_0$ is given by

$$T(y_0) = \inf_{y \in \Omega_Y \setminus \{y_0\}} \sum_{i=1}^{n} \log \frac{p(y_0|X_i)}{p(y|X_i)}.$$

To apply RS the observation $X = x$ is counted with multiplicity given by the sampling measure

$$L(x) = \frac{\tilde{p}(x)}{p(x)} = \frac{M(x)}{1 - p(\nabla)}$$

and otherwise the inference proceed as usual. The relevance weighted log likelihood ratio test statistic T_{RS} and the associated number N_{RS} of counted measurements are given by

$$T_{\text{RS}}(y_0) = \inf_{y \neq y_0} \sum_{i=1}^{n} L(X_i) \log \frac{p(y_0|X_i)}{p(y|X_i)},$$
$$N_{\text{RS}} = \sum_{i=1}^{n} L(X_i).$$

The slope of the weighted log likelihood ratio as a function of the number of counted measurements equals the slope of the ordinary log likelihood ratio, i.e.

$$\frac{L(X_i) \log \frac{p(y_0|X_i)}{p(y|X_i)}}{L(X_i)} = \log \frac{p(y_0|X_i)}{p(y|X_i)}.$$

Furthermore, the mean number of counted measurements equals the actual number of measurements, i.e.

$$\int_{\Omega_X} L(x) p(x) \, \mathrm{d}x = \int_{\Omega_X} \frac{M(x) p(x)}{1 - p(\nabla)} \, \mathrm{d}x = 1.$$

Thus, by taking steps of length $L(X_i)$ the relevance sampling gives a non uniform weighting of the measurements according to their relevance for the task variable. Especially, the log likelihood ratio is sampled with the density $L(x)p(x|y_0)$ under the null-hypothesis that y_0 is the true task.

This methodology can be extended to the case where the observation $X_1, \dots, X_n$ are weighted according to associated features $Q_1, \dots, Q_n$. The feature weighted log likelihood ratio test statistic with sampling measure $L_Q(q) = \frac{M_Q(q)}{1 - p_Q(\nabla)}$ is given by

$$T_Q(y_0) = \inf_{y \neq y_0} \sum_{i=1}^{n} L_Q(Q_i) \log \frac{p(y_0|X_i)}{p(y|X_i)},$$
$$N_Q = \sum_{i=1}^{n} L_Q(Q_i).$$

4.1. Example: Mean shift in a Gaussian distribution

Let $\Omega_Q = \{0, 1\}$, $\Omega_X = \mathbb{R}$, $\Omega_Y = \{-1, 1\}$ and

$$p(x, y) = \frac{\mathrm{e}^{-\frac{1}{2}(x-y)^2}}{2\sqrt{2\pi}}, \qquad Q_i = 1_{\{|X_i| > 1\}}.$$

Thus, Y is uniformly distributed on the two point set Ω_Y and the conditional distribution of X_i given $Y = y$ is $\mathcal{N}(y, 1)$. The feature Q_i states whether the numerical value of X_i is large. Intuition says that larger numerical values of X_i are more relevant for deciding whether $Y = -1$ or $Y = 1$. Before seeing what the developed theory says about this we remark that $p(y|\nabla) = p(y)$ by symmetry, i.e. relevance sampled inference is unbiased.

The conditional probabilities are given by $p(y|q) = \frac{1}{2}$ and

$$p(y|x) = \frac{\mathrm{e}^{-\frac{1}{2}(x-y)^2}}{\mathrm{e}^{-\frac{1}{2}(x+1)^2} + \mathrm{e}^{-\frac{1}{2}(x-1)^2}}.$$

The relevance $R_{Y|X}(x) = \log(2) + \sum_{y \in \Omega_Y} p(y|x) \log p(y|x)$ is depicted in Fig. 1 and we have $R_{Y|Q}(q) = 0$. Furthermore, the conditional entropies are given by

$$H(X|Q=0) = 0.6929, \quad I(X,Y|Q=0) = 0.1268,$$
$$H(X|Q=1) = 1.4020, \quad I(X,Y|Q=1) = 0.5284.$$

Using $\beta = 5$ and the quantities stated above we find

$$p(\nabla) = 0.6318, \qquad p_Q(\nabla) = 0.6714$$

and the densities $p(x)$, $\tilde{p}(x)$, $\tilde{p}_Q(x)$ and sampling measures $L(x)$, $L_Q(x)$ depicted in Fig. 2 and Fig. 3.

To test relevance sampling against ordinary likelihood ratio we chose $y_0 = 1$ and let $X_1, X_2, \dots$ be i.id. samples from $p(x|y_0)$. The likelihood ratio test statistic for y_0 given the measurements $X_1, \dots, X_n$ is given by

$$T = \sum_{i=1}^{n} \log \frac{p(Y=1|X_i)}{p(Y=-1|X_i)}.$$

The corresponding relevance sampled and feature based relevance sampled likelihood ratio test statistic are given by

$$T_{\text{RS}} = \sum_{i=1}^{n} L(X_i) \log \frac{p(Y=1|X_i)}{p(Y=-1|X_i)}, \qquad N_{\text{RS}} = \sum_{i=1}^{n} L(X_i),$$
$$T_Q = \sum_{i=1}^{n} L_Q(Q_i) \log \frac{p(Y=1|X_i)}{p(Y=-1|X_i)}, \qquad N_Q = \sum_{i=1}^{n} L_Q(Q_i).$$

The statistics T, T_{RS} and T_Q are depicted in Fig. 4. We see that the relevance weighted likelihood ratio test statistics increase more rapidly than the ordinary likelihood ratio test statistic. This property can be quantified. Under the null-hypothesis $Y = 1$ the central limit theorem and the Delta method gives the asymptotic distributions

$$\frac{1}{n} T \sim \text{as}\mathcal{N}\Big(2, \frac{4}{n}\Big),$$
$$\frac{1}{n}\begin{pmatrix} T_{\text{RS}} \\ N_{\text{RS}} \end{pmatrix} \sim \text{as}\mathcal{N}\Big(\begin{pmatrix} 2.9921 \\ 1 \end{pmatrix}, \frac{1}{n}\begin{pmatrix} 12.1234 & 2.4244 \\ 2.4244 & 0.7464 \end{pmatrix}\Big),$$
$$\frac{1}{n}\begin{pmatrix} T_Q \\ N_Q \end{pmatrix} \sim \text{as}\mathcal{N}\Big(\begin{pmatrix} 3.3197 \\ 1 \end{pmatrix}, \frac{1}{n}\begin{pmatrix} 15.9619 & 3.0308 \\ 3.0308 & 0.9121 \end{pmatrix}\Big)$$

and

$$\frac{T_{\text{RS}}}{N_{\text{RS}}} \sim \text{as}\mathcal{N}\Big(2.9921, \frac{4.2979}{n}\Big),$$
$$\frac{T_Q}{N_Q} \sim \text{as}\mathcal{N}\Big(3.3197, \frac{4.0124}{n}\Big).$$

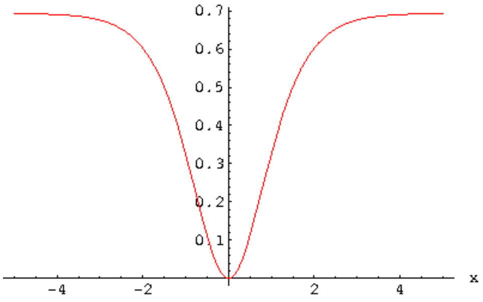


Figure 1. The relevance function $R_{Y|X}(x)$.

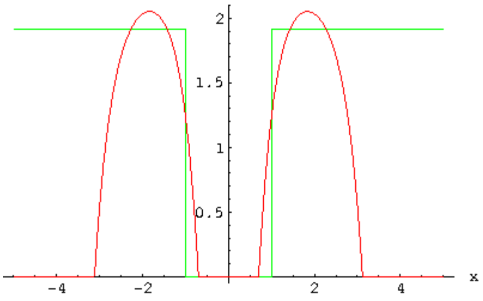


Figure 3. Sampling measures: $L(x) = \frac{M(x)}{1-p(\nabla)}$ (red) and $L_Q(q) = \frac{M_Q(q)}{1-p_Q(\nabla)}$ taken at $q = 1_{\{|x|>1\}}$ (green).

Apparently the relevance weighted likelihood ratio test statistic give stronger evidence for the true null-hypotheses $Y = 1$ than the ordinary likelihood ratio test statistic, even more so for the feature based method. This paradoxically contradicts the fact that the ordinary likelihood ratio statistic uses the most information. However, the relevance sampling methods indeed only use part of the observations. This is helpful if the number of available observations is so huge that the whole is difficult to analyse and comprehend.

