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
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

3
Contents

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

Plenary Talks

Gerhard Kramer: Upper Bounds on the Capacity of Fiber Channels ................. 5
Steffen Lauritzen: Proper Local Scoring Rules ........................................... 6

Invited Papers

Pekka Astola and Ioan Tabus: Model Structure Selection for Sparse Predictive Coding Utilizing Warped Views .................................................. 7
Łukasz Dębowski: Regular Hilberg Processes: Nonexistence of Universal Redundancy Ratios ................................................................. 11
Søren Forchhammer and Huynh Van Luong Distributed Source Coding of Video ........ 15
Peter Harremoës: Proper Scoring and Sufficiency ........................................ 19
Jørn Justesen: Iterative Decoding of Product Codes ....................................... 23
Ioannis Kontoyiannis and Maria Skoularidou: Causality and Directed Information Estimation as a Hypothesis Test .......................................... 26
Mokshay Madiman and Ioannis Kontoyiannis: The Ruzsa Divergence on Groups .... 30
Mikhail Malyutov, Paul Grosu, and Tong Zhang: SCOT Modeling, Training and Statistical Inference ...................................................... 35
Mads Nielsen, Bo Markussen, and Marco Loog: Relevance Sampling ................ 39
Boris Ryabko and Nadezhda Savina: The One-Time Pad is Robust to Small Deviations from Randomness ............................................. 43
Kazuho Watanabe: Rate-Distortion Analysis for Kernel-Based Distortion Measures . 46
Metodi Yankov and Søren Forchhammer: Achievable Information Rates on Linear Interference Channels with Discrete Input ......................... 50
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 \( \log(1 + \text{SNR}) \), 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.
MODEL STRUCTURE SELECTION FOR SPARSE PREDICTIVE CODING UTILIZING WARPED VIEWS

Pekka Astola and Ioan Tabus

Department of Signal Processing, Tampere University of Technology, Tampere, FINLAND, firstname.lastname@tut.fi

ABSTRACT

The color image encoding of one view, conditional on the other view, makes use of the disparity image for computing a warped partial approximation, to be used for sparse predictive coding. Different estimates of the disparity image, produced by different stereo matching algorithms, will result in different final codelength values for the conditional coding of the color image. In here we study how the ranking of stereo matching algorithms based on the predictive coding codelength, which does not need the knowledge of the ground-truth disparity, relates to the other rankings of the stereo matching algorithms, that require the availability of the ground truth disparity. We show that the two information-theoretic variables introduced herein correlate well with the most used quantity for stereo matching ranking.

1. INTRODUCTION

Parametric modeling and model order selection have been used in lossless image compression from the early pioneering work up to the most recent high performance solutions, see e.g., [1] and [2] for gray level images.

Recently there has been high interest in 3D image data, where besides the three color components there is also a depth, (or disparity) component, forming together RGB+D images. The depth component D is acquired directly by depth sensors, or is estimated from stereo or multiview images in the form of a disparity image. This has led to a plethora of compression methods, out of which many also expanded the 2D image compression solutions in which the parametric modeling plays an important role.

We have investigated in [3] a RGB modelling and compressing scheme where parametric modelling plays the major role: we considered the regions where the depth value is constant in D as tentative regions for using a specific predictor, and we have defined the predictors as acting on the color components of the current view and additionally having available the color components of a related view (as in stereo or multiview scenario). Hence, the predictors have to reconstruct the current color components by using the warped related view color components. A partition P of the image is formed by those regions where the cost of model plus cost of residuals of a specific predictor is better than the cost of a generic predictor (designed for the whole image). The partition, the specific predictors, and the residuals are encoding in a lossless way the view, conditional on another view. The end result is that the specific predictors using the warped views over many regions are able to improve significantly the compression performance, when compared to the performance of the predictors using the current view alone.

In here we expand the study of the above mentioned warping predictive scenario, by investigating the model structure selection involved in defining the partition P. For this, we study the change of the final compression ratio of a warped predictive scheme when changing the stereo matching algorithm by which the disparity image is estimated. This interest comes from the fact that disparity estimation is still an open problem and there is a plethora of currently competing methods (see [4] for a list of 150 algorithms), where the performance is expressed in percentage of pixels that have errors larger than a given threshold. It is important to find the ability of the stereo matching algorithms for producing disparity estimates that are useful in the above warped predictive scenario. The challenge in the modeling problem is finding the balance between the ability of the disparity estimation algorithm for finding each constant disparity region (corresponding to a distinct object, or part of it, located at a certain depth) and the ability of the predictive scheme of improving its performance by tuning sparse predictors for each constant disparity region, presumably part of a distinct object in the image.

In Section 2 the warped predictive compression scheme is briefly reviewed and the final codelength for encoding losslessly a color image from a stereo pair based on right-left correspondences is introduced. In Section 3 is shown the approach for ranking the stereo matching algorithms. In Section 4 an experimental study is presented, for analyzing the ranking of 50 disparity estimation algorithms, based on the the information theoretic quantities introduced in Section 3.

2. REVIEW OF THE GENERIC WARPED PREDICTIVE COMPRESSION SCHEME [3]

The goal is to encode a color image in a stereo pair, called here for exemplification the right view image, R, based on another view, called left view image, by knowing the disparity by which a pixel in L needs to be shifted to be in correspondence with a pixel in R. The disparity image...
is denoted $D^L$ (and similarly one can define a disparity $D^R$). We note that for some pixels in $R$ there is no correspondence with pixels in $L$, due to occlusions or different viewing angle, so only some regions of $D^L$ are useful for finding the warped image $R^w$, which is by definition the image obtained by translating each pixel in $L$ to the corresponding pixel in $R$, as described by the disparity $D^L$.

The coding scenario is the following: both the encoder and decoder have available one view, $L$. The encoder has access to the disparity $D^L$, from which it constructs and transmits to the decoder the warped disparity $D^w$, based on the useful parts of the disparity $D^L$. The decoder receives $D^w$ and constructs based on it a color warped image $R^w$, which is quite similar to $R$ for the locations that are attempted to be warped, and in general differs from $R$ as color values even at the warped locations. The warped predictive coding will use the causal neighbors of $(i,j)$ at the color components of $R$ and the non-causal neighbors at the color components of $R^w$ for forming long regressors, out of which optimal sparse predictor design selects the most relevant elements for linear prediction.

We call $\mathcal{P} = \{\Omega_1,\ldots,\Omega_N\}$ the partition of $D^w$ (and implicitly of $R^w$ as well) into $N$ regions, which can be given either in the form of a label image $\Omega$ having at each pixel $(i,j)$ the index $\ell$ of the region to which it belongs ($\ell = \Omega(i,j)$), or alternatively, it can be given in the form of the contours $\Gamma$ separating the regions of $R$. A contour image $\Gamma$ is encoded using the optimal context tree contour coding utilized in CERV [5], using $\mathcal{L}(\Gamma)$ bits. The model structure selection problem involved in designing $\mathcal{P}$ is discussed in Section 3.

The warped predictive coding operates then sequentially componentwise over the RGB components of the image $R$, encoding first the component $r_R$, then $g_R$, and finally $b_R$. At each pixel $(i,j)$ from the component $c \in \{r,g,b\}$ to be encoded, the prediction uses a causal template within the component $c$, and a non-causal template in the components already encoded. The regressor vector $\varphi^c(i,j)$ for the prediction of $(i,j)$ from the component $c \in \{r,g,b\}$ is thus formed of the values of the components in a causal template and in several noncausal templates. As an example, for computing the linear prediction $g_R(i,j) = (\theta^\gamma)^T \varphi^g(i,j)$, using the vector of parameters $\theta^\gamma$ one collects a regressor vector $\varphi^g(i,j)$ including the elements of the causal template $[g_R(i,j-1), g_R(i-1,j-1), g_R(i-1,j+1), g_R(i+1,j-1), g_R(i+1,j+1)]$ and the elements from the $3 \times 3$ non-causal template around $r_R(i,j)$ (that is already encoded). Additionally, the regressor vector $\varphi^g(i,j)$ includes the $3 \times 3$ non-causal templates around the components $r_R(i,j)$, $g_R(i,j)$, and $b_R(i,j)$, of the warped color image $R^w$.

The prediction templates become very large: for predicting $r_R(i,j)$ the regressor vector $\varphi^r(i,j)$ has 31 entries, for predicting $g_R(i,j)$ the regressor $\varphi^g(i,j)$ has 40 entries, and for $b_R(i,j)$ the regressor $\varphi^b(i,j)$ has 49 entries. Increasing the size of the individual templates to $5 \times 5$ will results even in much longer regressor vectors. Here comes the second model selection problem: design a sparse predictor for each component, to be used for all pixels of a given region, $\Omega_k$. In [3] we use a sparse design method faster than the greedy LS method, but having quite similar performance (see [3] for details). For example, the predictor $\hat{r}_R(i,j)$ will select $k^r$ components out of the template selected as above. We call the vectors of parameters for the three predictors over the region $\Omega_k$ as $\theta^r(\Omega_k), \theta^g(\Omega_k), \theta^b(\Omega_k)$, and for brevity we put $\Theta(\Omega_k) = \{\theta^r(\Omega_k), \theta^g(\Omega_k), \theta^b(\Omega_k)\}$. We simply denote $\Theta = \{\Theta(\Omega_1), \ldots, \Theta(\Omega_N)\}$. The cost of encoding each sparse predictor involves transmitting the mask of selected elements in the regressor, and the values of the coefficients (with the fractional part rounded to a precision of 16 bits) used in the linear prediction.

The residuals $e^c(i,j) = g_R(i,j) - \hat{g}_R(i,j)$ of each predictor, say $\theta^c(\Omega_k)$, over its region $\Omega_k$ are encoded by using a simple context encoding scheme, similar to the one used in JPEG-LS for encoding its residuals over the gradient based contexts.

