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

Edited by
Philippe Jacquet
Janne Leppä-aho
& Teemu Roos
PREFACE

The Tenth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2017) took place on September 11–13, 2017, in Paris, France. The workshop was organized jointly by Laboratory for Information, Networking and Communication Sciences (LINCS), the Helsinki Institute for Information Technology HIIT, Telecom ParisTech, INRIA, ETH Zürich, and the Center for Science of Information.

The WITMSE series started in 2008 and has continued annually at locations in Tampere (2008–2009), Helsinki (2011), Amsterdam (2012), Tokyo (2013), Honolulu (2014), Copenhagen (2015) and Helsinki (2016). 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 22 invited talks and two plenary talks that were given by Peter Shor and François Baccelli.

Outside the technical sessions, the programme included a welcoming reception at LINCS on Monday 11th September, and a banquet dinner at restaurant Alcazar on the following day.

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

December 5, 2017
Paris, Helsinki

Philippe Jacquet,
Janne Leppä-aho,
& Teemu Roos
Contents

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

Plenary Talks

François Baccelli: Stochastic Geometry in High Dimension .................. 6
Peter Shor: Information Theory, Quantum Mechanics, and Black Holes ........ 7

Invited Papers

David R. Bickel: The Maximum-Entropy and Minimax Redundancy Distribution Classes of Sufficiently Small Codelength ................................................. 8
Mérouane Debbah: Random Matrix Methods for Wireless Communications .......... 11
Tim van Erven: MetaGrad: Multiple Learning Rates in Online Learning .......... 12
Antonio Galves: Context Tree Models in the Brain .................................. 13
Jean-Marie Gorce, Philippe Mary, and Jean-Marc Kéfif: Towards Fundamental Limits of Bursty Multi-User Communications in Wireless Networks .......... 14
Peter Harremoës: Sharp Bounds on Tail Probabilities for Poisson Random Variables .... 18
Alexander Jung: Information-Theoretic Limits of Learning Networks for Big Data .... 21
Mokshay Madiman and Liyao Wang: Physical Implications of the Concentration of Information ................................................................. 22
Fabien Mathieu: Shannon, Turing and Hats: Information Theory Incompleteness .... 25
Olgica Milenkovic: New Approaches to Hypergraph Clustering and Community Detection .......................... 26
Boris Ryabko, Nadezhda Savina: Data Compressors Beyond Data Compression .... 27
Joe Suzuki: Conditional Mutual Information Estimation and its Application to Conditional Independence Detection ........................................ 29
Naftali Tishby: The Information Theory of Deep Learning ......................... 33
Kazuho Watanabe: Rate-distortion Dimension and Bayesian Learning Coefficient .... 34
Sumio Watanabe: Difference Between Bayes Cross Validation and WAIC for Conditional Independent Samples .......................................................... 38
Kenji Yamanishi: Gradual Model Change Detection via Descriptive Dimensionality . 42
Marek Zaionc: Asymptotic Density in Logic and Computability ................................. 43
STOCHASTIC GEOMETRY IN HIGH DIMENSION

François Baccelli

University of Texas at Austin

ABSTRACT

This talk will survey recent results on Boolean stochastic geometry in high dimensional Euclidean spaces.

A Boolean model in $\mathbb{R}^n$ consists of a homogeneous Poisson point process in $\mathbb{R}^n$, and independent and identically distributed random closed sets of $\mathbb{R}^n$ centered on each atom of this point process. The Shannon regime features a family of Boolean models indexed by $n \geq 1$, where the $n$-th model has a Poisson point process of intensity $e^{n\rho}$ and i.i.d. random compact sets with diameter of order $\sqrt{n}$, with $n$ tending to $\infty$. A typical example is that where each random compact set is an $n$-ball of radius distributed like $\bar{X}_n \sqrt{n}$, with $\bar{X}_n$ satisfying a large deviations principle.

The main focus of the talk will be on the asymptotic behavior of classical Boolean stochastic geometry quantities, like volume fraction, percolation threshold or mean cluster size, in this Shannon regime. More general high dimensional Particle Processes will also be discussed. For instance the case where the point process is hard core or determinantal rather than Poisson, and compact sets are still i.i.d.

This work is motivated by problems in information theory in the Poltyrev regime. The Boolean case corresponds to random coding, and the Particle Process case to more general coding assumptions. It leads to new results on error exponents.

Joint work with V. Anantharam (UC Berkeley), and E. O’Reilly (UT Austin).
INFORMATION THEORY, QUANTUM MECHANICS, AND BLACK HOLES

Peter Shor

Massachusetts Institute of Technology

ABSTRACT

The black hole information paradox is the question of whether information which enters a black hole can ever escape. Quantum mechanics says that it must escape when a black hole evaporates. General relativity says that it cannot. Recently, physicists have been looking to information theory to try to resolve this paradox.

Several fundamental modifications have to be made to classical information theory in order to apply it to quantum mechanical systems. Researchers have been studying how to modify it for some time, and we now have a robust theory of quantum information. After introducing the fundamentals of quantum information theory, I will discuss some proposals for how quantum information theory might help resolve the black hole information paradox.
THE MAXIMUM-ENTROPY AND MINIMAX REDUNDANCY DISTRIBUTION CLASSES
OF SUFFICIENTLY SMALL CODELENGTH

David R. Bickel

University of Ottawa

1. AVAILABILITY

The paper (Bickel, 2015) and slides (Bickel, 2017) corresponding to this extended abstract are available via http://www.davidbickel.com.