5. REFERENCES

[1] Claus Bundesen, “A theory of visual attention,” *Psycological Review*, vol. 97, no. 4, pp. 523–547, 1990.

[2] A Marshall, “The use of multi-stage sampling schemes in monte carlo computations,” in *Symposion on Monte Carlo Methods*, M Meyer, Ed. 1956, pp. 123–140, Wiley, New York.

[3] Peter W Glynn and Donald L Iglehart, “Importance sampling for stochastic simulations,” *Management Science*, vol. 35, no. 11, pp. 1367–1392, 1989.

[4] H S Witsenhausen and A D Wyner, “A conditional entropy bound for a pair of discrete random variables,” *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 495–501, 1975.

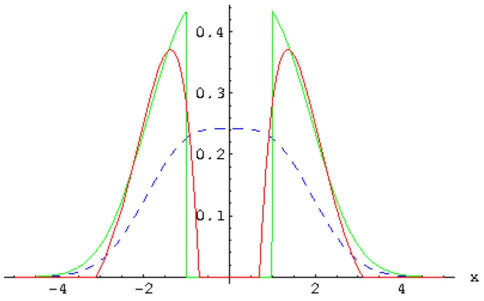


Figure 2. Sampling densities: $p(x)$ (blue, dashed), $\tilde{p}(x) = L(x)p(x)$ (red) and $\tilde{p}_Q(x) = L_Q(x)p(x)$ (green).

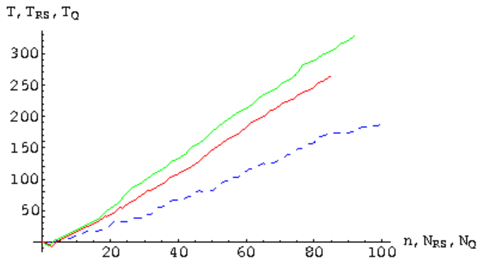


Figure 4. Likelihood ratio statistics: Ordinary (blue,dashed), relevance sampling (red) and feature based relevance sampling (green).

[5] N Tischby, F Pereira, and W Bialek, “The information bottleneck method,” in *Proc. of 37th Allerton Conference on Communication and Computation*, 1999.

THE ONE-TIME PAD IS ROBUST TO SMALL DEVIATIONS FROM RANDOMNESS

Boris Ryabko[1,2] *and Nadezhda Savina*[1]

[1]Institute of Computational Technologies of SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
boris@ryabko.net savina_nn@mail.ru

ABSTRACT

The one-time pad(or Vernam cipher) has played an important role in cryptography because it is a perfect secrecy system. For example, if an English text (presented in binary system) $X_1X_2...$ is enciphered according to the formula $Z_i = (X_i + Y_i) \mod 2$, where $Y_1Y_2...$ is a key sequence generated by the Bernoulli source with equal probabilities of 0 and 1, anyone who knows $Z_1Z_2...$ has no information about $X_1X_2...$ without the knowledge of the key $Y_1Y_2....$ (The best strategy is to guess $X_1X_2...$ not paying attention to $Z_1Z_2....$)

But what should one say about secrecy of an analogous method where the key sequence $Y_1Y_2...$ is generated by the Bernoulli source with a small bias, say, $P(0) = 0.49$, $P(1) = 0.51$? To the best of our knowledge, there are no theoretical estimates for the secrecy of such a system, as well as for the general case where $X_1X_2X_3...$ (the plaintext) and key sequence are described by stationary ergodic processes. We consider the running-key ciphers where the plaintext and the key are generated by stationary ergodic sources and show how to estimate the secrecy of such systems. In particular, it is shown that the Vernam cipher is robust to small deviations from randomness.

1. INTRODUCTION

We consider the classical problem of transmitting secret messages from Alice (a sender) to Bob (a receiver) via an open channel which can be accessed by Eve (an adversary). It is supposed that Alice and Bob (and nobody else) know a so-called key K which is a word in a certain alphabet. Before transmitting a message Alice encrypts it. In his turn, Bob, after having received the encrypted message (ciphertext), decrypts it to recover the initial text (plaintext); see, for ex., [3].

We consider so-called running-key ciphers where the plaintext $X_1...X_t$, the key sequence $Y_1...Y_t$ and ciphertext $Z_1...Z_t$ belong to one alphabet A (without loss of generality we suppose that $A = \{0, 1, ..., n-1\}$, where $n \geq 2$). The $i-th$ letter of the ciphertext is defined by $Z_i = c(X_i, Y_i),\ i = 1, ..., t$, whereas the deciphering rule is by $X_i = d(Z_i, Y_i),\ i = 1, ..., t$, i.e. $d(e(X_i, Y_i), Y_i) = X_i$. Here c and d are functions called coder and decoder, correspondingly. Quite often the following particular formula are used

$$Z_i = (X_i + Y_i) \mod n\,, \quad X_i = (Z_i - Y_i) \mod n\,, \tag{1}$$

i.e. $c(X_i, Y_i) = (X_i + Y_i) \mod n$, $d(Z_i, Y_i) = (Z_i - Y_i) \mod n$. In a case of two-letter alphabet (1) can be presented as follows:

$$Z_i = (X_i \oplus Y_i)\,,\ X_i = (Z_i \oplus Y_i) \tag{2}$$

where $a \oplus b = (a + b) \mod 2$.

The running-key cipher (1) is called the one-time pad (or Vernam cipher) if any word $k_1...k_t$, $k_i \in A$, is used as the key word with probability n^{-t}, i.e. $P(Y_1...Y_t = k_1...k_t) = n^{-t}$ for any $k_1...k_t \in A^t$. In other words, we can say that the key letters are independent and identically distributed (i.i.d.) and probabilities of all letters are equal.

The one-time pad has played an important role in cryptography, especially since C.Shannon proved that this cipher is perfectly secure [5]. If the plaintext is generated by a stationary ergodic source, this property can be interpreted as follows. According to the Shannon-McMillan-Breiman theorem, the set of all sequences $X_1...X_t$ for large t can be represented as two following subsets. The first subset contains $2^{h(X_1...X_t)}$ sequences whose probabilities are close and their sum is almost 1. The second one contains all other sequences whose total probability is almost 0. So, Eva knows that, with overwhelming probability, the ciphered text belongs to the first subset, whose sequences have close probabilities. Moreover, the number of such sequences grows exponentially (as 2^{ht}, where h is the entropy of the plaintext source). That is why Eva cannot find the ciphered text.

In this paper we consider the running-key ciphers (1) in the case where the plaintext $X_1...X_t$ and the key sequence $Y_1...Y_t$ are independently generated by stationary ergodic sources, and the entropy of the key can be smaller than the maximum of $\log n$ per letter (here and below $\log \equiv \log_2$). (In particular, if the entropy is close to $\log n$, we can say that the cipher is close to the one-time pad.) It will be shown that, in a certain sense, if a cipher is close to the one-time pad, their cryptographic security is also close.

It is worth noting that Shannon in his famous paper [5] mentioned that the problem of deciphering of a ciphertext and the problem of signal denoising are very close from mathematical point of view. In this paper we use some results obtained in [4] considering the problem of denoising.

2. PRELIMINARIES

We consider the case where the plaintext $X = X_1, X_2, \ldots$ and the key sequence $Y_1, Y_2, \ldots$ are independently generated by stationary ergodic processes with the finite alphabets $A = \{0, 1, ..., n-1\}$, $n \geq 2$.

The $m-$order Shannon entropy and the limit Shannon entropy are defined as follows:

$$h_m(X) = -\frac{1}{m+1} \sum_{u \in A^{m+1}} P_X(u) \log P_X(u),$$

$$h(X) = \lim_{m\to\infty} h_m(X) \qquad (3)$$

where $m \geq 0$, $P_X(u)$ is the probability that $X_1X_2...X_{|u|}$ $= u$ (this limit always exists, see, for ex., [1, 2]). Introduce also the conditional Shannon entropy

$$h_m(X|Z) = h_m(X,Z) - h_m(Z),\ h(X|Z) = \lim_{m\to\infty} h_m(X|Z) \qquad (4)$$

The Shannon-McMillan-Breiman theorem for conditional entropies can be stated as follows.

Theorem 1 (Shannon-McMillan-Breiman). $\forall \varepsilon > 0, \forall \delta > 0$, *for almost all* $Z_1, Z_2, \ldots$ *there exists* n' *such that if* $n > n'$ *then*

$$P\left\{\left|-\frac{1}{n}\log P(X_1..X_n|Z_1..Z_n) - h(X|Z)\right| < \varepsilon\right\} \geq 1-\delta, \qquad (5)$$

where $P(X_1..X_n|Z_1..Z_n)$ *is a conditional probability.*

The proof can be found in [1, 2].