Finally, encoding the color image $R$ using warped predictive coding is summarized as:

$$\mathcal{L}(R; \mathcal{P}, \Theta) = \mathcal{L}(\Gamma(\mathcal{P})) + \sum_{k=1}^{N} \mathcal{L}(D^w_{\Omega_k}) + \sum_{k=1}^{N} \mathcal{L}(\Theta(\Omega_k))$$

$$+ \sum_{k=1}^{N} \sum_{(i,j) \in \Omega_k} \mathcal{L}(e^c(i,j)) + \mathcal{L}(e^g(i,j)) + \mathcal{L}(e^b(i,j))$$

where the term $\mathcal{L}(\Gamma(\mathcal{P}))$ accounts for encoding the contours of the partition $\mathcal{P}$ (encoded as in Algorithm D of [5]), the term $\mathcal{L}(D^w_{\Omega_k})$ encodes the (constant) disparity value over the region $\Omega_k$ (encoded as in Algorithm Y of [5]), and the term $\mathcal{L}(\Theta(\Omega_k))$ accounts for encoding the predictor parameters.

### 3. RANKING STEREO MATCHING ALGORITHMS BASED ON INFORMATION THEORETIC QUANTITIES

In [3] we have studied the above scheme, where the partition $\mathcal{P}$ was obtained by considering the regions of constant disparity in the ground truth disparity images provided in the Middlebury database. The warped sparse predictive scheme has produced significant compression ratio improvements, compared with the sparse predictive encoding without considering the warped image components in the regressors, and where implicitly the whole image is just the only region. However, we note that the ground truth image having such a high quality as provided in the Middlebury data set is not currently available with the common depth acquisition sensors (e.g., Kinect sensor), nor is it available with the highly sophisticated disparity image estimation algorithms existing in the literature.

In here we address a further question: what is in practice achievable with the disparity image estimates given by the more than 150 different stereo matching algorithms benchmarked in Middlebury database.

Hence we repeat the warped sparse design for the partitions induced by each stereo matching algorithm benchmarked in [4]. For the $k$th algorithm, that produces a
Table 1. Spearman rank correlation of the variable $p_{n-o}$ with the variables $MSE_D$, $MAE_D$, $MSE_{LR}$, $L$, and $A$ (all which require the ground truth disparity to be known) and additionally with the variables $MSE_{LR}$, $L(R; P, \Theta))$ and $A(R; P, \Theta))$ (that do not require the ground truth disparity to be known) when ranking the 50 stereo matching algorithms. The ground truth disparity is the one given in [4].

<table>
<thead>
<tr>
<th>Image</th>
<th>$\rho(p_{n-o}, MSE_D)$</th>
<th>$\rho(p_{n-o}, MAE_D)$</th>
<th>$\rho(p_{n-o}, MSE_{LR})$</th>
<th>$\rho(p_{n-o}, L)$</th>
<th>$\rho(p_{n-o}, A)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>“teddy”</td>
<td>0.86</td>
<td>0.95</td>
<td>-0.02</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>“cones”</td>
<td>0.86</td>
<td>0.93</td>
<td>0.09</td>
<td>0.74</td>
<td>0.58</td>
</tr>
</tbody>
</table>

disparity image $D^k$ (publicly available in [4]), where the constant disparity regions form the partition $P$, we record the overall compression, $L(R; P, \Theta)$ (given by (1)) of the right view, resulted with sparse predictive coding using the warped right view $R^w$. The values for each algorithm are shown in Figure 2.

In addition we also consider the final area, $A$, of those constant disparity regions where the sparse predictors specifically designed for each region outperformed the general predictor, in terms of compression ratio. We denote $A(R; P, \Theta) = 1 - A_1/n_o n_s$, the missing area in percents (shown in Figure 3). This has the information theoretic significance, of percentage of the pixels in the image where the sparse class of models using warping was not better than the sparse class of models not using warping, in terms of description length measured by an implementable coding method.

A typical algorithm for stereo matching estimates the disparity $D$ by minimizing an energy function, e.g., in [6] the energy is

$$E(D) = \sum_{i,j} \text{Score}(L, R(D(i,j))) +$$

$$+ \sum_{(i',j') \in N(i,j)} \lambda_1 |D(i',j') - D(i,j)| = 1$$

$$+ \sum_{(i',j') \in N(i,j)} \lambda_2 |D(i',j') - D(i,j)| > 1$$  (2)

The $\text{Score}(L, R(D(i,j)))$ is computed based on right-left matching, as dictated by the disparity $D(i,j)$. The smoothness of $D$ is enforced by the two terms weighted with $\lambda_1$ and $\lambda_2$.

We briefly rephrase the significance of the terms in (1): the term $L(\Gamma(P)) + L(D^w)$ encourages the smoothness of the disparity $D$ (of the partition $P$); the matching cost right-left comprises the rest of the terms in (1). Hence the costs in (1) and (2) have a close correspondence. Therefore it is justified to attempt ranking the disparities by their overall performance (1).

4. EXPERIMENTAL RESULTS

We consider in this section the application of the two information-theoretic quantities for studying the disparity maps $D_1, \ldots , D_{n_o}$ saved in the benchmark database for $n_o = 50$ stereo matching algorithms (we experiment with the images “teddy” and “cones”). The database provides a “ground truth” disparity map $D^*$, obtained experimentally using projection of structured light over the scene.

We first illustrate in Figure 1 the ranking of the algorithms in the disparity domain, when comparing with the ground truth disparity, as recommended in [4], by using the proportion $p_{n-o}$ of non-occluded pixels $(i,j)$ at which $|D_k(i,j) - D^*(i,j)|$ exceeds the acceptable threshold of one. The evaluation is done only over the non-occluded part $\Omega_{\text{nonocc}}$ of the image formed of the $N_n = |\Omega_{\text{nonocc}}|$ pixels where there is right-left correspondence (no occlusion occurs), which is also provided in the benchmark. Additionally we plot in Figure 1 the mean square error in disparity domain

$$MSE_D(k) = \sum_{(i,j) \in \Omega_{\text{nonocc}}} |D_k(i,j) - D^*(i,j)|^2 / N_n$$

and the mean absolute error in disparity domain,

$$MAE_D(k) = \sum_{(i,j) \in \Omega_{\text{nonocc}}} |D_k(i,j) - D^*(i,j)| / N_n$$.

Ranking of the algorithms by the three criteria differs, however, the Spearman rank correlation coefficients $\rho(p_{n-o}, MSE_D)$ and $\rho(p_{n-o}, MAE_D)$ between the ranking provided by $p_{n-o}$ and the rankings based on $MSE_D$ and $MAE_D$ are quite strong, as presented in the second and third column of Table 1. Of the three criteria, $p_{n-o}$ seems more widely accepted in computer vision studies, since it reflects the number of useful disparity pixels (once an error in disparity is larger than 1, that disparity becomes useless in warping) and for this reason we take $p_{n-o}$ as a reference ranking criterion in the following. The indices of the algorithms in the following are chosen so that their $p_{n-o}$ is ordered increasingly. The sum of square distances $MSE_D$ and the sum of absolute distances $MAE_D$ are very often used as a score function in (2) and therefore they are directly minimized as part of the energy in (2) for some of the algorithms. This explains the overfitting with respect to these two criteria by several algorithms, that succeed in having low $MSE_D$ or $MAE_D$, although their $p_{n-o}$ are not very good.

All above criteria required the availability of a ground truth disparity image. A ranking criterion which does not require the ground truth $D^*$ is the right-left distortion, measured between the warped and the original image

$$MSE_{LR}(k) = \sum_{(i,j) \in \Omega} ||R(i,j) - R^w(i,j)||^2 / N_n$$  (3)

which is the most relevant measure of quality in 3D video coding, reflecting the capability of synthesizing new views
from a given view and its disparity. However, ranking the stereo matching algorithms based on $MSE_{LR}$ apparently is not correlated with the ranking by $p_{n-o}$, as seen in the fourth column of Table 1.

Finally we illustrate in Figures 2 and 3 the information theoretic rankings, based on the codelength $\mathcal{L}(R; P, \Theta)$ from (1) and on the unsuccessful area $A(R; P, \Theta)$, both of them not requiring the ground truth disparity to be known. Although the rankings differ, the Spearman rank correlation with the reference ranking of $p_{n-o}$ is reasonably high for both information theoretic quantities, as shown in the fifth and sixth columns of Table 1, being much better than the ranking $MSE_{LR}$.

5. CONCLUSION

We have shown that the ranking of the stereo matching algorithms by the proportion of “errorless” disparity pixels, $p_{n-o}$, is reasonably correlated with the squared, $MSE_{D}$, or absolute, $MAE_{D}$, distortions in the disparity domain, but unfortunately all these scoring variables require the knowledge of the ground truth disparity. When ground truth is not available, a surrogate scoring variable can be given by any of the two information theoretic quantities introduced here, $\mathcal{L}(R; P, \Theta)$ and $A(R; P, \Theta)$, that both are significantly correlated with $p_{n-o}$. Surprisingly, the most used scoring criteria used in lossy disparity coding, the right-left distortion $MSE_{LR}$ (very closely connected to the so-called view synthesis MSE), does not significantly correlate with $p_{n-o}$.

6. REFERENCES


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 \((\mathbb{A}^\mathbb{N}, \mathbb{A})\) from a finite alphabet \(\mathbb{A} \subset \mathbb{N}\). The random symbols will be denoted as \(x_k : \mathbb{A}^\mathbb{N} \ni (x_i)_{i \in \mathbb{N}} \mapsto x_k \in \mathbb{A}\), 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

\[
L(\xi_{1:k}) := \max\{m : x_{1:m} \text{ is repeated in } \xi_{1:k}\} \tag{1}
\]

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

\[
H_{\text{top}}(m|\xi_{1:\infty}) := \log \text{card}\{x_{1:m} : x_{1:m} \subset \xi_{1:\infty}\} \tag{2}
\]

where we write \(a \subset b\) 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 \(H_{\text{top}}(m|\xi_{1:\infty}) < \log(k - m + 1)\) then \(L(\xi_{1:k}) \geq m\).