2. EXTENDED ABSTRACT

Inference has to proceed in some way even after a Bayesian model is found to be inadequate, perhaps due to excessive code length. Should the researcher infer that no conclusions can be drawn? If not, what conclusion may be drawn and with what posterior probability does the conclusion hold?

To address this issue, a two-stage procedure is proposed:

1. The first stage checks each model within a large class of models to assess which models have sufficiently small code length for purposes of data analysis (Bickel, 2015, §A.1).

2. The resulting set of small-code length models is then used in the second stage either for summarizing a combined posterior such as a maximum-entropy posterior or for inference according to decision rules of the robust Bayes approach and of imprecise probability more generally (Bickel, 2015, §3, §A.2).

3. ACKNOWLEDGMENTS

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009) and by the Faculty of Medicine of the University of Ottawa.

4. REFERENCES


URL https://wp.me/p5eoC-iX
MINIMUM ENTROPY JOINT DISTRIBUTIONS WITH FIXED MARGINALS: ALGORITHMS AND APPLICATIONS

Ferdinando Cicalese1, Luisa Gargano2, and Ugo Vaccaro2

1University of Verona, ITALY, cclfdn22@univr.it
2 University of Salerno, ITALY {lgargano,uvaccaro}@unisa.it

ABSTRACT
Given two discrete random variables $X$ and $Y$, with probability distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_m)$, respectively, denote by $C(p, q)$ the set of all joint distributions of $X$ and $Y$ that have $p$ and $q$ as marginals. In this paper, we study the problem of finding the joint probability distribution in $C(p, q)$ of minimum entropy (equivalently, the joint probability distribution that maximizes the mutual information between $X$ and $Y$), and we discuss several situations where the need for this kind of optimization naturally arises. Since the optimization problem is known to be NP-hard, we give an efficient algorithm to find a joint probability distribution in $C(p, q)$ with entropy exceeding the minimum possible by at most 1, thus providing an approximation algorithm with additive approximation factor of 1. We also discuss some related consequences of our findings.

1. INTRODUCTION AND MOTIVATIONS
Inferring an unknown joint distribution of two random variables (r.v.), when only their marginals are given, is an old problem in the area of probabilistic inference. The problem goes back at least to Hoeffding [11] and Frechet [8], who studied the question of identifying the extremal joint distribution of r.v. $X$ and $Y$ that maximizes (resp., minimizes) their correlation, given the marginal distributions of $X$ and $Y$. We refer the reader to [1, 6, 14] for a (partial) account of the vast literature in the area and the many applications in the pure and applied sciences.

In this paper, we consider the following case of the general problem described above. Let $X$ and $Y$ be two discrete r.v., distributed according to $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_m)$, respectively. We seek a minimum-entropy joint probability distribution of $X$ and $Y$, whose marginals are equal to $p$ and $q$. This problem arises in many situations. For instance, the authors of [12] consider the important question of identifying the correct causal direction between two arbitrary r.v. $X$ and $Y$, that is, they want to discover whether it is the case that $X$ causes $Y$ or it is $Y$ that causes $X$. In general, $X$ causes $Y$ if there exists an exogenous r.v. $E$ and a deterministic function $f$ such that $Y = f(X, E)$. In order to identify the correct causal direction (i.e., either from $X$ to $Y$ or from $Y$ to $X$), the authors make the reasonable postulate that the entropy of the exogenous r.v. $E$ is small in the true causal direction, and empirically validate this assumption. Additionally, they prove the interesting fact that the problem of finding the exogenous variable $E$ with minimum entropy is equivalent to the problem of finding the minimum-entropy joint distribution of properly defined random variables, given (i.e., fixed) their marginal distributions. This is exactly the problem we consider in this paper. The authors of [12] also observe that the latter optimization problem is NP-hard (due to results of [13, 18]), and evaluate experimentally a greedy approximation algorithm to find the minimum-entropy joint distribution, given the marginals. No proved performance guarantee is given in [12] for that algorithm. In this paper, we give a (different) greedy algorithm and we prove that it returns a correct joint probability distribution (i.e., with the prescribed marginals) with entropy exceeding the minimum possible by at most 1. Another work that considers the problem of finding the minimum-entropy joint distribution of two r.v. $X$ and $Y$, given the marginals of $X$ and $Y$, is the paper [18]. There, the author introduces (and applies it to the problem of order-reduction of stochastic processes) a pseudo-metric among discrete probability distributions in the following way: given arbitrary $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_m)$, $m \leq n$, the distance $D(p, q)$ among $p$ and $q$ is defined as the quantity $D(p, q) = 2W(p, q) - H(p) - H(q)$, where $W(p, q)$ is the minimum entropy of a bivariate probability distribution that has $p$ and $q$ as marginals, and $H$ denotes the Shannon entropy. The author of [18] observes that the problem of computing $W(p, q)$ is NP-hard and proposes another different greedy algorithm for its computation, based on some analogy with the problem of Bin Packing with over-stuffing. Again, no performance guarantee is given in [18] for the proposed algorithm. Our result directly implies that we can compute the value of $D(p, q)$, for arbitrary $p$ and $q$, with an additive error of at most 1.