3. ESTIMATIONS OF SECRECY

Theorem 2. *Let a plaintext* $X = X_1X_2, \ldots$ *and the key sequence* $Y = Y_1Y_2, \ldots$ *be independent with a finite alphabet* $A = \{0, 1, ..., n-1\}$, $n \geq 2$, *and* (X, Y) *be a two-dimensional stationary ergodic process. Let a running-key cipher be applied to* X *and* Y *and* $Z = Z_1, Z_2, \ldots$ *be the ciphertext. Then, for any* $\varepsilon > 0$ *and* $\delta > 0$ *there is such an integer* n' *that, with probability 1, for any* $t > n'$ *and* $Z = Z_1, Z_2, \ldots Z_t$ *there exists the set* $\Psi(Z)$ *for which the following properties are valid:*

i) $P(\Psi(Z)) > 1 - \delta$

ii) for any $X^1 = X^1_1, \ldots, X^1_t$, $X^2 = X^2_1, \ldots, X^2_t$ *from* $\Psi(Z)$

$$\frac{1}{t}\left|\log P(X^1|Z) - \log P(X^2|Z)\right| < \varepsilon$$

iii) $\liminf_{t\to\infty} \frac{1}{t} \log |\Psi(Z)| \geq h(X|Z)$.

Proof. According to Shannon-McMillan-Breiman theorem for any $\varepsilon > 0, \delta > 0$ and almost all $Z_1, Z_2, \ldots$ there exists such n' that for $t > n'$

$$P\left\{\left|-\frac{1}{t}\log P(X_1X_2...X_t|Z_1Z_2...Z_t) - h(X|Z)\right| < \varepsilon/2\right\}$$

$$\geq 1 - \delta. \qquad (6)$$

Let us define

$$\Psi(Z) = \{X = X_1X_2...X_t :$$

$$|P(X_1...X_t|Z_1...Z_t) - h(X|Z)| < \varepsilon/2\}\,. \qquad (7)$$

The first property i) immediately follows from (6). In order to prove ii), note that for any $X^1 = X^1_1, \ldots, X^1_t$, $X^2 = X^2_1, \ldots, X^2_t$ from $\Psi(Z)$ we obtain from (6), (7)

$$\frac{1}{t}\left|\log P(X^1|Z) - \log P(X^2|Z)\right| \leq$$

$$\frac{1}{t}\left|\log P(X^1|Z) - h(X|Z)\right| +$$

$$\frac{1}{t}\left|\log P(X^2|Z) - h(X|Z)\right| < \varepsilon/2 + \varepsilon/2 = \varepsilon\,.$$

From (7) and the property i) we obtain the following: $|\Psi(Z)| > (1-\delta)2^{t\,(h(X|Z)-\varepsilon)}$. Taking into account that it is valid for any $\varepsilon > 0, \delta > 0$ and $t > n'$, we obtain iii). □

So, we can see that the set of possible decipherings $\Psi(Z)$ grows exponentially, its total probability is close to 1 and probabilities of words from this set are close to each other.

Theorem 2 gives a possibility to estimate an uncertainty of a cipher based on the conditional entropy $h(X|Z)$. Sometimes it can be difficult to calculate this value because it requires knowledge of the conditional probabilities. In this case the following simpler estimate can be useful.

Corollary 1. *For almost all* $Z_1Z_2...$

$$\liminf_{t\to\infty} \frac{1}{t} \log |\Psi(Z)| \geq h(X) + h(Y) - \log n\,.$$

Proof. From the well-known in Information Theory equation $h(X, Z) = h(X) + h(Z|X)$ (see [1, 2]) we obtain the following:

$$h(X|Z) = h(X, Z) - h(Z) = h(Z|X) + h(X) - h(Z).$$

Having taken into account that $\max h(Z) = \log n$ ([1, 2]), where n is the number of alphabet letters, we can derive from the latest equation that $h(X|Z) \geq h(Z|X) + h(X) - \log n$. The definition of the running-key cipher (1) shows that $h(Z|X) = h(Y)$. Taking into account two latest inequalities and the third statement iii) of Theorem 2 we obtain the statement of the corollary. □

Comment. In Information Theory the difference between maximal value of the entropy and real one quite often is called the redundancy. Hence, from the corollary we have new following presentations for the value $\frac{1}{t}\log|\Psi(Z)|$:

$$\liminf_{t\to\infty} \frac{1}{t} \log |\Psi(Z)| \geq h(X) - r_Y\,,$$

$$\liminf_{t\to\infty} \frac{1}{t} \log |\Psi(Z)| \geq h(Y) - r_X\,,$$

$$\liminf_{t\to\infty} \frac{1}{t} \log |\Psi(Z)| \geq \log n - (r_X + r_Y) \,, \qquad (8)$$

where $r_Y = \log n - h(Y)$ and $r_X = \log n - h(X)$ are the corresponding redundancies.

Those inequalities give a quantitative assessment of the well-known in cryptography and Information Theory observation that reduction of the redundancy improves the safety of ciphers.

Let us return to the first question of this note about the one-time pad with a biased key sequence. More precisely, let there be a plaintext $X_1X_2...$, $X_i \in \{0,1\}$ and the key sequence $Y_1Y_2...$, $Y_i \in \{0,1\}$, generated by a source whose entropy $h(Y)$ is less then 1. ($h(Y) = 1$ if and only if $Y_1Y_2...$ generated by the Bernoulli source with letter probabilities $P(0) = P(1) = 0.5$, [1, 2]). From (8) we can see that the size of the set $\Psi(Z)$ of high-probable possible decipherings grows exponentially with exponent grater than $h(X) - r_Y$, where $r_Y = 1 - h(Y)$. So, if r_Y goes to 0, the size of the set of possible probable decipherings trends to the size of this set for the case of the one-time pad. Indeed, if $h(Y) = 1$ and, hence, $r_Y = 0$, the set $\Psi(Z)$ of high-probable possible decipherings grows exponentially with exponent $h(X)$, as it should be for the one-time pad. For example, it is true for the case where the key sequence $Y_1Y_2...$ is generated by the Bernulli source with biased probabilities, say $P(0) = 0.5-\tau$, $P(1) = 0.5+\tau$, where τ is a small number. If τ goes to 0, the redundancy r_Y goes to 0, too, and we obtain the one-time pad. So, we can say that the one-time pad is robust to small deviations from randomness.

4. REFERENCES

[1] T. M. Cover and J. A. Thomas. Elements of information theory, Wiley-Interscience, New York, NY, USA, 2006.

[2] R. G. Gallager. Information Theory and Reliable Communication, John Wiley & Sons, New York, 1968.

[3] B. Ryabko, A. Fionov. Basics of Contemporary Cryptography for IT Practitioners. World Scientific Publishing Co., 2005.

[4] B. Ryabko, D. Ryabko. Confidence Sets in Time - Series Filtering, In: Proceedings of 2011 IEEE International Symposium on Information Theory (ISIT'11), July, 31 - August, 5, 2011, Saint-Petersburg, Russia

[5] Shannon, C. E. (1949). Communication theory of secrecy systems. Bell system technical journal, 28(4), 656-715.

RATE-DISTORTION ANALYSIS FOR KERNEL-BASED DISTORTION MEASURES

Kazuho Watanabe

Department of Computer Science and Engineering, Toyohashi University of Technology,
1-1 Hibarigaoka Tempaku-cho Toyohashi, Aichi 441-8580, JAPAN, wkazuho@cs.tut.ac.jp

ABSTRACT

Kernel methods have been used for turning linear learning algorithms into nonlinear ones. These nonlinear algorithms measures distances between data points by the distance in the kernel-induced feature space. However, the rate-distortion tradeoffs associated with such distortion measures have not been evaluated theoretically. We provide bounds to the rate-distortion functions for two reconstruction schemes, reconstruction in input space and reconstruction in feature space. Comparison of the derived bounds to the quantizer performance obtained by the kernel $\mathcal{K}$-means method suggests that the rate-distortion bounds for input space and feature space reconstructions are informative at low and high distortion levels, respectively.