In particular, using the Big O notation, we have

\[
H_{\text{top}}(m|\xi_{1:\infty}) = O(m^\beta) \Rightarrow L(\xi_{1:m}) = \Omega((\log m)^{1/\beta}) \tag{3}
\]

\[
L(\xi_{1:m}) = O((\log m)^{1/\beta}) \Rightarrow H_{\text{top}}(m|\xi_{1:\infty}) = \Omega(m^\beta) \tag{4}
\]

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

\[
L(\xi_{1:m}) = \Theta((\log m)^{1/\beta}) \tag{5}
\]

\[
H_{\text{top}}(m|\xi_{1:\infty}) = \Theta(m^\beta) \tag{6}
\]

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 \(\mu\) is denoted as \(E_\mu\). We also use shorthand \(\mu(x_{1:m}) = \mu(\xi_{1:m} = x_{1:m})\). The block entropy of measure \(\mu\) is

\[
H_\mu(m) := E_\mu[-\log_\mu(\xi_{1:m})] \tag{5}
\]

and the entropy rate of \(\mu\) is the limit

\[
h_\mu := \inf_{m \in \mathbb{N}} \frac{H_\mu(m)}{m} = \lim_{m \to \infty} \frac{H_\mu(m)}{m} \tag{6}
\]

Now will show the mentioned result.

**Theorem 2** \(h_\mu = 0\) for a regular Hilberg process \(\mu\).
Proof: The argument involves the random ergodic measure $F = \mu(\cdot|I)$, where $I$ is the shift-invariant algebra [9, 10]. By the ergodic theorem [9], we have $\mu$-almost surely

$$H_F(m) \leq H_{top}(m|\xi_{1,\infty}),$$

(7)

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

$$h_\mu = E_\mu h_F,$$

(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 $\mu$ will be called asymptotically deterministic when each symbol $\xi_i$ is $\mu$-almost surely a function of the infinite past $\xi_{-\infty,-1}$, cf. [11]. Secondly, the process $\mu$ will be called infinitely compressible when for every universal code $C$ the compression rate $|C(\xi_{1,m})| / m$ tends to zero $\mu$-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 $\mu$ with exponent $\beta$, $\mu$-almost surely

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

(9)

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

$$H_\mu(m) = H_F(m) \leq H_{top}(m|\xi_{1,\infty}) = O\left(m^\beta\right).$$

(10)

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

$$|C(\xi_{1,m})| \geq \frac{m}{L(\xi_{1,m}) + 1} \log \frac{m}{L(\xi_{1,m}) + 1}$$

$$= \Omega\left(\frac{m}{(\log m)^{1/\beta-1}}\right).$$

(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 arbitrarily 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 $2^0, 2^1, 2^2, \ldots$ 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 $(k_n)_{n \in [0,\infty)}$ be some sequence of strictly positive natural numbers that satisfy

$$k_{n-1} \leq k_n \leq 2 k_{n-1}^2.$$  

(12)

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

$$P((L_{n,1}, R_{n,1}, \ldots, L_{n,k_n}, R_{n,k_n}) = (l_{n,1}, l_{n,2}, \ldots, l_{n,k_n}, r_{n,k_n})) = \frac{1}{k_n^2}.$$  

(13)

Subsequently we define random variables

$$Y_j^0 = j, \quad j \in \{1,\ldots,k_n\},$$  

(14)

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

(15)
where $a \times b$ denotes concatenation. Hence $Y_n^j$ are $k_n$ distinct blocks of $2^n$ natural numbers, selected by some sort of random hierarchical concatenation.

Step 2: Variables $Y_n^j$ will be the building blocks of yet another process. Let $(C_n)_{n \in \{1, \ldots, k_n\}}$ be independent random variables, independent from $(L_n^j, R_n^j)_{n \in \{1, \ldots, k_n\}}$, with uniform distribution

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

Definition 2 The random hierarchical association (RHA) process $X$ with perplexities $(k_n)_{n \in \{1, \ldots, k_n\}}$ is defined as

$$X = Y_0^1 \times Y_0^2 \times Y_2^1 \times \ldots \quad (17)$$

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 $X$ will be parsed into a sequence of numbers $X_j$, where

$$X = X_1 \times X_2 \times X_3 \times \ldots \quad (18)$$

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

$$X_{k,l} = X_k \times X_{k+1} \times \ldots \times X_l. \quad (19)$$

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 $T: \mathcal{A}^N \ni (x_i)_{i \in \mathbb{N}} \mapsto (x_{i+1})_{i \in \mathbb{N}} \in \mathcal{A}^N$. We recall this definition:

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

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

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

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

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

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

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

exist for every block $x_{1,m} \in \mathcal{A}^+$. For an AMSB measure $\nu$ over a finite alphabet $\mathcal{A}$, function $\mu$, extended via $\mu(\xi_{1,m} = x_{1,m}) := \mu(x_{1,m})$, is also a stationary measure on $(\mathcal{A}^N, \mathcal{A}^N)$. We shall continue to call this $\mu$ a stationary mean of $\nu$. However, an AMSB measure need not be AMS, cf. [15, Example 6.3]. In particular, for an AMSB measure $\nu$ 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 $m \leq 2^n$ and $k \in \mathbb{N}$, the stationary mean is

$$\mu(x_{1,m}) = \frac{1}{2^n} \sum_{j=0}^{2^n-1} P(X_k2^n+jk2^n+jm-1 = x_{1,m}). \quad (22)$$

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

$$k_n = \lfloor \exp(2^{2n}) \rfloor, \quad (23)$$

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

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

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

The measure $\mu$ 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 $F = \mu(\mathcal{I})$. The measure $F$ is $\mu$-almost surely an ergodic regular Hilberg process with the exponent $\beta$.

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

$$\limsup_{m \to \infty} |E_F[C(\xi_{1,m})] - H_F(m) - \rho(m)| > 0. \quad (25)$$

Shields’ result concerns nonexistence of a universal sublinear bound for the difference $E_F[C(\xi_{1,m})] - H_F(m)$. Some way of strengthening this result is to investigate ratio $E_F[C(\xi_{1,m})]/H_F(m)$. 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 $E_F[C(\xi_{1,m})]/H_F(m)$, 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 \(\mu\) of the RHA process with perplexities (23) and its random ergodic measure \(F = \mu(\cdot | I)\), we have
\[
E_\mu E_F \frac{|C(\xi_{1:m})|}{H_F(m)} = \Omega \left( \frac{m^{1-\beta}}{(\log m)^{\beta+1}} \right),
\]
(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
\[
E_\mu |C(\xi_{1:m})| \geq H_\mu(m).
\]
\((\star)\)

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 the following relationship
\[
H_\mu(m) = \Theta(m^3),
\]
(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
\[
2H_\mu(m) - H_\mu(2m) = \Theta(m^3)
\]
(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) and (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

DISTRIBUTED SOURCE CODING OF VIDEO

Søren Forchhammer1 and Huynh Van Luong2

1DTU Fotonik, Technical University of Denmark
2Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg

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 address 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 (4x4 DCT), quantized and decomposed into bitplanes.
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 from 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 subband. 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 codec (Figure 2).
Figure 2. TDWZ decoder with the motion and reconstruction reestimation (MORE). The combination of initial side information and re-estimated 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 Bjøntegaard 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

<table>
<thead>
<tr>
<th>Codec</th>
<th>Crossband</th>
<th>SING</th>
<th>MORE</th>
<th>MORE(AMD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman</td>
<td>0.65</td>
<td>1.52</td>
<td>3.00</td>
<td>2.93</td>
</tr>
<tr>
<td>Hall</td>
<td>0.39</td>
<td>0.99</td>
<td>1.42</td>
<td>1.95</td>
</tr>
<tr>
<td>Soccer</td>
<td>1.33</td>
<td>2.70</td>
<td>4.19</td>
<td>4.182</td>
</tr>
<tr>
<td>Coast</td>
<td>0.36</td>
<td>0.41</td>
<td>0.65</td>
<td>0.85</td>
</tr>
<tr>
<td>Average</td>
<td>0.64</td>
<td>1.49</td>
<td>2.47</td>
<td>2.58</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Codec</th>
<th>Crossband</th>
<th>SING</th>
<th>MORE</th>
<th>MORE(AMD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman</td>
<td>0.33</td>
<td>0.75</td>
<td>1.43</td>
<td>1.41</td>
</tr>
<tr>
<td>Hall</td>
<td>0.19</td>
<td>0.40</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td>Soccer</td>
<td>0.73</td>
<td>1.51</td>
<td>2.26</td>
<td>2.23</td>
</tr>
<tr>
<td>Coast</td>
<td>0.19</td>
<td>0.22</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>Average</td>
<td>0.33</td>
<td>0.76</td>
<td>1.22</td>
<td>1.22</td>
</tr>
</tbody>
</table>

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 re-
quested.

In [9], a rate-adaptive BCH (RA-BCH) code was in-
troduced 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 syn-
dromes used to confirm a decoding adaptive to the num-
er 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 $10^{-3}$, the average code lengths where $\sim 0.10$
for RA-BCH for length 1023 and $\sim 0.144$ for LDPCA of
length 1584. For fixed rate coding, a bound of $\sim 0.146$
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 to-
wards 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

information sources", IEEE Trans. Inf. Theory, vol. 19,
source coding with side information at the decoder",
"Distributed video coding", Proc. IEEE, vol. 93, no. 1,
http://www.discoverdvc.org/
refinement for transform domain Wyner-Ziv video cod-
[6] H. V. Luong, L. L. Rakêt, X. Huang, and S. Forchham-
mer, "Side Information and Noise Learning for Distribut-
2012.
[7] H. V. Luong, L. L. Rakêt, and S. Forchhammer, "Re-
estimation of Motion and Reconstruction for Distributed
[8] H. V. Luong, S. Forchhammer, J. Slowack, J. De Cock,
and R. Van de Walle, "Adaptive Mode Decision with Re-
sidual Motion Compensation for Distributed Video Cod-
ing," 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,”
codes for distributed source coding,” EURASIP Signal
ert, A. Munteanu, and R. Van de Walle, “Rate-distortion
driven decoder-side bitplane mode decision for distributed
(2010), 660–673.
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, and 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 $P$ denote the set of preparations. Let $p_1$ and $p_2$ denote two preparations. For $t \in [0, 1]$ we define $(1 - t) \cdot p_1 + t \cdot p_2$ 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 $M$ denote a set of feasible measurements. The state space $S$ 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 $m \in M$.