There are many other problems that require the computation of the minimum-entropy joint probability distribution of two random variables, whose marginals are equal to $p$ and $q$. We shall limit ourselves to discuss a few additional examples, postponing a more complete examination in the full version of the paper. To this purpose, let us write the joint entropy of two r.v. $X$ and $Y$, distributed according to $p$ and $q$, respectively, as $H(XY) = H(X) + H(Y) - I(X; Y)$, where $I(X; Y)$ is the mutual...
information between $X$ and $Y$. Then, one sees that our original problem can be equivalently stated as the determination of a joint probability distribution of $X$ and $Y$ (having given marginals $p$ and $q$) that maximizes the mutual information $I(X;Y)$. In the paper [13] this maximal mutual information is interpreted, in agreement with Renyi axioms for a *bona fide* dependence measure [17], as a measure of the *largest possible* dependence of the two r.v. $X$ and $Y$. Since the problem of its exact computation is obviously NP-hard, our result implies an approximation algorithm for it. Another situation where the need to maximize the mutual information between two r.v. (with fixed probability distributions) naturally arises, is in the area of medical imaging [16]. Finally, our problem could also be seen as a kind of “channel-synthesis” problem, where it is given pair of r.v. $(X, Y)$, and the goal is to construct a memoryless channel that maximizes the mutual information $I(X;Y)$ between $X$ and $Y$.

### 2. OUR RESULT

Given two discrete random variables $X$ and $Y$, with probability distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_m)$, respectively, denote by $C(p, q)$ the set of all joint distributions of $X$ and $Y$ that have $p$ and $q$ as marginals. For our purposes, each element in $C(p, q)$ can be seen as an $n \times m$ matrix $M = [m_{ij}] \in \mathbb{R}^{n \times m}$ such that its row-sums give the elements of $p$ and its column-sums give the elements of $q$, that is,

$$C(p, q) = \left\{ M = [m_{ij}] : \sum_j m_{ij} = p_i, \sum_i m_{ij} = q_j \right\}. \tag{1}$$

The goal is to approximate the quantity

$$OPT = \min_{N \in C(p, q)} H(N).$$

We have already observed that computing above minimum value is NP-hard. We can give an algorithm that, having in input distributions $p$ and $q$, constructs in polynomial time an $M \in C(p, q)$ such that

$$H(M) \leq OPT + 1. \tag{2}$$

The algorithm and all related proofs are contained in the report [5].

### 3. REFERENCES


RANDOM MATRIX METHODS FOR WIRELESS COMMUNICATIONS

Mérouane Debbah

Mathematical and Algorithmic Sciences Lab, Huawei, France

ABSTRACT

The following talk aims to provide a comprehensive understanding on how random matrix theory can model the complexity of the interaction between wireless devices. One of the useful features of the large dimensional random matrix theory approach, is its ability to predict, under certain conditions, the behavior of the empirical eigenvalue distribution of products and sums of matrices. The results are striking in terms of accuracy compared to simulations with reasonable matrix sizes, and the theory has been shown to be an efficient tool to predict the behavior of wireless systems with only few meaningful parameters. In this talk, we will introduce the recent results in the field, in particular the "Deterministic Equivalent" approach and its application to the design of Massive MIMO systems.
ABSTRACT
In online convex optimization (known as prediction with individual sequences in information theory) it is well known that certain subclasses of loss functions are much easier than arbitrary convex functions. We are interested in designing adaptive methods that can automatically get fast rates in as many such subclasses as possible, without any manual tuning. Previous adaptive methods are able to interpolate between strongly convex and general convex functions. We present a new method, MetaGrad, that adapts to a much broader class of functions, including exp-concave and strongly convex functions, but also various types of stochastic and non-stochastic functions without any curvature. For instance, MetaGrad can achieve logarithmic regret on the unregularized hinge loss, even though it has no curvature, if the data come from a favourable probability distribution, which satisfies a Bernstein/margin-type condition.

MetaGrad’s main feature is that it simultaneously considers multiple learning rates, which control the amount of regularization. Unlike all previous methods with provable regret guarantees, however, its learning rates are not monotonically decreasing over time and are not tuned based on a theoretically derived bound on the regret. Instead, they are weighted directly proportional to their empirical performance on the data using a tilted exponential weights master algorithm.

This is joint work with Wouter Koolen and Peter Grünwald [1, 2].

1. REFERENCES
CONTEXT TREE MODELS IN THE BRAIN

Antonio Galves

University of S.Paulo, Brazil

ABSTRACT

Galves and Loecherbach (2013) introduces a new class of stochastic processes describing networks of spiking neurons. These processes are systems of interacting chains with memory of variable length. Each chain in the system describes the spiking activity of a single neuron. Processes in this class are not markovian as the spiking rate of each neuron depends on the history of the system after the last spiking time of that neuron. This class of models extends the class of context tree models introduced in the seminal paper Rissanen (1983). In this talk some recent results about these models will be presented, including a consistent statistical selection procedure to retrieve the graph of interactions characterizing the system.
ABSTRACT
Considering an isolated wireless cell containing a high density of nodes, the fundamental limit can be defined as the maximal number of nodes the associate base station can serve under some system level constraints including maximal rate, reliability, latency and transmission power. This limit can be investigated in the downlink, modeled as a spatial continuum broadcast channel (SCBC) as well as in the uplink modeled as a spatial continuum multiple access channel (SCMAC).

In this short paper, we summarize the different steps towards the characterization of this fundamental limit, considering four figures of merit: energy efficiency, spectral efficiency, latency, reliability.

1. INTRODUCTION

The recent evolution of the cellular market towards internet of things (IoT) connecting objects instead of humans induces significant changes in the mathematical modeling of wireless networks. Indeed, the IoT paradigm relies on bursty but massive distributed transmissions, where billions of communicating objects will be spread over large radio cells, transmitting only few packets per day, per month or even per year. In such a scenario, the classical fundamental limit of communication systems derived since the seminal work of Claude E. Shannon [1] needs to be revised. As we will see, the capacity, or the capacity region in case of multi-user communications, becomes less important in regard to other metrics.