1. INTRODUCTION

Kernel methods have been widely used for nonlinear learning problems combined with linear learning algorithms such as the support vector machine and the principal component analysis [1]. By the so-called kernel trick, kernel-based methods can use linear learning methods in the kernel-induced feature space without explicitly computing the high-dimensional feature mapping. Kernel-based methods measure the dissimilarity between data points by the distance in the feature space, which in input space, corresponds to a distance measure involving the feature mapping [2]. If a kernel-based learning method is used as a lossy source coding scheme, its optimal rate-distortion tradeoff is indicated by the rate-distortion function associated with the distortion measure defined by the kernel feature map [3]. However, the rate-distortion function of such a distortion measure has yet to be evaluated analytically. Although there are several kernel-based approaches to vector quantization [4, 5], their rate-distortion tradeoffs have been unknown.

In this paper, we derive bounds to the rate-distortion functions for kernel-based distortion measures. We consider two schemes to reconstruct inputs in lossy coding methods. One is to obtain a reconstruction in the original input space. Since kernel methods usually yield results of learning by the linear combination of vectors in feature space, we need an additional step to obtain the reconstruction in input space, such as preimaging [6]. We derive lower and upper bounds to the rate-distortion function of this scheme (Section 4.1 and Section 4.2). The other is to consider the linear combination of feature vectors as the reconstruction and measure the distortion in the feature space directly. We provide an upper bound to the rate-distortion function for this distortion measure (Section 4.3).

We train the vector quantizer using the kernel $\mathcal{K}$-means method and compare its performance with the derived rate-distortion bounds (Section 5). It is demonstrated that the rate-distortion bounds of reconstruction in input space are accurate at low distortion levels while the upper bound for reconstruction in feature space is informative at high distortion levels.

2. RATE-DISTORTION FUNCTION

Let X and Y be random variables of input and reconstruction whose domains are $\mathcal{X}$ and $\mathcal{Y}$, respectively. For the non-negative distortion measure between x and y, $d(x, y)$, the rate-distortion function $R(D)$ of the source $X \sim p(x)$ is defined by

$$R(D) = \inf_{q(y|x):E[d(X,Y)]\leq D} I(q), \tag{1}$$

where

$$I(q) = \int\int q(y|x)p(x) \log \frac{q(y|x)}{\int q(y|x)p(x)dx} dxdy$$

is the mutual information and E denotes the expectation with respect to $q(y|x)p(x)$. $R(D)$ shows the minimum achievable rate for the i.i.d. source with the density $p(x)$ under the given distortion measure d [3, 7]. The distortion-rate function is the inverse function of the rate-distortion function and denoted by $D(R)$.

If the conditional distribution q_s achieves the minimum of the following Lagrange function parameterized by $s \geq 0$,

$$L(q) = I(q) + s\left(E[d(X, Y)] - D\right),$$

then, the rate-distortion function is parametrically given by

$$\begin{aligned} R(D_s) &= I(q_s), \\ D_s &= \int q_s(y|x)p(x)d(x, y)dxdy. \end{aligned}$$

The parameter s corresponds to the (negated) slope of the tangent of $R(D)$ at $(D_s, R(D_s))$ and hence is referred to as the slope parameter [3].

From the properties of the rate-distortion function $R(D)$, we know that $R(D) > 0$ for $0 < D < D_{\max}$, where

$$D_{\max} = \inf_{y} \int p(x)d(x,y)dx, \quad (2)$$

and $R(D) = 0$ for $D \geq D_{\max}$ [3, p. 90]. Hence, $D_{\max} = \lim_{R \to 0} D(R)$.

3. KERNEL-BASED DISTORTION MEASURES

In kernel-based learning methods, data points in input space $\mathcal{X}$ are mapped into some high-dimensional feature space F by a feature mapping ϕ. Then the similarity between the two points x and y in $\mathcal{X}$ is measured by the inner product $\langle \phi(x), \phi(y) \rangle$ in F.

The inner product is directly evaluated by a nonlinear function in input space

$$K(x,y) = \langle \phi(x), \phi(y) \rangle \,, \quad (3)$$

which is called the kernel function. Mercer's theorem ensures that there exists some ϕ such that Eq. (3) holds if K is a positive definite kernel [1]. This enables us to avoid explicitly computing the feature map ϕ in the potentially high-dimensional space F, which is called the *kernel trick*. A lot of learning methods which can be expressed by only the inner products between data points have been kernelized [1].

3.1. Reconstruction in Input Space

If we restrict ourselves to the reconstruction in input space, that is, the reconstruction $y \in \mathcal{X} \subset \mathbf{R}^d$ is computed for each input $x \in \mathcal{X}$, the distortion measure is naturally defined by

$$\begin{aligned} d_{\text{inp}}(x,y) &= ||\phi(x) - \phi(y)||^2 \\ &= K(x,x) + K(y,y) - 2K(x,y). \quad (4) \end{aligned}$$

To obtain a reconstruction in input space, we need a technique such as preimaging [6].

This is a difference distortion measure if and only if the kernel function is translation invariant, $K(x+a, y+a) = K(x,y)$ for any $a \in \mathcal{X}$. That is,

$$d_{\text{inp}}(x,y) = \rho(x-y), \quad (5)$$

where $\rho(z) = 2(C - K(z,0))$ and $C = K(0,0)$. The rate-distortion function (distortion-rate function, resp.) for this distortion measure is denoted by $R_{\text{inp}}(D)$ ($D_{\text{inp}}(R)$, resp.) and the maximum distortion $D_{\max}$ in Eq. (2) is denoted by $D_{\max,\text{inp}}$.

3.2. Reconstruction in Feature Space

Suppose we have a sample of length n in input space, $\{x_1, ..., x_n\}$. If we compute the reconstruction by the linear combination $\sum_{i=1}^n \alpha_i \phi(x_i)$ for $\alpha_i \in \mathbf{R}, i = 1, ..., n$, and consider it as the reconstruction in feature space, the distortion can be measured by

$$\begin{aligned} d_{\text{fea}}(x, \boldsymbol{\alpha}) &= \left\| \phi(x) - \sum_{i=1}^n \alpha_i \phi(x_i) \right\|^2 \\ &= K(x,x) - 2\boldsymbol{\alpha}^T \boldsymbol{k}(x) + \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \quad (6) \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T \in \mathbf{R}^n$,

$$\boldsymbol{k}(x) = (K(x_1, x), ..., K(x_n, x))^T,$$

and $\boldsymbol{K} = (K(x_i, x_j))_{ij}$ is the Gram matrix. Note that the reconstruction is identified with the coefficients $\boldsymbol{\alpha}$ whose domain is not identical to the input space $\mathcal{X} \subset \mathbf{R}^d$.

The rate-distortion function (distortion-rate function, resp.) for this distortion measure is denoted by $R_{\text{fea}}(D)$ ($D_{\text{fea}}(R)$, resp.) and the maximum distortion $D_{\max}$ in Eq. (2) is given by

$$D_{\max,\text{fea}} = E[K(x,x)] - E[\boldsymbol{k}(x)]\boldsymbol{K}^{-1}E[\boldsymbol{k}(x)]^T.$$

4. RATE-DISTORTION BOUNDS

4.1. Lower Bound to $R_{\text{inp}}(D)$

Although the Shannon lower bound to $R(D)$ is defined for difference distortion measures in general [3, p. 92], it diverges to $-\infty$ for the distortion measure (5) since $\int e^{-s\rho(z)}dz$ diverges to ∞. Hence, we consider an improved lower bound, which is discussed in [3, p. 140]. Let Q_B be the probability that $\|X\| \leq B$. Then, $R(D)$ is lower-bounded as

$$R(D) \geq Q_B \left\{ h(p_B) - \max_{g \in G_{B,D}} h(g) \right\}, \quad (7)$$

where h denotes the differential entropy,

$$p_B(x) = \frac{1}{Q_B} p(x) u(B - \|x\|),$$

and u is the step function. $G_{B,D}$ is the set of all probability densities $g(\cdot)$ for which $g(x) = 0$ for $\|x\| > B$ and $\int \rho(z)g(z)dz \leq D/Q_B$.