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 $A$ denote a subset of the feasible measurements $M$
such that $a \in A$ maps $S$ into a distribution on the real
numbers i.e. a random variable. The elements of $A$ may
represent actions like the score of a statistical decision,
the energy extracted by a certain interaction with the sys-
tem, (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 $s \in S$ we define
$$F(s) = \sup_{a \in A} E[a(s)].$$
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
$$D_F(s_1, s_2) = F(s_1) - \sup_{(a_n)_n} \lim_{n \to \infty} \sup E[a_n(s_1)]$$
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 $s_1 = s_2$.
- $\sum t_i \cdot D_F(s_1, \tilde{s}) \geq \sum t_i \cdot D_F(s_1, \tilde{s}) + D_F(\tilde{s}, \tilde{\tilde{s}})$
  where $(t_1, t_2, \ldots, t_\ell)$ is a probability vector and $\tilde{s} = \sum t_i \cdot s_i$.
- $\sum t_i \cdot D_F(s_1, \tilde{s})$ is minimal when $\tilde{s} = \sum t_i \cdot s_i$.

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
$D_F(s_1, s_2) = E[a_2(s_1)] - E[a_3(s_1)]$. If unique op-
timal 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
$$D_F(s_1, s_2) = F(s_1) - (F(s_2) + \langle s_1 - s_2, \nabla F(s_2) \rangle).$$

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
$$F(s_1) = D_F(s_1, s_2) + E[a_2(s_1)]$$
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 $(s_0)_a$ denote a family of states and let $\Phi$ denote a completely positive transformation $S \to T$ where $S$ and $T$
state spaces. Then $\Phi$ is said to be sufficient for $(s_0)_a$ if there exists a completely positive transformation
$\Psi : T \to S$ such that $\Psi(\Phi(s_0)) = s_0$.

We say that the regret $D_F$ on the state space $S$ sat-
sifies the sufficiency property if $D_F(\Phi(s_1), \Phi(s_2))$
for any completely positive transformation $S \to S$
that is sufficient for $(s_1, s_2)$ . The notion of suffi-
cency as a property of divergences was introduced in [10]. The

Theorem 3. Assume that $S$ is a state space. If the diver-
gence $D_F$ 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
$$D_F(\Phi(s_1), \Phi(s_2)) = D_F(s_1, s_2).$$

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

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

5. APPLICATIONS

5.1. Statistics

Consider an experiment with $X = \{1, 2, \ldots, \ell\}$ as sample
space. A scoring rule $f$ is defined as a function with domain
$X \times M^+_1(X) \to \mathbb{R}$ such that the score is $f(x, Q)$
when the prediction was given by $Q$ and $x \in X$ has been
observed. A scoring rule is proper if for any probability
measure $P \in M^+_1(X)$ the score $\sum_{x \in X} P(x) \cdot f(x, Q)$
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 $f(x, Q) =
D_F(\delta_x, Q) + f(x)$.

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 scor-
ing 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 condi-
tion 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 \( \kappa \). 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 \( \ell \) are allowed then \( b \) 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 fixed under all interactions \( F \) is called Helmholtz free energy. Any sufficient transformation \( \Phi \) for \( s_1 \) and \( s_2 \) is quasi-static and can be approximately realized by a physical process \( \Psi \) that is reversible in the thermodynamic sense of the word.

\[
D_F (\Phi (s_1), \Phi (s_2)) = a_{\Phi (s_1)} (\Phi (s_1)) - a_{\Phi (s_2)} (\Phi (s_1)).
\]  

(5)

Now

\[
a_{\Phi (s_2)} (\Phi (s_2)) = (a_{\Phi (s_2)} \circ \Phi) (s_2) \\
\leq a_2 (s_2) = a_2 (\Psi (\Phi (s_2))) \\
= (a_2 \circ \Psi) (\Phi (s_2)) \leq a_{\Phi (s_2)} (\Phi (s_2)).
\]  

(6)

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

\[
D_F (\Phi (s_1), \Phi (s_2)) \\
= (a_1 \circ \Psi) (\Phi (s_1)) - (a_2 \circ \Psi) (\Phi (s_1)) \\
= a_1 (s_1) - a_2 (s_1) = D_F (s_1, s_2).
\]  

(7)

The amount of extractable energy \( E_x \) 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

\[
E_x = kT \cdot D (s_1 \| s_2)
\]  

(8)

where \( k = 1.381 \cdot 10^{-23} \cdot J/K \) 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 \( b_2 = 0.3 \) 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

$$S_n = \prod_{i=1}^{n} \langle \vec{X}_i, \vec{b} \rangle$$  \hspace{1cm} (9)

According to law of large numbers

$$\frac{1}{n} \log (S_n) \to E \left[ \log \left( \frac{\vec{X}}{\vec{b}} \right) \right]$$  \hspace{1cm} (10)

Here $E \left[ \log \left( \frac{\vec{X}}{\vec{b}} \right) \right]$ 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 is to maximize $W \left( \vec{b}, P \right)$ by choosing an appropriate portfolio $\vec{b}$.

Let $b_P$ denote the portfolio that is optimal for $P$. As proved in [7]

$$W \left( \vec{b}_P, P \right) - W \left( \vec{b}_Q, P \right) \leq D \left( P \parallel Q \right).$$  \hspace{1cm} (11)

**Theorem 9.** The Bregman divergence

$$W \left( \vec{b}_P, P \right) - W \left( \vec{b}_Q, P \right)$$  \hspace{1cm} (12)

satisfies the equation

$$W \left( \vec{b}_P, P \right) - W \left( \vec{b}_Q, P \right) = D \left( P \parallel Q \right).$$  \hspace{1cm} (13)

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

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 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


ITERATIVE DECODING OF PRODUCT CODES

Jørn Justesen
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 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 $t = [(d - 1)/2]$ 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 $r$ 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 $2j$ 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

$$\left\lfloor \sum \deg(v) - e(S', S'') \right\rfloor / 2$$

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 $2E$, we can write the condition as

$$\sum_{S'+S''} \deg(v) - \left\lfloor \sum_{S'} \deg(v) + e(S', S'') + jd/2 \right\rfloor \geq 0$$
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

\[
\sum_{j=0}^{\infty} \binom{n}{j} 2^{j/2} p^{j/2}
\]

This probability is very small for \(p < d/(2n)\) and the largest term occurs for \(j=d\). This term can 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 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 columns by decoding errors. However, the probability of decoding error is low in these cases. However, the important part of the algorithm is decoding the critical subset, where most of the rows and columns are chosen to have error patterns of low weight.

If a component decoder correctly locates the error positions in a row, the corresponding messages from the columns may be missing, since a significant fraction 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 few messages are 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 \(t+1\) 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


CAUSALITY AND DIRECTED INFORMATION ESTIMATION
AS A HYPOTHESIS TEST

Ioannis Kontoyiannis1 and Maria Skoularidou2

1Department of Informatics, Athens University of Economics and Business, Patission 76, Athens 10434, Greece. yiannis@auieb.gr
2Department 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 \( \chi^2 \) 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 \( \chi^2 \) 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 \( \chi^2 \) test, we compute the normalized \( \chi^2 \) distance between these two distributions,

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

Under the null hypothesis – assuming, that is, that the data are independent – the distribution of the statistic \( \chi^2_n \) for large \( n \) is approximately \( \chi^2 \) with \((mn-1)(\ell-1)\) degrees of freedom, where \( m, \ell \) are the sizes of the alphabets of \( X \) and \( Y \), respectively. Therefore, computing the probability of the statistic \( \chi^2_n \) 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 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, \( \Delta_n \) has the exact same distribution as \( \chi^2_n \), 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,

\[
\Delta_n = 2n I(\hat{X}; \hat{Y}) = 2n D(\hat{P}_{XY,n} \parallel \hat{P}_{X,n}\hat{P}_{Y,n}), \tag{2}
\]

where \( \hat{X}, \hat{Y} \) are distributed according to the empirical distribution \( \hat{P}_{XY,n} \). One way to look at the difference between \( \chi^2_n \) and \( \Delta_n \) is that, instead of the \( \chi^2 \) 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 \( \Delta_n \) 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 \( \Delta_n \) 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 \( X = \{X_n\} \) and \( Y = \{Y_n\} \) are two finite-valued random process, then the directed information \( I(X^n_1 \rightarrow Y^n) \) between \( X_n = (X_1, X_2, \ldots, X_n) \) and \( Y^n = (Y_1, Y_2, \ldots, Y_n) \) is defined as,

\[
I(X^n_1 \rightarrow Y^n) = H(Y^n_1) - \sum_{i=1}^{n} H(Y_i|Y_i^{i-1}, X_i), \tag{3}
\]

and the directed information rate between \( X \) and \( Y \) is,

\[
I(X \rightarrow Y) = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n_1 \rightarrow Y^n), \tag{4}
\]

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 $X^n = (X_1, X_2, \ldots, X_n)$ 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 $X^i_j = (X_i, X_{i+1}, \ldots, X_j)$, $i \leq j$, for vectors of random variables and similarly $a^i_j = (a_i, a_{i+1}, \ldots, a_j) \in A^{i-j+1}$, $i \leq j$, 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 $X = \{X_n : n \geq 0\}$ and $Y = \{Y_n : n \geq 0\}$ 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 $I(X^n \to Y^n)$ in (3).

This is zero exactly when $Y_i$ is conditionally independent of $X_i$, given its past $Y_{i-1}^k$, for each $i = 1, 2, \ldots, n$. The natural interpretation of this equivalence is to say that the directed information is zero if and only if $X$ has no causal influence on $Y$. We are interested in the problem of estimating the directed information rate, $I(X \to Y)$, 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 $\{(X_n, Y_n)\}$ 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 $H(Y)$ of the univariate process $Y = \{Y_n\}$ exists and,

$$H(Y) = \lim_{n \to \infty} \frac{1}{n} H(Y^n) = \lim_{n \to \infty} \frac{1}{n} H(\tilde{Y}^n).$$

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

$$H(Y) - H(Z_0, \tilde{X}_k, \tilde{Y}_{-k}^{-1}),$$

where $H(Y|X) = H(X, Y) - H(X)$ denotes the conditional entropy.
(iii) If $Y = \{Y_n\}$ is also a Markov chain of order no larger than $k$, then,

$$I(X \rightarrow Y) = I(\bar{Y}_0; \bar{X}_0^{k-1})$$

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 $a_k \in A^{k+1}, b_k \in B^{k+1}$. As discussed below, this assumption can be significantly relaxed.