It is worth that Shannon’s second theorem on capacity can be expressed as a fundamental tradeoff between energy efficiency (EE) and spectral efficiency (SE). For a real Gaussian channel the received power needs to verify \( P \geq (2^{\eta_S}/2 - 1) N_0 \), where \( \eta_S \), the spectral efficiency, is the number of bits-per-channel-use and \( N_0 \) the receiver noise. The normalized energy efficiency is given by \( \eta_E = \eta_S/(P/N_0) \). But this fundamental limit is achievable in the asymptotic regime, i.e. when the encoding time spread over an infinite number of channel uses. This is nothing but a steady-state analysis and relies on a set of assumptions, not further valid in the context of this study:

1. A unique source and a unique destination are considered.
2. The traffic is characterized by a fixed and continuous data rate.
3. Synchronization and transmitter identification costs are dropped off since the communication is assumed to be established.
4. The encoding time is infinite and so is the latency.

Information theory has been extended since its infancy to multi-user scenarios including the multiple access channel (MAC) or the broadcast channel (BC) well appropriate to study the uplink and the downlink, respectively. Most noticeable results with these schemes have been provided in the asymptotic regime, providing the expression of achievable rates region [2] characterizing achievable joint rates. For Gaussian channels, these capacity regions are known even for a large number of nodes and strategies based on superposition coding and successive decoding are capacity achieving [2].

But in the context of IoT cells with bursty traffic, these joint rates are not representative of the real scenarios that have to be supported. It comes out to be necessary to study these fundamental limits in the finite block length (FBL) regime where the information transmitted in the network is divided in finite size quantities that have to be transmitted under finite time transmission constraints. Yuri Polyanskiy et al [3] recently studied the fundamental limits of the point to point Gaussian channel in the FBL regime, paving the way to the study of latency and reliability constraints from a fundamental point of view. At the best of our knowledge the unique extension of this work for Gaussian channels has been done by Molavian-Jazi and Laneman [4] for the Gaussian MAC. We analysed the Gaussian BC in [5].

In addition to the FBL regime it is also necessary to study the behavior of BC/MAC models when the nodes are modeled through a spatial distribution as we proposed in [6, 7] to establish the fundamental EE-SE limit under equal rates condition in the asymptotic regime. We called these models SCMAC or SCBC, where SC stands for spatial continuum. This fundamental limit can be interpreted as the equivalent of the asymptotic Shannon capacity for a wireless cell.
If these results provided interesting insights, latency and reliability where kept off the study. Indeed, the asymptotic regime relies on error free but infinite latency transmission. So, the next question relates on the price of in-transient regime necessary to synchronize and detect which nodes are transmitting cannot be neglected. This corresponds to an additional information to be transmitted that should be considered in large scale low rate networks. Secondly, transmitting small packets over finite time slots leads necessarily to transmission errors. The minimal error can be modeled in information theory by working in the finite block length (FBL) regime, introducing an additional latency-reliability tradeoff.

2. SYSTEM MODEL

Let consider a unique BS, serving a cell area denoted by \( \Omega \subset \mathbb{R}^2 \). We denote by \( (\Omega, A, m) \) the corresponding measurable space with \( A \) the Lebesgue \( \sigma \)-algebra and \( m \) the Lebesgue measure. Let \( x \) be a point in \( \Omega \).

Without lack of generality, the BS is assumed to be located at point \((0,0)\). Nodes appear randomly in time and space on \( \Omega \). As such, they are not described by a discrete set but through a spatial density \( u(x) \) (the density of users per \( m^2 \) per time unit). Denoting by \( \tilde{u}(x,t) \) a realization of the ergodic random process with probability density function (pdf) \( u(x) \), then, for any subset \( B \subset A \), the average number of users per time unit is given by

\[
U(B) = \lim_{T \to \infty} \frac{1}{T} \int_{x \in B} \tilde{u}(x,t) \cdot dx \cdot dt.
\]

The global average number of nodes associated with the whole cell service area is denoted by \( U_T \).

Definition 2.1 (Requested rate density) The requested rate density \( \rho(x) : \Omega \to \mathbb{R} \) is a measurable function that represents the information rate spatial density requested at point \( x \).

- Note 1: \( \rho(x) \) is normalized\(^1\) by the system bandwidth and is expressed in bits per channel use (bpcu) per \( m^2 \).
- Note 2: The requested rates are either in downlink (SCBC) or uplink (SCMAC) modes.

The cell sum-rate per channel use is denoted the spectral efficiency (SE) of the cell:

\[
\eta_s = \int_{\Omega} \rho(x) \cdot m(dx). \tag{2}
\]

\(^1\)\( \rho(x) \) and related quantities (\( R(\cdot,\cdot) \)) are given in bit-per-channel-use (bpcu) throughout this paper. For practical applications, the results can be converted in bits or in bps, with appropriate system parameters.

The uniform rate condition corresponds to the case where each node requests the same quantity of information denoted by \( I_0 \), with

\[
\rho(x) = \frac{I_0 \cdot u(x)}{N_{cu}}, \tag{3}
\]

where \( N_{cu} \) is the number of channel uses per time unit, allocated to the system. In this special case, one have \( \eta_s = I_0 / N_{cu} \).