The maximum in Eq. (7) is explicitly given by

$$g_s(z) = \frac{1}{C_{B,s}} \exp\left(2sK(z,0)\right) u(B - \|z\|), \quad (8)$$

where $C_{B,s} = \int_{\|z\| \leq B} e^{2sK(z,0)}dz$ for s related to D by $\int \rho(z) g_s(z) dz = D/Q_B$. Since its differential entropy is

$$h(g_s) = -s\frac{\partial \log C_{B,s}}{\partial s} + \log C_{B,s}, \quad (9)$$

$R_{\text{inp}}(D)$ is parametrically lower-bounded by

$$\begin{aligned} R_{\text{inp},L}(D_s) &= Q_B \left\{ h(p_B) + s\frac{\partial \log C_{B,s}}{\partial s} - \log C_{B,s} \right\}, \\ D_s &= Q_B \left\{ 2C - \frac{\partial \log C_{B,s}}{\partial s} \right\}. \end{aligned}$$

4.2. Upper Bound to $R_{\text{inp}}(D)$

If d_{inp} in Eq. (4) is a difference distortion measure, that is, K is translation invariant, by choosing $q(y|x) = g_s(y-x)$ for the density g_s in Eq. (8), the following upper bound is obtained,

$$\begin{aligned} R_{\text{inp}}(D_s) \leq R_{\text{inp},U}(D_s) &= h(g_s * p) - h(g_s) \quad (10) \\ D_s &= 2C - \frac{\partial \log C_{B,s}}{\partial s}. \quad (11) \end{aligned}$$

where $h(g_s)$ is given by Eq. (9) and $(g_s * p)(y) = \int g_s(y-x)p(x)dx$ is the convolution between g_s and p.

By the maximum entropy principle of the Gaussian distribution, $R_{\mathrm{inp},U}(D)$ is further upper-bounded by

$$R_{\mathrm{inp},G}(D_s) = \frac{d}{2}\log(2\pi e(v_p + v_s)) - h(g_s),$$

where

$$\begin{aligned} v_p &= \frac{1}{d}\int \|x-m\|^2 p(x)dx, & (12)\\ m &= \int xp(x)dx & (13)\\ v_s &= \frac{1}{d}\int \|x\|^2 g_s(x)dx, & (14)\\ &= \frac{A(d)}{dC_{B,s}}\int_0^B r^{d+1}e^{2sk(r)}dr. & (15)\end{aligned}$$

where $A(d) = \frac{d\sqrt{\pi}^d}{\Gamma(d/2)+1}$ is the area of the d-dimensional unit sphere. Here, we have further assumed that the kernel function is radial, that is, $K(x,y) = K(x-y,0) = k(\|x-y\|)$ for some function k.

4.3. Upper Bound to $R_{\mathrm{fea}}(D)$

We construct an upper bound to the rate-distortion function $R_{\mathrm{fea}}(D)$. We choose the conditional distribution of the reconstruction by

$$q(\boldsymbol{\alpha}|x) = N(\boldsymbol{\alpha}; \boldsymbol{m}_K(x), \tilde{\boldsymbol{K}}^{-1}/2s),$$

where $\tilde{\boldsymbol{K}} = \boldsymbol{K} + c\boldsymbol{I}$,

$$\boldsymbol{m}_K(x) = \tilde{\boldsymbol{K}}^{-1}\boldsymbol{k}(x),$$

and $N(\cdot;\boldsymbol{m},\boldsymbol{\Sigma})$ denotes the n-dimensional normal density with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, we have introduced the regularization constant $c \geq 0$ with the $n \times n$ identity matrix $\boldsymbol{I}$. This reconstruction distribution yields the following upper bound,

$$\begin{aligned} R_{\mathrm{fea}}(D_s) &\leq R_{\mathrm{fea},U}(D_s)\\ &= h(M_p) - h(N(\boldsymbol{\alpha};\boldsymbol{m}_K(x),\tilde{\boldsymbol{K}}^{-1}/2s)),\\ D_s &= \int p(x)q(\boldsymbol{\alpha}|x)d_{\mathrm{fea}}(x,\boldsymbol{\alpha})dxd\boldsymbol{\alpha}\\ &= \frac{n}{2s} + D_{\min}(c),\end{aligned}$$

where $M_p(\boldsymbol{\alpha}) = \int N(\boldsymbol{\alpha};\boldsymbol{m}_K(x),\tilde{\boldsymbol{K}}^{-1}/2s)p(x)dx$,

$$h(N(\boldsymbol{\alpha};\boldsymbol{m}_K(x),\tilde{\boldsymbol{K}}^{-1}/2s)) = \frac{n}{2}\log\left(\frac{\pi}{s}|\tilde{\boldsymbol{K}}|^{1/n}\right),$$

which is independent of the input x, and

$$\begin{aligned} D_{\min}(c) &= E[K(x,x)] - \mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}E[\boldsymbol{k}(x)\boldsymbol{k}(x)^T]\}\\ &\quad + c\mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}E[\boldsymbol{k}(x)\boldsymbol{k}(x)^T]\tilde{\boldsymbol{K}}^{-1}\}.\end{aligned}$$

If $c = 0$, $D_{\min}$ is the mean of the variance of the prediction by the associated Gaussian process [8].

Further upper-bounding the differential entropy $h(M_p)$ by the Gaussian entropy, we have

$$R_{\mathrm{fea}}(D) \leq R_{\mathrm{fea},G}(D) = \frac{1}{2}\log\left|\boldsymbol{I} + \frac{n\tilde{\boldsymbol{K}}^{-1}\boldsymbol{C}}{D - D_{\min}(c)}\right|, \tag{16}$$

where $\boldsymbol{C} = E[\boldsymbol{k}(x)\boldsymbol{k}(x)^T] - E[\boldsymbol{k}(x)]E[\boldsymbol{k}(x)]^T$.

5. EXPERIMENTAL EVALUATION

We numerically evaluate the rate-distortion bounds obtained in the previous section. Designing a quantizer by the kernel $\mathcal{K}$-means algorithm, we compare its performance with the bounds.

We focus on the case of the Gaussian kernel,

$$K(x,y) = e^{-\gamma\|x-y\|^2}$$

with the kernel parameter $\gamma > 0$. As a source, we assumed the uniform distribution on the union of the two regions, $C_1 = \{x \in \mathbf{R}^d; A(d)\|x\|^d \leq d/2\}$ and $C_1 = \{x \in \mathbf{R}^d; d^2 \leq A(d)\|x\|^d \leq d(d+1/2)\}$, where C_1 and C_2 have equal volumes and $C_1 \cup C_2$ has volume 1.

We used the trapezoidal rule to compute the integrations in the lower bound $R_{\mathrm{inp},L}$ and the upper bound $R_{\mathrm{inp},G}$. We generated i.i.d sample of the size $n = 4000$ from the source to compute $\boldsymbol{k}(x)$ and $\boldsymbol{K}$ for $R_{\mathrm{fea},G}$ in Eq. (16). Generating another 4000 data points, we approximated the required expectations.

Using the same data set of the size 4000 as a training data set, we run the kernel $\mathcal{K}$-means algorithm 10 times with random initializations to obtain the minimum distortion for each rate. Varying the number $\mathcal{K}$ of quantized points from 2^1 to 2^{10} , for each $\mathcal{K}$, we counted the effective number $\mathcal{K}_{\mathrm{eff}}$ of quantized points which have at least one assigned data point and computed rates by $\log_2 \mathcal{K}_{\mathrm{eff}}$. The kernel parameter γ was chosen so that the clear separation of C_1 and C_2 is obtained when $\mathcal{K} = 2$. We optimized the regularization coefficient c to minimize the upper bound $R_{\mathrm{fea},G}$ for each $\mathcal{K}$.

After the training, we computed the distortion and rate for the test data set, by assigning each of 20000 test data generated from the same source to the nearest quantized points in the feature space.

The obtained bounds and the quantizer performance are displayed in Figure 1 and Figure 2 for $d = 2$ and $d = 10$, respectively, in the forms of distortion-rate functions.

In both dimensions, the upper bound $D_{\mathrm{fea},G}$ is smaller than $D_{\mathrm{inp},G}$ at low rates while the bound is above the quantizer performance. However, the value of $D_{\mathrm{max,fea}}$ suggests that the bound is informative at low rates. As the rate becomes higher, the lower and upper bounds of the input-space-reconstruction, $D_{L,\mathrm{inp}}$ and $D_{G,\mathrm{inp}}$, approach each other. In fact, they sandwich the quantizer performance tightly in the 2-dimensional case, which suggests that the rate-distortion function for the feature space reconstruction, $R_{\mathrm{fea}}(D)$ is close to the rate-distortion function of the input space reconstruction $R_{\mathrm{inp}}(D)$ at high rates.

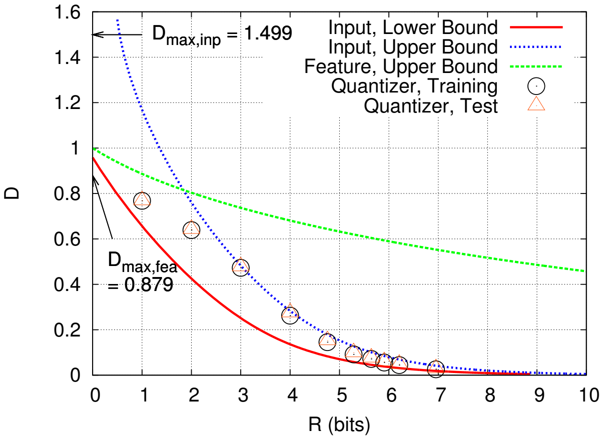


Figure 1. Rate-distortion bounds and quantizer performance for $d = 2$.