2. The directed information rate $I(X \rightarrow Y)$ admits important operational interpretations. For example, in the case of a stationary $k$th order Markov chain $\{(X_n, Y_n)\}$ such that $\{Y_n\}$ is also a 4th order chain, we can use the data processing property of mutual information in the result of part (iii) of the proposition to see that $I(X \rightarrow Y)$ equals,

$$I(Y_0; X_0^k|Y_{-k}^{-1}) - I(Y_0; X_0^k|Y_{-k}^{1-k})$$

This is zero if and only if each $Y_i$, given its past $Y_{i-1}$, is conditionally independent of $X_{i-1}$, 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,

$$I(Y_0; X_0^k|Y_{-k}^{-1}) = I(Y_0; X_0^{\infty}|Y_{-k}^{-1})$$

by data processing; this is zero if and only if $Y_0$, conditionally only on its $k$-past $Y_{-k}^{-1}$, is independent of $X_0^{\infty}$. In this case the quantity $I(Y_0; X_0^k|Y_{-k}^{-1})$ is not enough to entirely characterize the absence of causal influence from $X$ to $Y$, 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 $X$ has no (or little) causal influence on $Y$.

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 $I(Y_0; X_0^k|Y_{-k}^{-1})$. 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 $I(X \rightarrow Y)$

Given a sample $(X_n^{n+k-1}, Y_n^{n+k-1})$ 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,

$$\hat{P}_{X_n^k Y_n^k \sim n}(a_0^k, b_0^k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_{i+k-1} = a_k; Y_{i+k-1} = b_k\}}$$

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

$$\hat{i}^{(k)}_n(X \rightarrow Y) = I(\hat{Y}_0; \hat{X}_0^k|\hat{Y}_{-k}^{-1})$$

where $\hat{X}_0^k, \hat{Y}_0^k \sim \hat{P}_{X_n^k Y_n^k \sim n}$. Since all the transition probabilities of the bivariate chain $\{(X_n, Y_n)\}$ are non-zero, it is easy to see that the plug-in estimator $\hat{i}^{(k)}_n(X \rightarrow Y)$ converges almost surely to the desired value, $I(X \rightarrow Y)$. 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 $A \times B = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, l\}$, 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 $I(X \rightarrow Y) > 0$ then, as $n \rightarrow \infty$,

$$\sqrt{n} \left( \hat{i}^{(k)}_n(X \rightarrow Y) - I(X \rightarrow Y) \right) \overset{D}{\rightarrow} N(0, \sigma^2),$$

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \text{Var} \left( \log \left( \prod_{i=1}^{n} \frac{P_{X_0^k Y_0^k \sim n}(X_{i-1}^k, Y_{i-1}^k)}{P_{\hat{Y}_0^k Y_{-k}^{1-k}}(Y_{i-1}^k, Y_{i-1}^k)} \right) \right).$$

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

$$n \hat{i}^{(k)}_n(X \rightarrow Y) \overset{D}{\rightarrow} \chi^2((\ell k + k^2 - 1)(\ell - 1)).$$

**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 2.2 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 $I(X \rightarrow Y) = 0$ 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 $I(X \rightarrow Y)$ is strictly positive, and the plug-in estimator converges at the slower rate $O(1/\sqrt{n}).$

Finally, the proof of the $\chi^2$ 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 \( \{Y_n\} \). In this context, as discussed above, this translates to testing the null hypothesis that each random variable \( Y_i \) is conditionally independent of \( X_{i-k} \) given \( Y_{i-k} \), within the larger hypothesis that the pair process \( \{(X_n, Y_n)\} \) is a 4th 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 \( \theta \) taking values in an \( m^k \ell (m^\ell - 1) \)-dimensional open set \( \Theta \). Informally, the null hypothesis corresponding to each random variable \( Y_i \) being conditionally independent of \( X_{i-k} \) given \( Y_{i-k} \), is described by transition matrices \( Q_\theta \) which can be decomposed as,

\[
Q_\theta(a_0, b_0|a_{-k}^{-1}, b_{-k}^{-1}) = Q_\theta^0(a_0|a_{-k}^{-1}, b_{-k}^{-1}) Q_\theta^0(b_0|b_{-k}^{-1}).
\]

Formally, this can be described by a lower-dimensional parameter set \( \Phi \), which will be embedded in \( \Theta \) via a map \( h : \Phi \rightarrow \Theta \), 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 \( \Phi \) within the general model \( \Theta \), we employ a likelihood ratio test. Specifically, we define the log-likelihood \( L_n(X_n^{k+1}, Y_n^{k+1}|\theta) \) of the sample \( \{X_n^{k+1}, Y_n^{k+1}\} \) under the distribution corresponding to \( \theta \),

\[
\log \left[ P_\theta(X_1^n, Y_1^n | X_0^{k+1}, Y_0^{k+1}) \right],
\]

so that the likelihood ratio test statistic is simply,

\[
\Delta_n = 2 \max_{\theta \in \Theta} L_n(X_n^{k+1}, Y_n^{k+1}|\theta) - \max_{\phi \in \Phi} L_n(X_n^{k+1}, Y_n^{k+1}|h(\phi)),
\]

The key observation here is that:

\[
\Delta_n = 2n \hat{I}_n^{(k)}(X \rightarrow Y).
\]

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 \( \Delta_n \) is approximately \( \chi^2 \) with \( \ell k(m^\ell - 1)(\ell - 1) \) 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 \( \Delta_n \) 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

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 independent 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 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 \textit{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” \(e\) such that \(x + e = e + x = x\) for all \(x \in G\), and \(x + (-x) = (-x) + x = e\) for all \(x \in G\). For any \(f : G \to \mathbb{R}\), define the \textit{entropy} of \(f\) by

\[ h(f) := -\int_G f(x) \log f(x) dx \]

where \(dx\) is the Haar measure on \(G\). The differential entropy of a random vector \(X\) with density \(f(x)\) on \(G\) is defined by

\[ h(X) = h(f) := -\int_G f(x) \log f(x) dx \]

where \(dx\) is the Haar measure on \(G\).
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 $\lambda$ that is translation-invariant, i.e., such that

$$\lambda(A + x) = \lambda(A) \quad \forall A \subseteq G, \forall x \in G$$

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 $\lambda$ be a Haar measure on $G$. If $\mu$ is a probability measure on $G$ that is absolutely continuous with respect to $\lambda$, then there exists a nonnegative function $f : G \to \mathbb{R}$ with total integral of 1 such that

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

which we call the density (or probability density function) of $X$ or $\mu$. 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 $\lambda$ is finite, and so we can normalize it to get the “uniform” probability measure on $G$. Then, for every $G$-valued random variable $X$,

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

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

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 $X_1 \leftrightarrow Y \leftrightarrow X_2$ 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 $X_1 \leftrightarrow Y \leftrightarrow X_2$ 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 $h(Y + Y') \leq h(Y + Y' + Z),$

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

Also the Ruzsa triangle inequality can be rewritten:

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

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 \(\Delta_+ / \Delta_-\) 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. We believe that these results and implications for probability, convex geometry, and geometric functional analysis. We believe that these results 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


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.

For the simplest SCOT of Fig. 1 contexts are \{0\}, \{01\}, \{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 bifaction 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 \((1/2,1/4,...,(1/2)^{n-1},(1/2)^n)\), and entropy rate (ER) for SCOT reduced to 1-MC between contexts is by the well-known formula for 1-MC: \(ER=\sum_{i\in A} \sum_{j:A} q_{ij}p_{ij}\log p_{ij}\) 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:
\[ P(Y_{n+1}|Y_n) = P(x_{n+m}|Y_n), \text{ if } x_{n+1}, \ldots, x_{n+m-1} \text{ coincide in both sides, and } 0 \text{ otherwise.} \]

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

\[ C(x_0) = x_{-1}, \ldots, x_{-k}, k \leq m := x_{-k}^k; \quad (1) \]

the top part of the preceding \( m \)-gram of minimal length such that the conditional probability

\[ P(x_0|x_{-j}^{-1}) \equiv P(x_0|x_{-j}^{-1}), \forall r > k; k = |C(x_0)| \quad (2) \]

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 of \( T \) whose root is \( a_i \), thus \( T^*_i := \{u|u \circuit \in T^*\}; \) For all pairs \( a_i \in A, a_j \in A, T_{ij} : = T(T_{ij}), \) thus we have \( T^*_j := \{u|u \circuit \in T^*\}. \)

A complete context tree \( T \) is called "tailclosed" if \( \forall c \in T^*, i \in \{1, \ldots, d\}, \exists u \in T^*, \text{ s.t. } \circuit \in \circuit \), 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) \( \forall 1 \leq i, j \leq d, T_{ij} \subseteq T_i \)
(c) \( \forall 1 \leq i, j \leq d, c \in T_{ij}, \exists c' \in T^*_i, \text{ s.t. } c' = \circuit \), where \( u \) is a string.
(d) \( \forall 1 \leq i, j \leq d, c' \in T^*_i, \exists c \in T_{ij}, \text{ s.t. } c' = \circuit \), 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 \( [\mathcal{F}(n)] \times [\mathcal{F}(n)] \) matrix \( P \) of \( m \)-MC transition probabilities between contexts.

In our first example, this definition gives the following transition probability matrix \( P \) between contexts:

<table>
<thead>
<tr>
<th>( P )</th>
<th>( (0,0) )</th>
<th>( (0,1) )</th>
<th>( (1,1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

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 \( P \), we approximate the theoretical stationary distribution better.

Namely, the Euclidean norm of the approximation error shrinks exponentially with the power of \( P. \) If a context probability shrinks, we remove it.