3. UNIFORM CAPACITY IN THE ASYMPTOTIC REGIME

The access capacity region is defined in [6] for the uplink and the downlink, as the set of rate spatial distributions for which a joint encoding-decoding scheme exists such that the transmission error tends to 0 when the encoding time tends to infinity (characterized by the number of channel uses \( n \)). Compared to the classical Shannon’s approach, the asymptotic regime herein integrates another dimension: when \( n \to \infty \), the number of nodes tends to infinity but individual rates tend to 0 since each packet is spread over an infinite time. Note that the cell’s sum-rate tends to its average spectral efficiency. In [6], based on an iterative splitting process, the sum-rate that the cell can achieve when a continuum of users is considered has been obtained. The corresponding fundamental limit is expressed through the total transmission power required to serve a given distribution and was found identical for the GSCBC and the GSCMAC with transferable powers [7]:

Theorem 3.1 (GSCBC and GSCMAC fundamental limit) The achievable EE-SE tradeoff for a given rate spatial density \( \rho(x) \) is given by

\[
\eta_c \leq \left[ a \int_{\nu_{M}}^{\nu_{m}} t \cdot f_\nu(t) \cdot e^{a \cdot \eta_s \cdot G_\nu(t) \cdot dt} \right]^{-1}, \tag{4}
\]

where
\[
\begin{align*}
\bullet \quad a &= 2 \log(2), \\
\bullet \quad f_\nu(t) \quad \text{and} \quad G_\nu(t) \quad \text{are respectively the probability density function (pdf) and the complementary cumulated density function (ccdf) of the rate as a function of the equivalent noise (an exemple is represented in Fig.1)}. 
\end{align*}
\]

4. EXTENSIONS IN THE NON ASYMPTOTIC REGIME

The former result was derived in the asymptotic regime under no latency constraint. Modeling IoT networks requires additional constraints that will reduce the fundamental limit.

4.1. Finite Time Transmission

The asymptotic regime used in Th. 3.1 means that all packets are transmitted simultaneously in the network to or from all nodes and spread over an infinite time. Let now be imposed a finite time constraint formalized as follows:
Figure 1. The rate request is distributed according to the density \( f_\nu(.) \), where \( \nu(x) \) for a node located in \( x \) is the equivalent noise power given by the ratio between the receiver noise and the channel gain \( \nu(x) = N_0/g(x) \). Then, \( G_\nu(t) \) is the fraction of the cell with an equivalent noise power greater or equal to \( t \). This distribution for a regular circular cell is represented where the \( x \) axis is the equivalent noise and the \( y \) axis represents the rate density.

**Property 4.1** A multi-user network with packets of constant information quantities \( I_0 \), is said FTT constrained if each transmission is constrained to be done in at most \( n^* \) channel uses.

The average spectral efficiency (i.e. the sum-rate) of the cell is noted \( \bar{\eta}_c \) and is nothing but:

\[
\bar{\eta}_c = \frac{I_0 \cdot U(\Omega)}{N_{cu}}.
\]

(5)

In the downlink, when BS transmits packets of \( I_0 \) bits in \( n^* \) channel uses, individual rates are given by \( \eta_n = \frac{I_0}{n^*} \). The BS can achieve the target spectral efficiency by using superposition coding (SC) to simultaneously transmit several packets. Therefore, optimizing the transmission strategy relies on optimizing the scheduling of the different packets in order to minimize the total power.

Let us define a frame of length \( N_{cu} \), made of \( L \) slots, each slot \( s_l \) being made of \( n^* \) channel uses. For consistency, all time units are counted in channel uses. In the downlink, the BS has in its queue a random number of packets to be transmitted to a set of nodes \( N_{U,L} \), with \( \lim_{L \to \infty} \frac{|N_{U,L}|}{L} = U(\Omega) \).

In each slot \( s_l \), a subset of nodes \( N_{u,l} \subset N_{U,L} \) is selected and the corresponding packets are simultaneously transmitted. Let \( N_{u,l} = |N_{u,l}| \) be the number of nodes allocated to the slot \( s_l \), the spectral efficiency is:

\[
\eta_u(l) = N_{u,l} \cdot \eta_u.
\]

(6)

To achieve the desired rate \( \eta_u = I_0/n^* \), the SINR at each decoding step has to verify \( \gamma^* \geq e^{\eta_u} - 1 \) according to the Shannon channel capacity. To be optimal, the power allocation is computed such that each node gets a SINR exactly equal to \( \gamma^* \). Using a superposition coding strategy, each node decodes its signal after having decoded all greater signals.

Under these conditions of equal rate, the minimal power associated to each slot is obtained as:

**Theorem 4.1 (Sum-power in the downlink)** For a given set of users indexed by \( \{1, \ldots, N_{u,l}\} \), the minimal sum-power necessary to transmit reliably to each node independent informations of size \( I_0 \) in \( n^* \) channel uses, is given by:

\[
P_m(l) = \gamma^* \sum_{k=1}^{N_{u,l}} \nu_{k,l} \cdot (1 + \gamma^*)^{k-1}
\]

with \( \gamma^* = e^{I_0/n^*} - 1 \), and \( k = N_{u,l} - k + 1 \).

Therefore, comparing numerically Th. 4.1 to Th. 3.1 allows to evaluate the performance loss due to the FTT constraint.

**4.2. Finite Block Length Coding**

In the former section, we assumed \( n^* \) to be large enough to have error-free transmission. This is not the case in a band limited system.