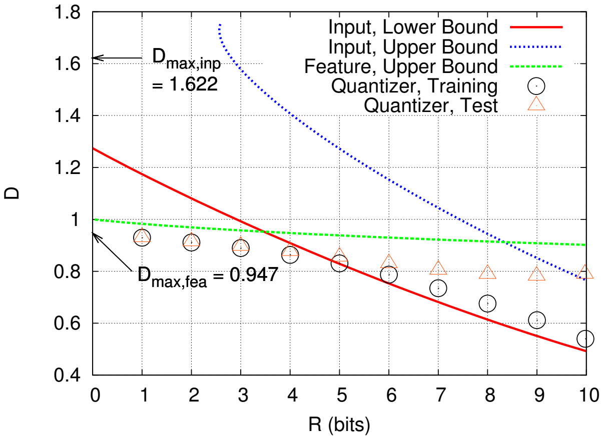


Figure 2. Rate-distortion bounds and quantizer performance for $d = 10$.

At low distortion levels, each source output should be reconstructed within a small neighborhood in the feature space where we can find another point y in the input space whose feature map $\phi(y)$ is sufficiently close to the reconstruction. This suggests that the rate-distortion function of feature space reconstruction is well approximated by the rate-distortion function of input space reconstruction. In other words, combining multiple input points to make a reconstruction in feature space does not do any good for reducing distortion and only a single input point is enough when it is mapped into feature space. Hence, the rate-distortion bounds of input space reconstruction may be informative at low distortion levels.

In the 10-dimensional case, the distortion in the test data set is close to $D_{\text{inp},G}(R)$ or above it at high rates. This may be due to overfitting of the kernel $\mathcal{K}$-means to the training data set of the size, 4000. That is, as the the rate grows, the distortion in the training data set decreases and the discrepancy between the distortions in the training and test sets increases. If the quantizer is designed with more training data, its performance would lie between the bounds of reconstruction in the input space, $D_{\text{inp},L}$ and $D_{\text{inp},G}$, as in the 2-dimensional case.

6. CONCLUSION

In this extended abstract, we have shown upper and lower bounds for the rate-distortion functions associated with kernel-feature mapping. As suggested in Section 5, the upper bound for the reconstruction in feature space is informative at high distortion levels while the bounds for the reconstruction in input space are informative at low distortion levels. Our future directions include deriving tighter bounds and exact evaluation of the rate-distortion function in some special cases. In particular, it is an important undertaking to derive a lower bound to the rate-distortion function of the reconstruction in feature space.

7. ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI Grant Numbers 25120014 and 15K16050.

8. REFERENCES

[1] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2001.

[2] M. A. Aizerman, E. A. Braverman, and L. Rozonoer, “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[3] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[4] M. Girolami, “Mercer kernel-based clustering in feature space,” *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 780–784, 2002.

[5] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, “A survey of kernel and spectral methods for clustering,” *Pattern Recognition*, vol. 41, no. 1, pp. 176–190, 2008.

[6] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, G. Ratsch, and A. J. Smola, “Input space versus feature space in kernel-based methods,” *IEEE Transactions on Neural Networks*, vol. 10, pp. 1000–1017, 1999.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 1991.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

ACHIEVABLE INFORMATION RATES ON LINEAR INTERFERENCE CHANNELS WITH DISCRETE INPUT

Metodi Yankov and Søren Forchhammer [1]

[1]Department of Photonics Engineering, Technical University of Denmark,
Bldng. 343, Ørsteds Plads, 2800 Kgs. Lyngby, Denmark
meya@fotonik.dtu.dk

ABSTRACT

In this paper lower bound on the capacity of multi-dimensional linear interference channels is derived, when the input is taken from a finite size alphabet. The bounds are based on the QR decomposition of the channel matrix, and hold for any input distribution that is independent across dimensions. Calculation of the bounds can be performed on a per-dimensions basis via look-up tables of the information rates of 1D channels.

1. INTRODUCTION

The capacity of a set of linearly interfering channels when the input is taken from a finite size alphabet has been a long standing problem in information theory. In the case of the Multiple Input Multiple Output (MIMO) channel with Gaussian input the capacity has been found [1]. When the transmitter has perfect knowledge of the channel, it can align the input to the channel eigen modes and allocate the power based on the water-filling strategy. When the channel is known at the receiver only, i.i.d. Gaussian input is optimal. It has been shown in [2] that when the input is discrete, both orthogonalization and water-filling power allocation are sub-optimal. Low and high SNR asymptotic expressions for the capacity in the discrete case are derived based on the Mutual Information (MI) - Minimum Mean Squared Error (MMSE) relation [3][4][5]. Due to the requirement for high spectral efficiency on current communication systems, the mid-SNR is usually where they operate. The MIMO Constellation Constrained Capacity (CCC) in this region remains unknown. The capacity of a standard impulse response channel with discrete input is another open problem in the area of linear interference channels. The general method for computing it relies on trellis processing [6], which quickly becomes intractable when the channel memory increases. Some extensions and simplifications exist, e.g. [7], which usually attempt to shorten the memory length, however, they still suffer from the inherent complexity of the trellis description.

In [8] we derived a lower bound on the CCC of the ergodic MIMO channel with i.i.d. matrix elements using the QR Decomposition (QRD) of the channel. Here we generalize this result to the single channel realization case, and we use it to also bound the Achievable Information rate (AIR) on a general impulse response channel.

2. CHANNEL MODEL AND COMPLEXITY PROBLEM

Consider a standard MIMO channel model:

$$Y = \mathbf{H}X + W, \tag{1}$$

where X is M-dimensional complex random variable vector $X = [X_1, X_2, \dots X_M]^T$, which is discrete and takes values from the complex-valued set $\mathcal{X}^M$, obtained as the Cartesian product of the basic 1D set $\mathcal{X}$. This can be a QAM, APSK, etc. complex-valued set. The matrix $\mathbf{H}$ represents the $[N\mathbf{x}M]$ complex-valued channel, W is N dimensional complex AWGN, assumed here to have unit variance and Y is the N dimensional channel observation. We assume the channel realization is known at the receiver, but not at the transmitter. The realization of a random variable, e.g. X, at time k will be denoted as $\mathbf{x}_k$ (x_k in the case of 1D variable), and the sequence from time t to k as $\mathbf{x}_t^k = [\mathbf{x}_t, \mathbf{x}_2, \dots \mathbf{x}_k]^T$.

The AIR on the channel when signaling with $\mathcal{X}^M$, having Probability Mass Function (PMF) $p(X)$, and averaging among the possible channel realizations is given by the MI:

$$\begin{aligned}\mathcal{I}(X;Y) = \mathrm{E}_{\mathbf{H}}\left[\mathcal{I}(X;Y|\mathbf{H})\right] = \\ \mathcal{H}(X) - \mathrm{E}_{\mathbf{H}}\left[\mathcal{H}(X|Y,\mathbf{H})\right].\end{aligned} \tag{2}$$

The standard method for calculating the MI is to generate a long enough pair of input-output sequences, and use the fact, that the entropy converges [6]:

$$\mathcal{H}(X|Y,\mathbf{H}) = -\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\log_2 p(\mathbf{x}_k|\mathbf{y}_k,\mathbf{H}). \tag{3}$$

The probability above is calculated from Bayes theorem:

$$p(\mathbf{x}_k|\mathbf{y}_k,\mathbf{H}) = \frac{p(\mathbf{y}_k|\mathbf{x}_k,\mathbf{H})p(\mathbf{x}_k)}{\sum_{\mathbf{x}_k\in\mathcal{X}^M} p(\mathbf{y}_k|\mathbf{x}_k,\mathbf{H})p(\mathbf{x}_k)} \tag{4}$$

Since the normalization term in (4) must be calculated, the complexity grows exponentially with M. Furthermore, in order to see the convergence in (3), K must also be increased with M. Going beyond e.g. 64QAM on a 3x2 channel on a standard computer becomes challenging.