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

\[ Q^* P = Q^*. \quad (3) \]

If the induced \( 1\)-MC is ergodic, then **Ergodic theorem** holds: The solution to (1) exists, is unique, and iterations \( Q_m = Q_0 P^m \) 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 \( \pm 1 \) 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 \( \pm \) Poisson processes.

The family \( X_N^n = \sum_{i=1}^{n} r_i \) of Spike models is a reflected Random walk on large interval \([-l, l], l > N^{1/2} \). Regular part of \( X_N^n \) has increments \( \pm 1 \) 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 \( X^*_N = X^*_{N-1} \), or \( X^*_N > X^*_{N-1} \), or the opposite inequality holds.

\[ X_N^n = \sum_{i=1}^{n} r_i, r_i \text{ in a regular part is an equally likely sequence of independent identically distributed (IID) } \pm 1, i = 1, \ldots, \text{ inside } (-l,l), \text{ while irregular part is a SCOT model specified above.} \]

### 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 \([Nt]\) of \( Nt \)). We study the weak convergence of \( w_N(t) \) as \( N \to \infty \).

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

\[ E(X_{k+1} - X_k) = 0, \]
Var(\(X_{k+1}^N - X_k^N\)) = (1 - 2/N)/N.

Let \(\tau_k\) be the k-th spike time. Obviously, \(\tau_k - \tau_{k-1}\) are IID, independent of \(\sigma\)-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 independent of \(w^N(t)\) compound Poisson process of equally likely magnitudes \(\pm 1\):

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

\(\tau\) and \((x_j), t < \tau\), 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 \(u^N(t)\) be as above. Then \(E r_i = 0). Var[\(u^N(t)\)] = \(N t_i^2/[a^2N^{2b-1}N^{-2b} + (1 - N^{-2b})] \rightarrow a^2 + 1\). The equality of summands preceding a spike can be neglected. The covariance of \(u^N(t)\) 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.** \(u^N(t)\) 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 \(\{y_t\}\). If m-MC \(X_S\) is ergodic with finite alphabet and the sequence \(\{Y_t\}\) is an ergodic 1-MC, then:

- this 1-MC \(\{Y_t\}\) is a Harris invariant ([9], chapter 17), with respect to a probability distribution. Let \(g\) be a Borel function on \(R\).

- Define \(f(Y_t) := f(x_{i-1}, \ldots, x_{i-m+1}) = \sum \delta_{g(x_{i-k})}\).

- Define \(\bar{f}_N := (1/N) \sum f(Y_t), \bar{g}_N := (1/N) \sum g(x_i)\).

- If \(\pi\) is the stationary distribution and \(E_x|f|^2 < \infty\), then the ergodic theorem ([9], 17.3) guarantees that \(\bar{f}_N \rightarrow E_xf\) with probability 1 as \(N \rightarrow \infty\) and the central limit theorem holds for \(\bar{f}_N\) ([9], 17.4):

\[
\sqrt{N}(\bar{f}_N - E_x f) \Rightarrow N(0, \sigma^2_f) \text{ weakly,}
\]

\[
\sqrt{N}(1/\sigma_x^2_f f(Y_t)) \Rightarrow N(0, \sigma_x^2_f) \text{ weakly,}
\]

\[
\sqrt{N}(\sigma_{Y^{-\infty}} g(x_{i-k}) - E_x f) \Rightarrow N(0, \sigma^2_f) \text{ weakly,}
\]

\[
\sqrt{N}(m\bar{g}_N(x) - E_x f) \Rightarrow N(0, \sigma^2_f) \text{ weakly.}
\]

- 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:

\[
P(N^{-1/2}(\sum x_i) = \Phi_2(x) + \Phi(q(x))N^{-1/2} + O(N^{-1})
\]

and finds explicit expression for the Hermite polynomial \(q(x)\). Here \(\phi\) and \(\Phi\) are pdf and CDF of the central Normal RV with StD \(\sigma\).

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 divergence (cross entropy) \(D(y_i || P_0) = E_1 log(P_1/P_0)\) 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 \(0 < \epsilon < 1\), then the error probabilities of Likelihood Ratio Test (LRT) satisfy simultaneously

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

and

\[
\lim P_1(L_0 - L_1 > n\lambda)) \geq 1 - \epsilon > 0.
\]

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

10. Nonparametric version of the ‘Stein’ lemma

Generate an artificial \(n\)-sequence \(z^N\) independent of \(y^N, z^N\) 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 \(T\) is \(T > T_0 > N\lambda\). 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:

\[
ESI(s) = \sum_{i \in A} \sum_{f \in A} N(i.s.f) \cdot \log_2 \left( \frac{N(i.s.f) \cdot N(s)}{N(i.s) \cdot N(x.f)} \right)
\]

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

![Matrix showing ESI calculation](image)

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 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:

\[
P(i|s) = \frac{N(s.i)}{N(s)}
\]

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

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


RELEVANCE SAMPLING

Mads Nielsen1, Bo Markussen2 and Marco Loog3

1Department of Computer Science, University of Copenhagen,
Sigurdsgade 41, 2200 Copenhagen N, DENMARK, madsn@di.ku.dk
2Department of Mathematical Sciences, University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen OE, DENMARK, bomar@math.ku.dk
3Pattern 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 ˜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 (ΩS × ΩX × ΩY , FS ⊗ FX ⊗ FY , 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:

\[ X \rightarrow S \rightarrow Y. \]

Now let s ∈ ΩS 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 ˜X ∈ ΩX to model the situation that an observation is thrown away. The probability of retaining the observation x ∈ ΩX is denoted by M(x) ∈ [0, 1]. If ˜X denotes the outcome of this procedure, then we have the Markov diagram

\[ X \rightarrow ˜X \rightarrow Y \]

with Ω ˜X = ΩX ∪ {∇} and transition probabilities from X to ˜X given by

\[ p(˜x|x) = \begin{cases} M(x) & \text{for } ˜x = x, \\ 1 - M(x) & \text{for } ˜x = ∇, \\ 0 & \text{otherwise}. \end{cases} \]

The probability of not seeing anything is given by p(∇) = 1 − ∫ΩX M(x)p(x)dx, and the sampling density on the original sample space is given by

\[ p(x) = P(˜X = x|X ≠ ∇) = \frac{M(x)p(x)}{1 - p(∇)}. \]

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

\[ L(x) = \frac{M(x)}{1 - p(∇)}. \]

The conditional density of Y given ˜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 ˜X = ∇ is given by

\[ p(y|∇) = \int_{ΩX} p(y|x)p(x|∇)dx = \frac{p(y)}{p(∇)} - \int_{ΩX} p(y|x)M(x)p(x)dx \frac{p(∇)}{p(∇)}. \]
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
\[
\int_{\Omega_x} p(y|x) \frac{M(x)p(x)}{1 - p(y)} \, dx = \frac{1 - p(y)p(\tilde{p}(x)/p(y))}{1 - p(y)} \, p(y).
\]
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
\[
M^*(x) = \arg \min_{M(x) \in [0,1]} I(\tilde{X}, X) - \beta I(\tilde{X}, Y). \tag{1}
\]

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
\[
R_{Y|X}(x) = H(Y) - H(Y|X = x) = \int_{\Omega_y} p(y|x) \log \frac{p(y|x)}{p(y)} \, dy.
\]

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
\[
M(x) = 1 - \min \left\{ 1, \frac{p(\nabla)p(x)}{p(x)} e^{-\beta R_{Y|X}(x)} \right\}, \quad p(\nabla) = \int_{\Omega_x} \min \left\{ p(x), p(\nabla) e^{-\beta R_{Y|X}(x)} \right\} \, dx.
\]

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
\[
H(X) - \beta I(X, Y) - p(\nabla) \log p(\nabla) + \int_{\Omega_x} p(\nabla, x) \left( \log p(\nabla, x) + \beta R_{Y|X}(x) \right) \, dx
\]
with \( p(\nabla, x) = p(x) - M(x)p(x) \). Using \( \frac{\delta \mathcal{L}}{\delta p(\nabla, x)} = 1 \) 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 \( p(\nabla) e^{-\beta R_{Y|X}(x)} \). 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
\[
p(\nabla) = \int_{\Omega_x} p(\nabla, x) \, dx = \int_{\Omega_x} \left( p(x) - M(x)p(x) \right) \, dy
\]
\[
= 1 - \int_{\Omega_x} p(x) \left( 1 - \min \left\{ \frac{p(\nabla)}{p(x)} e^{-\beta R_{Y|X}(x)} \right\} \right) \, dx
\]
\[
= \int_{\Omega_x} \min \left\{ p(x), p(\nabla) e^{-\beta R_{Y|X}(x)} \right\} \, dx \overset{\text{def}}{=} F(p(\nabla)).
\]

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. \( \square \)

### 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 \( (\Omega_Q, F_Q, P) \) 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
\[
M(q, x) = M_Q(q), \quad p(\nabla, q, x) = p(\nabla, q) p(x|q).
\]

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

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.i.d. measurements \( X_1, \ldots, X_n \). 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

\[ L(x) = \frac{p(x)}{\tilde{p}(x)} = \frac{M(x)}{1 - p(\nabla)}, \]

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

\[ T_{RS}(y_0) = \inf_{y \neq y_0} \sum_{i=1}^{n} L(x_i) \log \frac{p(y_0|x_i)}{p(y|x_i)}, \]

\[ N_{RS} = \sum_{i=1}^{n} L(x_i). \]

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 measures, i.e.

\[ \int_{\Omega_X} L(x)p(x) \, dx = \int_{\Omega_X} M(x)p(x) \, dx = 1. \]

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 \( X_1, \ldots, X_n \) are weighted according to associated features \( Q_1, \ldots, Q_n \). The feature weighted log likelihood ratio test statistic with sampling measure \( L_Q(q) = \frac{M_Q(q)}{1-p(\nabla)} \) 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

\[ p(x,y) = \frac{e^{-\frac{1}{2}(x-y)^2}}{2\sqrt{2\pi}}, \quad Q_i = 1_{\{|x_i|>1\}}. \]