Strassen [8] studied the finite blocklength regime and has shown for a Discrete Memoryless Channel (DMC) that:

\[
\log M^*(n,\epsilon) = nC - Q^{-1}(\epsilon)\sqrt{nV} + O(\log n)
\]

(8)

where \( n, \epsilon, M \) are the blocklength, the error probability and the codebook size, respectively. The second term in the right hand side of (8) is the rate penalty, namely the back-off from the channel capacity. This expression gives the fundamental trade-off between latency and reliability for short packet transmissions. Unfortunately, Strassen’s result was not compliant to be generalized, particularly to the AWGN channel, generally to any type of channel with input constraints. Almost half a century later, Polyanisky et al. [3] has generalized Strassen’s approach, i.e. treating the mutual information density as a random variable and using second order statistics to approximate the channel coding rates in the FBL regime, for various channel models such as the AWGN channel, for which the following holds:

**Theorem 4.2 (Polyanskiy et al., 2010)** For the AWGN channel with signal-to-noise ratio (SNR) \( \gamma \), for \( \epsilon \in (0,1) \) and for equal, maximal and average-power constraints, the following holds:

\[
\log M^*(n,\epsilon,\gamma) = nC(\gamma) - Q^{-1}(\epsilon)\sqrt{nV(\gamma)} + O(\log n)
\]

(9)

where \( C(\gamma) \) is the Shannon capacity defined as

\[
C(\gamma) = \frac{1}{2} \log(1 + \gamma)
\]

(10)

and

\[
V(\gamma) = \frac{\gamma + 2}{2(\gamma + 1)^2} \log^2 e.
\]

(11)
However, the extension of this result to the case of multi-user channels is not straightforward. The exact capacity regions of Gaussian MAC or Gaussian BC are not known but achievable regions for the 2-user BC or 2-user MAC have been characterized recently in [4, 5]. As a first approximation, the use of a superposition coding strategy [5] allows to extrapolate the FBL expression in Th. 4.2 to compute a modified SNR threshold $\gamma^*$ as a function of some latency and reliability constraints. This $\gamma^*$ can be used in Th. 4.1 to modify the fundamental limit.

4.3. Additional real time constraints

In the former paragraph, we detailed how to combine the dense nodes distribution assumption and a finite time transmission constraint. To model real-time transmissions, where the nodes have to transmit or receive instantaneously, additional issues have to be considered. In the downlink, the main challenge is to indicate to the nodes if they have something to decode and how it is encoded. Clearly, signaling cannot be ignored and is even a part of the information to be transmitted. In the uplink, the randomness of the active nodes distribution and their identification is necessary. Stochastic geometry coupled with FBL information theory can provide insights about the impact of randomness [9]. From a protocol point of view, integrating data transmission and signaling as a whole, approaches based on compressive sensing [10, 11] are appealing although their optimality is not granted.

5. CONCLUSION

In this paper, we paved the way to derive fundamental limits of IoT cells where dense distributions of users are spread over large cells and where a sporadic traffic has to be supported. Starting from the EE-SE fundamental limit in the asymptotic regime, we defined the different steps towards the derivation of the EE-SE loss in the FBL regime and also how to balance it with respect to the reliability-latency trade-off.

6. ACKNOWLEDGMENTS

This work has been partly supported by Orange Labs under CRE n\textsuperscript{F}O5151, and by the French National Agency for Research (ANR) under project Arburst.

7. REFERENCES


SHARP BOUNDS ON TAIL PROBABILITIES FOR POISSON RANDOM VARIABLES

Peter Harremoës

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

ABSTRACT

Recently it has been shown that the signed log-likelihood of a Poisson random variable is quite close to a standard Gaussian distribution in the sense that one can prove sharp upper and lower bounds on the tail probabilities. Here we prove a one-sided bound that is even sharper.

1. INTRODUCTION

For a Goodness-of-Fit test one may use the \( \chi^2 \)-statistic or the \( G^2 \)-statistic and for about 100 year it has been unclear which of these statistics is better and in what sense. In this short note we shall focus on the simplest possible case where we have a single count that is assumed to follow a Poisson distribution with unknown mean. It is already known that the distribution of the \( G^2 \)-statistic is closer to a \( \chi^2 \)-distribution than the distribution of the \( \chi^2 \)-statistic is in [1]. Here we will provide a new sharp bound on the tail distribution of the \( G^2 \)-statistic that strengthen the conclusion in [1]. Instead of working with the \( G^2 \)-statistic we state related bounds in terms of tail bounds on the signed log-likelihood statistic \( G \).

In this note \( \Phi \) denotes the distribution function of a standard Gaussian variable and \( \phi = \Phi' \) denotes the density of a standard Gaussian variable.

2. NEW INEQUALITIES

Let \( L \) denote a Poisson random variable \( Po(\lambda) \) and let \( G_{Po}(\lambda) \) denote the function given by

\[
G_{Po}(\lambda)(x) = \pm \left[ 2 \left( x \ln \frac{x}{\lambda} + \lambda - x \right) \right]^{1/2}.
\]

In [1, 2] the following inequality was proved

\[
\Pr(L < k) \leq \Phi \left( G_{Po}(\lambda)(k) \right) \leq \Pr(L \leq k)
\]

for \( k = 0, 1, 2, \ldots \). The inequality is equivalent to

\[
\Phi \left( G_{Po}(\lambda)(\ell - 1) \right) \leq \Pr(L < \ell) \leq \Phi \left( G_{Po}(\lambda)(\ell) \right)
\]

for \( \ell = 1, 2, 3, \ldots \). The following improved inequality was conjectured.

\[
\Phi \left( G_{Po}(\lambda) \left( k + \frac{1}{2} \right) \right) \leq \Pr(L \leq k)
\]

for \( k = 0, 1, 2, \ldots \). The conjecture is equivalent to the main theorem that we will prove later in this paper.