3. LOWER BOUNDS

Let $\mathbf{H} = \mathbf{QR}$ be the QR decomposition of $\mathbf{H}$, where $\mathbf{Q}$ is unitary and $\mathbf{R}$ is upper-triangular. A well known MIMO receiver utilizes the form of $\mathbf{R}$ to successively cancel the interference from previously detected layers, hence Successive Interference Cancellation (SIC), in the following manner: the received samples are pre-processed as $\hat{Y} = \mathbf{Q}^H Y$, and the channel model becomes $\hat{Y}_i = \sum_{j=i}^{M} \mathbf{R}_{i,j} X_j$. Assuming the layers $i+1$ to M are correctly decoded by the following channel code, the symbols can be remodulated and subtracted from the current layer i. Here we use a similar technique to derive a lower bound on the channel capacity.

Since $\mathbf{Q}$ is unitary and doesn't change the entropy of Y, and thus the MI, we can write:

$$\begin{aligned}\mathcal{I}(X;Y|\mathbf{H}) = \mathcal{H}(X) - \mathcal{H}(X|\hat{Y}|\mathbf{H}) = \\ \mathcal{H}(X) - \sum_{i=1:M} \mathcal{H}(X_i|\hat{Y}, X_{i+1}^M, \mathbf{H}) \geq \\ \mathcal{H}(X) - \sum_{i=1:M} \mathcal{H}(X_i|\hat{Y}_i, X_{i+1}^M, \mathbf{H}) = \underline{\mathcal{I}}(X;Y|\mathbf{H}), \quad (5)\end{aligned}$$

where we have used the fact, that conditioning does not increase the entropy. In order to calculate the terms in the sum, we express the posterior probabilities similar to (4):

$$p(X_i|\hat{Y}_i, X_{i+1}^M, \mathbf{H}) = \frac{p(X_i)p(\hat{Y}_i|X_i, X_{i+1}^M, \mathbf{H})}{\sum_{X_i} p(X_i)p(\hat{Y}_i|X_i, X_{i+1}^M, \mathbf{H})} \quad (6)$$

Since we condition on the following layers, the likelihood above can be expressed as:

$$\begin{aligned}p(\hat{Y}_i|X_i, X_{i+1}^M, \mathbf{H}) = \mathcal{N}(\hat{Y}_i| \sum_{j=i:M} R_{i,j}X_j, 1) = \\ \mathcal{N}(\hat{Y}_i - \sum_{j=i+1:M} R_{i,j}X_j|X_i, 1), \quad (7)\end{aligned}$$

where $R_{i,j}$ is the element on the $i-$th row and $j-$th column of $\mathbf{R}$, and $\mathcal{N}(x|\mu, \sigma^2)$ is a 1D Gaussian function at x, with mean and variance μ and σ^2, respectively. Using (7), lower bound on the MI on each layer can be calculated independently from an SNR-MI Look-Up Table (LUT), where the SNR is given by $|R_{i,i}|^2 \mathrm{E}\left[X_i^2\right]$. When $M <= N$, the achievable rate on the $M-$th layer coincides with the actual capacity for that layer. However, when $M > N$, there is residual interference on the $N+1$-st to the $M-$th layers from layers, which are not yet decoded, and the resulting lower bound becomes poorer. In order to improve it, we model the residual interference as noise, which is a standard practice in communications engineering. The likelihood we use on layers $i > N$ is then:

$$\mathcal{N}(\hat{Y}_N - \sum_{j=i+1:M} R_{N,j}X_j|R_{N,i}X_i, \hat{\sigma}_i), \quad (8)$$

where $\hat{\sigma}_i = 1 + \sum_{j=N:i-1} |R_{N,j}|^2 \mathrm{E}\left[X_i^2\right]$. In this case it is clear, that in the asymptotically high SNR we have:

$$\lim_{\mathrm{E}\left[X_i^2\right]\to\infty} \mathcal{I}(X;Y|\mathbf{H}) = \mathcal{H}(X), \quad (9)$$

whereas:

$$\begin{aligned}\lim_{\mathrm{E}\left[X_i^2\right]\to\infty} \underline{\mathcal{I}}(X;Y|\mathbf{H}) = \\ \mathcal{H}(X) - \sum_{i=N+1:M} \mathcal{H}(X_i|Y, H, SNR_i), \quad (10)\end{aligned}$$

where the conditional entropy is larger than zero, because $\lim_{\mathrm{E}\left[X_i^2\right]\to\infty} SNR_i = \frac{R_{N,i}}{\sum_{j=N:i-1} R_{N,j}}$, which is a finite number.

3.1. Relation to auxiliary channel lower bounds

A simple upper bound on the entropy of a variable X with PDF $p(X)$ can be obtained by using an *auxiliary* probability function $\bar{p}(X) \neq p(X)$. If X is generated by its original PDF, then the upper bound is found by calculating the entropy function from X, but using $\bar{p}(X)$ [6]:

$$\bar{\mathcal{H}}(X) = -\frac{1}{K}\sum_k \log_2 \bar{p}(x_k) \geq \mathcal{H}(X)$$

A lower bound on the MI is derived in a similar manner. Say there is a channel with input-output sequence pair $x \to y$, governed by the laws $p_{Y|X}(Y|X)$, and $p_{X|Y}(X|Y) = \frac{p_{Y|X}(Y|X)p_X(X)}{\sum_X p_{Y|X}(Y|X)p_X(X)}$. Then if y is generated by the law $p_{Y|X}(Y|X)$, the lower bound is calculated as:

$$\begin{aligned}\underline{\mathcal{I}}(X;Y) = \mathcal{H}(X) - \bar{\mathcal{H}}(X|Y) \leq \\ \mathcal{H}(X) - \mathcal{H}(X|Y) = \mathcal{I}(X;Y), \quad (11)\end{aligned}$$

where $\bar{\mathcal{H}}(X|Y)$ is calculated using some valid PMF $\bar{p}_{X|Y}(X|Y) \neq p_{X|Y}(X|Y)$.

Turning back to the $\mathbf{R}$ channel, we use the auxiliary probability distribution $\bar{p}(X|Y, \mathbf{H})$:

$$\begin{aligned}\bar{p}(X|\hat{Y}, \mathbf{H}) = \prod_{i=1}^{M} \bar{p}(X_i|\hat{Y}, X_{i+1}^M, \mathbf{H}) \\ = \prod_{i=1}^{M} \frac{\bar{p}(\hat{Y}|X_i, X_{i+1}^M, \mathbf{H})p(X_i)}{\sum_{X_i} \bar{p}(\hat{Y}|X_i, X_{i+1}^M, \mathbf{H})p(X_i)},\end{aligned}$$

where:

$$\begin{aligned}\bar{p}(\hat{Y}|X_i, X_{i+1}^M, \mathbf{H}) = \\ \mathcal{N}(\hat{Y}_i - \sum_{j=i+1:M} \mathbf{R}_{i,j}X_j|\mathbf{R}_{i,i}X_i, 1), \quad (12)\end{aligned}$$

which leads to the same lower bound.

3.2. Impulse response channels

Consider a standard impulse response channel:

$$y_k = \sum_{i=0:l} h_i x_{k-i} + w_k, \quad (13)$$

where $\mathbf{h} = [h_0, h_1, \ldots h_l]^T$ is the impulse response. Equivalently, the channel may be expressed in its matrix form:

$$y_1^k = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ h_l & h_{l-1} & \ddots & \vdots \\ 0 & h_l & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & h_1 & h_0 \end{bmatrix} \times x_1^k + w_1^k. \qquad (14)$$

The MI with channel knowledge at the receiver is now calculated as:

$$\mathcal{I}(X;Y|\mathbf{h}) = \mathcal{H}(Y|\mathbf{h}) - \mathcal{H}(Y|X,\mathbf{h}) =$$
$$= -\frac{1}{K}\log_2 p(y_1^K|\mathbf{h}) - H(W) \qquad (15)$$

The standard approach to calculating (15) is to use a trellis to calculate $p(y_1^K) = \prod_{1:K} p(y_k|y_{1:K-1})$. One section of such trellis is given in Fig. 1. The interfering symbols are cast into the state: $S_k = \{X_{k-l}, \ldots X_{k-1}\}$, and the current symbol governs the transition. Marginalizing the state, the desired probability at time k is $p(y_1^k) = \sum_{s_k} p(s_k, y_1^k)$, where each term is calculated recursively [6]:

$$p(s_k, y_1^k) = \sum_{x_k}\sum_{s_{k-1}} p(s_{k-1}, y_1^{k-1})p(y_k|x_k, s_k)p(x_k|s_k)$$

Since the number of states is given by $|\mathcal{S}| = |\mathcal{X}|^l$, the dimensionality problem is the same as for the MIMO channel. The equivalent of the above mentioned 3x2 64QAM here is 64QAM with maximum 2 taps, or similarly - 16QAM with maximum 3 taps, for a standard PC. Trellis pruning techniques may be utilized both in case of MIMO and impulse response channels, leading to the so-called sphere detection [9]. Sphere detection is popular, but is still limited in the number of nodes which can be pruned before the performance degrades significantly. Another approach for the impulse response channel is to use an auxiliary channel of shorter length [7]. The same problem exist here - the more the channel is shortened, the worse auxiliary channel we can find, and thus worse lower bounds.