Thus, \( Y \) is uniformly distributed on the two point set \( \Omega_Y \) and the conditional distribution of \( X_i \) given \( Y = y \) is \( N(y,1) \). 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

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

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

\[ p(\nabla) = 0.6318, \quad p_Q(\nabla) = 0.6714 \]

and the densities \( p(x), \tilde{p}(x), p_Q(x) \) 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 \( X_1, X_2, \ldots, \) be i.i.d. samples from \( p(x|y_0) \). The likelihood ratio test statistic for \( y_0 \) given the measurements \( X_1, \ldots, X_n \) 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

\[ T_{RS} = \sum_{i=1}^{n} L(x_i) \log \frac{p(Y = 1|X_i)}{p(Y = -1|X_i)}, \quad N_{RS} = \sum_{i=1}^{n} L(x_i), \]

\[ T_Q = \sum_{i=1}^{n} L_Q(Q_i) \log \frac{p(Y = 1|X_i)}{p(Y = -1|X_i)}, \quad N_Q = \sum_{i=1}^{n} L_Q(Q_i). \]

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

\[ \frac{1}{n} T \sim \text{as}N\left( \frac{2}{n}, \frac{4}{n} \right), \]

\[ \frac{1}{n} \left( \frac{T_{RS}}{N_{RS}} \right) \sim \text{as}N\left( \frac{2.9921}{1}, \frac{1}{n} \left( 12.1234, 2.4244 \right) \right), \]

\[ \frac{1}{n} \left( \frac{T_Q}{N_Q} \right) \sim \text{as}N\left( \frac{3.3197}{1}, \frac{1}{n} \left( 15.9619, 3.0308 \right) \right) \]

and

\[ \frac{T_{RS}}{N_{RS}} \sim \text{as}N\left( 2.9921, \frac{4.2979}{n} \right), \]

\[ \frac{T_Q}{N_Q} \sim \text{as}N\left( 3.3197, \frac{4.0124}{n} \right). \]
Figure 1. The relevance function $R_{Y|X}(x)$.

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 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).

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


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

Boris Ryabko\textsuperscript{1,2} and Nadezhda Savina\textsuperscript{1}

\textsuperscript{1}Institute of Computational Technologies of SB RAS, Novosibirsk, Russia  
\textsuperscript{2}Novosibirsk State University, Novosibirsk, Russia  
boris@ryabko.net  savina.nn@mail.ru

\section*{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 form) \(X_1X_2...\) 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.

\section{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 \(i\)\textsuperscript{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. \(d(c(X_i, Y_i), Y_i) = X_i\). Here \(c\) and \(d\) are functions called coder and decoder, correspondingly. Quite often the following particular formula are used

\[ Z_i = (X_i + Y_i) \mod n, \quad X_i = (Z_i - Y_i) \mod n, \quad i = 1, ..., t \]  
\[ a \oplus b = (a + b) \mod 2. \]

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. \(\text{P}(Y_1...Y_t = k_1...k_t) = n^{-t}\) 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 \( A = \{0, 1, \ldots, n-1\}, n \geq 2 \).

The \( m \)-order Shannon entropy and the limit Shannon entropy are defined as follows:

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

\[
h(X) = \lim_{m \to \infty} h_m(X) \tag{3}
\]

where \( m \geq 0 \), \( P_X(u) \) is the probability that \( X_1, X_2, \ldots, X_m = u \) (this limit always exists, see, for ex., [1, 2]). Introduce also the conditional Shannon entropy

\[
h_m(X|Z) = h_m(X, Z) - h_m(Z), \quad h(X|Z) = \lim_{m \to \infty} h_m(X|Z) \tag{4}
\]

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

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

where \( P(X_1 \ldots X_n|Z_1 \ldots Z_n) \) is a conditional probability.

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

3. ESTIMATIONS OF SECRECY

**Theorem 2.** Let a plaintext \( X = X_1, X_2, \ldots \) and the key sequence \( Y = Y_1, Y_2, \ldots \) be independent with a finite alphabet \( A = \{0, 1, \ldots, n-1\}, n \geq 2 \), and \( (X, Y) \) be a twodimensional 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 \( X^1 = X_1^1, X_1^2, \ldots, X_t^2 \) 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 \( t > n' \)

\[
P \left\{ \left| -\frac{1}{t} \log P(X_1 \ldots X_t|Z_1 \ldots Z_t) - h(X|Z) \right| < \varepsilon/2 \right\}
\]

\[
\geq 1 - \delta. \tag{6}
\]

Let us define

\[
\Psi(Z) = \{ X = X_1 X_2 \ldots X_t : \\
\mid P(X_1 \ldots X_t|Z_1 \ldots Z_t) - h(X|Z) \mid < \varepsilon/2 \}. \tag{7}
\]

The first property i) immediately follows from (6). In order to prove ii), note that for any \( X^1 = X_1^1, \ldots, X_t^1, X^2 = X_1^2, \ldots, X_t^2 \) 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 \( Z_1 Z_2 \ldots \)

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

**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 \( h(Z|X) \geq h(Z|X) + h(X) - \log n \). 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 \( \frac{1}{t} \log |\Psi(Z)| \):

\[
\liminf_{t \to \infty} \frac{1}{t} \log |\Psi(Z)| \geq h(X) - r_Y, \tag{8}
\]

\[
\liminf_{t \to \infty} \frac{1}{t} \log |\Psi(Z)| \geq h(Y) - r_X, \tag{9}
\]
\[
\liminf_{t \to \infty} \frac{1}{t} \log |\Psi(Z)| \geq \log n - (r_X + r_Y), \quad (8)
\]

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 \( X_1X_2\ldots, X_i \in \{0, 1\} \) and the key sequence \( Y_1Y_2\ldots, Y_i \in \{0, 1\} \), generated by a source whose entropy \( h(Y) \) is less than 1. \( (h(Y) = 1 \text{ if and only if } Y_1Y_2\ldots \text{ generated by the Bernoulli source with letter probabilities } P(0) = 0.5, [1, 2]). \) 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 \( Y_1Y_2\ldots \) is generated by the Bernoulli source with biased probabilities, say \( P(0) = 0.5 - \tau, P(1) = 0.5 + \tau \), where \( \tau \) is a small number. If \( \tau \) 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


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 asso-ciating 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 analyti-}

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

where

\[ I(q) = \int \int q(y|x) p(x) \log \frac{q(y|x)}{q(y|x) p(x)} dx dy \]

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 \( s \geq 0 \),

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

then, the rate-distortion function is parametrically given by

\[ R(D_s) = I(q_s), \]

\[ D_s = \int q_s(y|x) p(x) d(x,y) dy. \]

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 \( 0 < D < D_{\text{max}} \), where

\[
D_{\text{max}} = \inf_{y} \int p(x) d(x,y) dx,
\]

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

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 \( \phi \). Then the similarity between the two points \( x \) and \( y \) in \( \mathcal{X} \) is measured by the inner product \( (\phi(x), \phi(y)) \) in \( F \).

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

\[
K(x, y) = (\phi(x), \phi(y)),
\]

which is called the kernel function. Mercer’s theorem ensures that there exists some \( \phi \) such that Eq. (3) holds if \( K \) is a positive definite kernel [1]. This enables us to avoid explicitly computing the feature map \( \phi \) 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 \( y \in \mathcal{X} \subset \mathbb{R}^d \) is computed for each input \( x \in \mathcal{X} \), the distortion measure is naturally defined by

\[
d_{\text{inp}}(x, y) = ||\phi(x) - \phi(y)||^2 = K(x, x) + K(y, y) - 2K(x, y).
\]

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,

\[
d_{\text{inp}}(x, y) = \rho(x - y),
\]

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 \( D_{\text{max}} \) in Eq. (2) is denoted by \( D_{\text{max,inp}} \).

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 \( \alpha_i \in \mathbb{R}, i = 1, ..., n \), and consider it as the reconstruction in feature space, the distortion can be measured by

\[
d_{\text{lea}}(x, \alpha) = \left\| \phi(x) - \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|^2 = K(x, x) - 2\alpha^T k(x) + \alpha^T K \alpha,
\]

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

\[
k(x) = (K(x_1, x), ..., K(x_n, x))^T,
\]

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

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

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

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 \( \int e^{-\rho(z)} dz \) diverges to \(\infty\). 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

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

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 \( \int p(z) g(z) dz \leq D/Q_B \).

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

\[
g_s(z) = \frac{1}{C_{B,s}} \exp \left( 2s K(z, 0) \right) u(B - ||z||),
\]

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

\[
b(g_s) = -s \frac{\partial \log C_{B,s}}{\partial s} + \log C_{B,s},
\]

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

\[
R_{\text{inp,L}}(D_s) = Q_B \left\{ h(p_B) + s \frac{\partial \log C_{B,s}}{\partial s} - \log C_{B,s} \right\},
\]

\[
D_s = Q_B \left\{ 2C - \frac{\partial \log C_{B,s}}{\partial s} \right\}.
\]

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

If \( d_{\text{inp}} \) 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,

\[
R_{\text{inp}}(D_s) \leq R_{\text{inp,U}}(D_s) = h(g_s \ast p) - h(g_s)(10)
\]

\[
D_s = 2C - \frac{\partial \log C_{B,s}}{\partial s} \tag{11}
\]

47
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, $R_{inp,t}(D)$ is further upper-bounded by

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

where

$$v_p = \frac{1}{d} \int \|x - m\|^2 p(x)dx,$$

$$m = \int x p(x)dx,$$

$$v_s = \frac{1}{d} \int \|x\|^2 g_s(x)dx,$$

$$= \frac{A(d)}{dC_B s} \int_0^B r^{d+1} e^{2k(r)} dr,$$

where $A(d) = \frac{2^{d/2}}{(\sqrt{\pi}d)^{d/2}}$ 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 $R_{rea}(D)$

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

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

where $\tilde{K} = K + cI$,

$$m_K(x) = \tilde{K}^{-1}k(x),$$

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

$$R_{rea}(D_s) \leq R_{rea,t}(D_s)$$

$$= h(M_p) - h(N(\alpha; m_K(x), \tilde{K}^{-1}/2s)),$$

$$D_s = \int p(x)q(\alpha|x)d\gamma_{alfa}(x, \alpha)d\alpha$$

$$= \frac{n}{2s} + D_{\min}(c),$$

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

$$h(N(\alpha; m_K(x), \tilde{K}^{-1}/2s)) = \frac{n}{2} \log \left( \frac{\pi}{8} \tilde{K}^{1/n} \right),$$

which is independent of the input $x$, and

$$D_{\min}(c) = E[K(x, x)] - \text{tr} \left\{ \tilde{K}^{-1}E[k(x)k(x)^T] \right\}$$

$$+ \text{tr} \left\{ \tilde{K}^{-1}E[k(x)k(x)^T] \tilde{K}^{-1} \right\}.$$

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

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

where $C = E[k(x)k(x)^T] - E[k(x)]E[k(x)^T]$.