Before we can prove the main theorem we need to prove a related result that is of independent interest. In [1, 2] it was proved that the signed log-likelihood of a Gamma distributed random variable \( X \sim \Gamma(\ell, \theta) \) is stochastically dominated by the standard Gaussian, i.e.

\[
\Phi \left( G_{\Gamma(\ell, \theta)}(x) \right) \leq \Pr(X \leq x).
\]

This result gives a lower bound on the distribution function of a Gamma distribution. Here we will prove an upper bound on the distribution function.

Theorem 1. The distribution function of a Gamma distributed random variable satisfies the following inequality

\[
\Pr(X \leq x) \leq \Phi \left( G_{\Gamma(\ell, \theta)}(x) \right).
\]

Proof. Let \( Z \) denote a standard Gaussian random variable. Then the inequality can be rewritten as

\[
\Pr \left( G_{\Gamma(\ell, \theta)}(X) \leq G_{\Gamma(\ell, \theta)}(x) \right) \leq \Pr \left( Z \leq G_{\Gamma(\ell, \theta)}(x) \right)
\]

which means that \( Z \) is stochastically dominated by \( G_{\Gamma(\ell, \theta)}(X) \).

To see this we need the density of \( G_{\Gamma(\ell, \theta)}(X) \). Without loss of generality we may assume that \( \theta = 1 \). We introduce

\[
f(x) = \frac{x^{\ell-1}}{\Gamma(\ell)} \exp(-x)
\]

\[
\gamma(x) = \pm \left[ 2 \left( x - 1 - \ln(x) \right) \right]^{1/2}
\]

\[
G(x) = \left( \ell - \frac{1}{2} \right)^{1/2} \gamma \left( x \ell - \frac{1}{2} \right)
\]

so that \( G_{\Gamma(\ell, \theta)}(X) \) has density

\[
f \left( G^{-1}(x) \right) = \frac{G'(G^{-1}(x))}{G'(G^{-1}(x))}
\]
We need to prove that
\[
\frac{f(G^{-1}(x))}{G'(G^{-1}(x))}\phi(x)
\]
is increasing but \(G\) is increasing so it is sufficient to prove that \(\frac{f(y)}{\phi(G(y))}\) is increasing. We have
\[
\frac{f(y)}{G'(y) \cdot \phi(G(y))} = \frac{y^\gamma \exp(-y)}{\Gamma(\ell) \cdot \frac{1}{t-x} \cdot \frac{1}{v^\gamma} \exp\left(-\frac{G(y)^2}{2}\right)}
\]
\[
= \frac{\tau^{1/2} (\ell - 1/2)^{1/2} \cdot \frac{y^\gamma \exp\left(-\frac{G(y)^2}{2} - y\right)}{\Gamma(\ell)}}{\gamma^\gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}
\]
This can be rewritten as
\[
\frac{\tau^{1/2} (\ell - 1/2)^{1/2} \exp\left(-\frac{G(y)^2}{2} - y\right)}{\Gamma(\ell)} \cdot \frac{y^\gamma \exp\left(-\frac{G(y)^2}{2} - y\right)}{\gamma^\gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}
\]
\[
= \frac{\tau^{1/2} (\ell - 1/2)^{1/2} \exp\left(\frac{1}{2} - \ell \right) \left(\ell - 1/2\right)^{y^\gamma \gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}
\]
Now we make the substitution \(z = \frac{y}{t-x}\) so that we have to prove that
\[
2\tau^{1/2} (\ell - 1/2)^{1/2} \exp\left(-\ell \left(\frac{1}{2} - \ell \right) \left(\ell - 1/2\right)^{y^\gamma \gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}
\]
is increasing. The function \(z \to z^{1/2} \cdot 2(z)\) is positive so it is sufficient to prove that
\[
\frac{f(y)}{G'(y) \cdot \phi(G(y))} = \frac{\tau^{1/2} (\ell - 1/2)^{1/2} \exp\left(-\ell \left(\frac{1}{2} - \ell \right) \left(\ell - 1/2\right)^{y^\gamma \gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}}{\Gamma(\ell) \cdot \ell \left(\ell - 1/2\right)^{1/2} \cdot \frac{y^\gamma \exp\left(-\frac{G(y)^2}{2} - y\right)}{\gamma^\gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}}
\]
is increasing. The function \(z \to z^{1/2} \cdot 2(z)\) is positive so it is increasing. We have
\[
h(z) = z \cdot \left(\frac{y^\gamma \gamma \left(\frac{y}{t-x}\right)^{\gamma/2} \cdot \frac{y-(\ell - 1/2)}{y-(\ell - 1/2)}} \cdot \frac{1}{\pm 2 \left[\left(\frac{t-x}{v^\gamma - 1 - \ln\left(\frac{y}{t-x}\right)}\right)^{1/2}\right]}
\]
is increasing. We have
\[
h'(z) = 2 \cdot \frac{(z-1)^2(2z-1-\ln(z)-1)}{(z-1)^4} - \frac{2(z-1)(2z-\ln(z)-2)}{(z-1)^3} - \frac{2z^2-z\ln(z)-2z-2z+\ln(z)+2}{(z-1)^3} - \frac{2z\ln(z)+\ln(z)-2z+2}{(z-1)^3}
\]
The function \(m(z) = z\ln(z) + \ln(z) - 2z + 2\) has \(m(1) = 0\) so it is sufficient to prove that \(m\) is increasing.
\[
m'(z) = \ln(z) + 1 - \frac{1}{z} - 2
\]
\[
= \frac{1}{z} - 1 - \ln\left(\frac{1}{z}\right)
\]
\(\geq 0.\)

Now we are ready for the main theorem.
Theorem 2. Let $L \sim Po(\lambda)$ denote a Poisson random variable and let $G_{Po}$ denote the function given by (1). Then
\[ \Phi \left( G_{Po}(\lambda) \left( \ell - \frac{1}{2} \right) \right) \leq Pr (L < \ell) \]
for $\ell = 1, 2, 3, \ldots$

Proof. We have
\[ Pr (L < \ell) = 1 - Pr (L \geq \ell) = 1 - Pr (Y \leq \lambda) \]
where $Y \sim \Gamma (\ell, 1)$.