Instead we can use the QRD based lower bounds. If the channel is expressed as in (14), the QR decomposition may be performed, and a bound may be obtained by the above mentioned LUT. In this case $M = N$, and so Eq. (7) is used. This method is independent of the memory length. The only bottleneck is the QRD computation, which for very long sequences may become problematic. In this paper we used $K = 10^4$, which we found was enough to see convergence for 16QAM constellations. The QRD on the $\left[10^4\text{x}10^4\right]$ matrix was computed in a few seconds on the PC we used. We note that the channel matrix in this case is highly structured and periodic, and the $\mathbf{R}$ matrix therefore may be expected to also hold some structure. For example, in all our simulations the diagonal elements of the $\mathbf{R}$ matrix either converged to some value, or to some periodic pattern. However, exploiting this periodicity is left for future research.

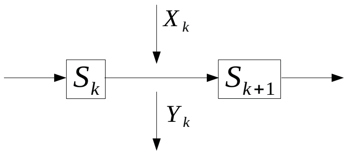


Figure 1. Trellis representation of the impulse response channel (13)

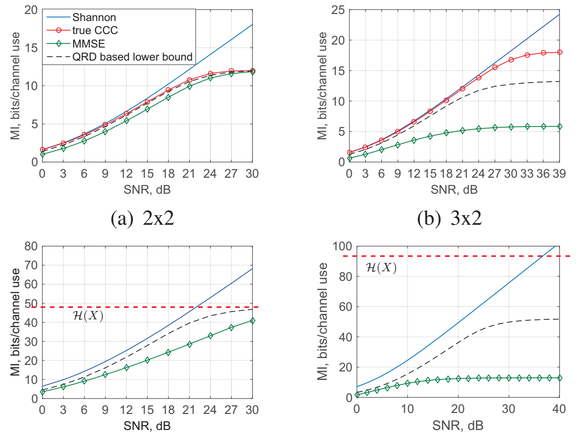
4. RESULTS

4.1. MIMO channel

In Fig. 2(a) the lower bounds from Eq. (5) are shown for a 2x2 MIMO with 64QAM input, together with the true MI, as calculated from Eq. (2). The input PMF is uniform. For comparison, we also plot the AIRs with the popular linear MMSE receiver processing [9]. We see that the true information rate is closely approached by the proposed method. The MMSE processing also calculates a lower bound, however, poorer than the QRD based one. As mentioned in Section 3, in the case of $M > N$, the bounds will not be as tight. In Fig. 2(b) the AIRs are shown for a 3x2 MIMO with 64QAM input. We see a significant underestimation, especially in the high SNR region. However, we note that the transmit diversity system is generally not used for maximizing throughput, and therefore a practical system would not operate at this high SNR region with an input of rank, which is larger than $rank(H) \leq \min(M, N)$. In the low-to-mid SNR, the QRD based bound may still be used. In Fig. 2(c) the AIRs on a 8x8 system are shown, where the full-complexity algorithm can no longer be used. The QRD based lower bound follows the slope of the Gaussian capacity, and converges to $\mathcal{H}(X)$. When we further increase M, more terms are added in the conditional entropy in Eq. (10), and the lower bound becomes worse. However, the slope at low-to-mid SNR is still the same as the Gaussian capacity. Finally in this section we note, that the uniform PMF is not a requirement. The bounds hold for any PMF, which is independent across dimensions. The consequence is that optimization can also be performed using the auxiliary function (7). The PMF, which is optimized for the auxiliary channel can then be used on the true channel, and the AIR in that case is still bounded by what is achieved in (5). Some results obtained by the well known Blahut-Arimoto algorithm for optimization of the input PMF on an ergodic MIMO channel may be found in [8].

4.2. Impulse response channel

We also analyze the QRD based lower bound on a fixed impulse response channel, where $\mathbf{h}$ is obtained from standard Gaussian distribution. In Figures 3(a) and 3(b) we see the AIRs on an impulse response channel with $l = 3$ and $l = 6$, respectively (channel as given in the caption). Without loss of generality, we sort the channel elements



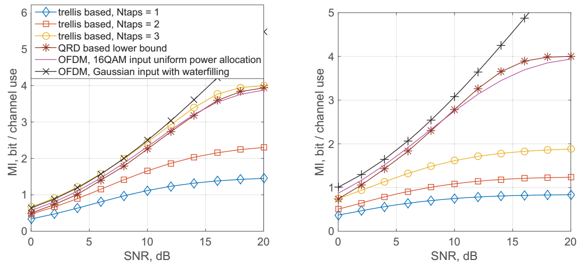
(a) 2x2

(b) 3x2

(c) 8x8

(d) 16x8

Figure 2. AIRs on MIMO channels of different size



(a) $l = 3$

(b) $l = 6$

Figure 3. Achievable information rates for the channels: a) $\mathbf{h} = [0.37 - 0.18i, -0.35 + 0.05i, -0.20 - 0.26i, -0.23 - 0.17i]^T$; b) $\mathbf{h} = [-0.18 - 0.47i, -0.35 - 0.26i, 0.37 - 0.06i, -0.26 + 0.25i, -0.17 + 0.27i, -0.23 + 0.18i, -0.20 + 0.05i]^T$

in descending order of their amplitude. The input symbols are i.i.d., and so this does not change the AIRs, but makes the implementation of the trellis simpler, since the state actually represents previous symbols. In the general case, we would like our state to represent the symbols, responsible for largest interference. We compare the QRD based bounds with the trellis based method, which casts $Ntaps$ previous symbols into the state, and the rest $l - Ntaps$ symbols are modeled as noise, similar to Eq. (8). When $Ntaps = l$ the AIR is the true CCC. For comparison we also include the AIRs using OFDM. We note that OFDM with Gaussian input and water-filling power allocation is the power constrained channel capacity. On the short channels, the QRD based lower bound closely approaches the constrained capacity, achieved with the trellis algorithm. It is slightly outperformed in the low SNR by OFDM, and slightly outperforms OFDM in the mid-to-high SNR. When we increase the channel length, the trellis based algorithm can be used with up to 3 taps on the PC we used for simulations. The QRD bound in this case is able to provide larger improvement over the OFDM. Both figures show that orthogonalization of the channel when the input is discrete can be sub-optimal, confirming the results from [2][8].

5. CONCLUSION

In this paper some of the more popular linear interference channels are studied. Lower bounds on the AIRs are derived using the QR decomposition of the channel. In case of linear 1D channels with memory, the QRD is performed on the matrix form of the channel. Based on the diagonal elements of the $\mathbf{R}$ matrix, an SNR-AIR look-up table can be efficiently used to find lower bounds on capacity. These bounds were shown to closely approach the true constellation constrained capacity, where the latter can be computed by standard PC, and were also shown to have good performance in terms of slope and distance to Gaussian capacity in most cases of interest.

6. REFERENCES

[1] G. J. Foschini and Gans M. J, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, 1998.

[2] F. Pereź-Cruz, M. R. D. Rodrigues, and S. Verdú, "MIMO gaussian channels with arbitrary inputs: Optimal precoding and power allocation," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1070–1084, 2010.

[3] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, pp. 1289–1301, 2012.

[4] A. Alvarado, F. Brännström, E. Agrell, and T. Koch, "High-SNR asymptotics of mutual information for discrete constellations with applications to BICM," *IEEE Trans. Inf. Theory*, vol. 60, pp. 1061–1076, 2014.

[5] M. Rodrigues, "Multiple-antenna fading channels with arbitrary inputs: Characterization and optimization of the information rate," *IEEE Trans. Inf. Theory*, vol. 60, pp. 569–585, 2014.

[6] D. M. Arnold, H. Loeliger, P. O. Vintobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, pp. 3498–3508, 2006.

[7] F. Rusek and D. Fertonani, "Bounds on the information rate of intersymbol interference channels based on mismatched receivers," *IEEE Trans. Inf. Theory*, vol. 58, pp. 1470–1482, 2012.

[8] M. P. Yankov, S. Forchhammer, K. J. Larsen, and L. P. B. Christensen, "Approximating the constellation constrained capacity of the MIMO channel with discrete input," in *Proc. IEEE International Conference on Communications.*, Jun. 2015, pp. 5634–5639.

[9] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*, Cambridge University Press, 2007.