5. EXPERIMENTAL EVALUATION

We numerically evaluate the rate-distortion bounds obtained in the previous section. Designing a quantizer by the kernel $K$-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, $C_1 = \{ x \in \mathbb{R}^d; A(d)\|x\|^d \leq d/2 \}$ and $C_1 = \{ x \in \mathbb{R}^d; d^2 \leq A(d)\|x\|^d \leq d(d + 1/2) \}$, 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 integrals in the lower bound $R_{inp,L}$ and the upper bound $R_{inp,G}$. We generated i.i.d sample of the size $n = 4000$ from the source to compute $k(x)$ and $K$ for $R_{rea,G}$ 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 $K$-means algorithm 10 times with random initializations to obtain the minimum distortion for each rate. Varying the number $K$ of quantized points from $2^4$ to $2^{10}$ for each $K$, we counted the effective number $K_{\text{eff}}$ of quantized points which have at least one assigned data point and computed rates by $\log_2 K_{\text{eff}}$. The kernel parameter $\gamma$ was chosen so that the clear separation of $C_1$ and $C_2$ is obtained when $K = 2$. We optimized the regularization coefficient $c$ to minimize the upper bound $R_{rea,G}$ for each $K$.

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 $D_{\text{rea},G}$ is smaller than $D_{\text{inp},G}$ at low rates while the bound is above the quantizer performance. However, the value of $D_{\text{max},\text{rea}}$ suggests that the bound is informative at low rates. As the rate becomes higher, the lower and upper bounds of the input-space-reconstruction, $D_{L,\text{inp}}$ and $D_{U,\text{inp}}$, 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, $R_{\text{rea},G}(D)$ is close to the rate-distortion function of the input space reconstruction $R_{\text{inp}}(D)$ at high rates.
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 $K$-means to the training data set of the size, 4000. That is, as 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


ACHIEVABLE INFORMATION RATES ON LINEAR INTERFERENCE CHANNELS WITH DISCRETE INPUT

Metodi Yankov and Søren Forchhammer

Department of Photonics Engineering, Technical University of Denmark, Bldg. 343, Ørsteds Plads, 2800 Kgs. Lyngby, Denmark
mey@fotonik.dtu.dk

1

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:

\[ Y = HX + W, \]

where \( X \) is \( M \)-dimensional complex random variable vector \( X = [X_1, X_2, \ldots, X_M]^T \), 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 \( H \) represents the \([N \times M]\) 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 \( x_k \) (in the case of 1D variable), and the sequence from time \( t \) to \( k \) as \( x_k^t = [x_t, x_{t+1}, \ldots, x_k]^T \).

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:

\[ I(X; Y) = E_H [I(X; Y|H)] = H(X) - E_H [H(X|Y,H)]. \]

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]:

\[ H(X|Y,H) = -\lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \log_2 p(x_k|y_k, H). \]

The probability above is calculated from Bayes theorem:

\[ p(x_k|y_k, H) = \frac{p(y_k|x_k, H)p(x_k)}{\sum_{x_{k'}} \in \mathcal{X}^M p(y_k|x_{k'}, H)p(x_{k'})} \]

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 $H = Q R$ be the QR decomposition of $H$, where $Q$ is unitary and $R$ is upper-triangular. A well known MIMO receiver utilizes the form of $R$ 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 $Y = Q^H Y$, and the channel model becomes $Y_i = \sum_{j=1}^{M} R_{i,j} X_j$. Assuming the layers $i + 1$ to $M$ are correctly decoded by the following channel code, the symbols can be re-modulated and subtracted from the current layer $i$. Here we use a similar technique to derive a lower bound on the channel capacity.

Since $Q$ is unitary and doesn’t change the entropy of $Y$, and thus the MI, we can write:

$$\mathcal{I}(X; Y| H) = \mathcal{H}(X) - \mathcal{H}(X| Y| H) = \mathcal{H}(X) - \sum_{i=1:M} \mathcal{H}(X_i|Y_i, X_{i+1}^M, H) \geq \mathcal{H}(X) - \sum_{i=1:M} \mathcal{H}(X_i|Y_i, X_{i+1}^M, H) = \mathcal{I}(X; Y| H), \quad (5)$$

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):

$$p(X_i|Y_i, X_{i+1}^M, H) = \frac{p(Y_i|X_i, X_{i+1}^M, H) p(X_i)}{\sum_X p(X_i) p(Y_i|X_i, X_{i+1}^M, H)}$$

(6)

Since we condition on the following layers, the likelihood above can be expressed as:

$$p(\hat{Y}_i|X_i, X_{i+1}^M, H) = \mathcal{N}(\hat{Y}_i - \sum_{j=i+1:M} R_{i,j} X_j| X_i, 1), \quad (7)$$

where $R_{i,j}$ is the element on the $i$-th row and $j$-th column of $R$, and $\mathcal{N}(x|\mu, \sigma^2)$ is a 1D Gaussian function at $x$, with mean and variance $\mu$ and $\sigma^2$, 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 $|R_{i,j}|^2 E[X_i^2]$. When $M \ll N$, 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:

$$\mathcal{N}(\hat{Y}_i - \sum_{j=i+1:M} R_{N,j} X_j| R_{N,i}, \hat{\sigma}_i), \quad (8)$$

where $\hat{\sigma}_i = 1 + \sum_{j=N+1-i}^{N} |R_{N,j}|^2 E[X_i^2]$. In this case it is clear, that in the asymptotically high SNR we have:

$$\lim_{E[X_i^2] \to \infty} \mathcal{I}(X; Y| H) = \mathcal{H}(X), \quad (9)$$

whereas:

$$\lim_{E[X_i^2] \to \infty} \mathcal{I}(X; Y| H) = \mathcal{H}(X) - \sum_{i=N+1:M} \mathcal{H}(X_i| Y, H, SNR_i), \quad (10)$$

where the conditional entropy is larger than zero, because $\lim_{E[X_i^2] \to \infty} SNR_i = \frac{R_{N,j}}{\sum_j R_{N,j}}$, 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 $\tilde{p}(X) \neq p(X)$. If $X$ is generated by its original PDF, then the upper bound is found by calculating the entropy function from $X$, but using $\tilde{p}(X)$ [6]:

$$\mathcal{H}(X) = - \frac{1}{K} \sum_k \log_2 \tilde{p}(x_k) \geq \mathcal{H}(X)$$

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 $p(x|y)$ and $p(y|x) = \frac{p_Y(Y|X)p_X(X)}{\sum_X p_Y(Y|X)p_X(X)}$. Then if $y$ is generated by the law $p_Y(Y|X)$, the lower bound is calculated as:

$$\mathcal{I}(X; Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) \leq \mathcal{H}(X) - \mathcal{H}(Y) = \mathcal{I}(X, Y), \quad (11)$$

where $\mathcal{H}(X|Y)$ is calculated using some valid PDF $\tilde{p}(X|Y)$.

Turning back to the $R$ channel, we use the auxiliary probability distribution $\tilde{p}(X|Y, H)$:

$$\tilde{p}(X|Y, H) = \prod_{i=1}^{M} \tilde{p}(X_i|Y_i, X_{i+1}^M, H)$$

$$= \prod_{i=1}^{M} \frac{\tilde{p}(Y_i|X_i, X_{i+1}^M, H)p(X_i)}{\sum_X \tilde{p}(Y_i|X_i, X_{i+1}^M, H)p(X_i)}$$

where:

$$\tilde{p}(Y_i|X_i, X_{i+1}^M, H) = \mathcal{N}(\hat{Y}_i - \sum_{j=i+1:M} R_{i,j} X_j| R_{i,j}, 1), \quad (12)$$

which leads to the same lower bound.

3.2. Impulse response channels

Consider a standard impulse response channel:

$$y_k = \sum_{i=0:t} h_i x_{k-i} + w_k, \quad (13)$$
where \( h = [h_0, h_1, \ldots, h_l]^T \) is the impulse response. Equivalently, the channel may be expressed in its matrix form:

\[
y_1^k = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\
0 & h_1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & h_l \end{bmatrix} x_1^k + w_1^k.
\]

The MI with channel knowledge at the receiver is now calculated as:

\[
I(X; Y | h) = \mathcal{H}(Y | h) - \mathcal{H}(Y | X, h) =
\]

\[
= - \frac{1}{K} \log_2 p(y_1^K | h) - H(W)
\]

(15)

The standard approach to calculating (15) is to use a trellis to calculate \( p(y_1^k) = \prod_{k=1}^K p(y_k | y_{1:k-1}) \). One section of such trellis is given in Fig. 1. The interfering symbols are cast into the state: \( S_k = \{X_{k-1}, \ldots, X_{k-1}\} \), 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]:

\[
p(s_k, y_1^k) = \sum_{s_{k-1}} \sum_{s_{k-2}} p(s_{k-1}, s_{k-2}) p(y_k | x_k, s_k) p(x_k | s_k)
\]

Since the number of states is given by \( |S| = |X|^l \), 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 \( [10^4 \times 10^3] \) 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 \( R \) matrix therefore may be expected to also hold some structure. For example, in all our simulations the diagonal elements of the \( R \) matrix either converged to some value, or to some periodic pattern. However, exploiting this periodicity is left for future research.

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 \( \text{rank}(H) \leq \min(M, N) \). 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 \( h \) 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
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 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