Now
\[ 1 - Pr (Y \leq \lambda) \geq 1 - \Phi \left( G_{\Gamma(\ell-\frac{1}{2},1)} (\lambda) \right) \]
\[ = 1 - \Phi \left( -G_{Po}(\lambda) \left( \ell - \frac{1}{2} \right) \right) \]
\[ = \Phi \left( G_{Po}(\lambda) \left( \ell - \frac{1}{2} \right) \right) \]

Combining the inequalities gives
\[ Pr (L < \ell) \geq \Phi \left( G_{Po}(\lambda) \left( \ell - \frac{1}{2} \right) \right) . \]

3. DISCUSSION

The inequality proved in this note corresponds to the so-called continuity correction often used for the $\chi^2$-statistic. A similar continuity correction does not lead to a sharp inequality for binomial distributions. Instead one should look for similar inequalities for the Bernoulli process, i.e. bounds on the signed log-likelihood of a negative binomial distribution as conjectured in [2]. Numerical calculations seem to verify such an inequality. A proof of such an inequality would lead to inequalities for the binomial distributions like the one conjectured in [3], which has been a great motivation for the present work.

4. REFERENCES


INFORMATION-THEORETIC LIMITS OF LEARNING NETWORKS FOR BIG DATA

Alexander Jung

Dept. of Computer Science, Aalto University

ABSTRACT

We consider the problem of learning the network structure underlying massive datasets. Our approach is based on modelling the raw data as non-stationary samples of a high-dimensional Gaussian random process. The underlying network structure is defined via a probabilistic graphical model which encodes conditional independence relations between individual data features as an undirected graph, i.e., the conditional independence graph (CIG). We interpret the problem of learning the network structure of the CIG, i.e., graphical model selection, as a communication problem where “nature” selects a particular CIG which makes itself felt by the conditional dependencies inherent to the observed samples. Using this communication picture, we use standard tools from information theory to obtain lower bounds on the required sample size in order to reliably learn the underlying CIG structure. These limits depend crucially on a certain average partial correlation, which quantifies the strength of the edges in the CIG. Moreover, the bounds indicate that the required sample size typically grows linear with the maximum node degree of the underlying CIG but only logarithmically with the system size and therefore allows for coping with the high-dimensional regime where the system model size is much larger than the number of observed samples. We verify the tightness of the lower bounds by deriving matching upper bounds on the sample size required by a particular GMS method which is based on a simple sparse linear regression procedure for identifying the edges of the underlying CIG. This is joint work with N. Goertz, G. Hannak and N. T. Quang [1, 2]. The slides of the talk are available at https://gitlab.com/alexjung/Talks.

1. REFERENCES


PHYSICAL APPLICATIONS OF THE CONCENTRATION OF INFORMATION

Mokshay Madiman1 and Liyao Wang2

1Department of Mathematical Sciences, University of Delaware, 501 Ewing Hall, Newark DE 19716, USA, madiman@udel.edu
2 J. P. Morgan, New York, USA, njuwangliyao@gmail.com

ABSTRACT

We discuss some physical applications of recently developed concentration inequalities for the information content of log-concave random vectors. For example, we show that when a physical system has a Hamiltonian that is convex (possibly up to some symmetries), first-order phase transitions are ruled out.

1. SUMMARY OF RESULTS

The framework of classical statistical mechanics, with Gibbs measures determined by a positive temperature parameter and a Hamiltonian or energy functional on the space of possible configurations of the system, is very standard and much of the notation is standard as well. For convenience, we now collect such largely standardized notations. Denoting by $\langle \rangle$ such that $Z$ is the partition function. In this note, all Hamiltonians are existence of thermodynamic limit, then there are additional

1. $f_\beta = \lim_n f_{n,\beta},$
2. $u_\beta = -\frac{\partial f_\beta}{\partial \beta},$
3. $c_{v,\beta} = \beta^2 \frac{\partial^2 f_\beta}{\partial \beta^2},$ specific heat.

To give some motivation, we recall the classical equipartition property for energy [1]. In the simplest setting, it states that if the Hamiltonian $\mathcal{H}$ has the form

$$\mathcal{H}(x_1, x_2, ..., x_n) = \sum_i |x_i|^k,$$

where $x_i \in \mathbb{R}$ are abstract state variables, then the specific heat in a canonical ensemble $C_{v,n,\beta} = \frac{k}{T}$. If we require the Hamiltonian to be convex, which is the same as requiring $k \geq 1$ in this case, then $C_{v,n,\beta} \leq n$.

Somewhat unexpectedly, underlying this simple observation is a much more general phenomenon. Specifically, as long as the Hamiltonian is convex,

$$C_{v,n,\beta} \leq n,$$

which is optimal. Note that since the heat capacity measures energy fluctuations, the above result means that we have a uniform control of these fluctuations. Rigorous upper bounds for specific heat have appeared in a number of places in the literature (see, e.g., [2, 3]). Each of them deals with specific systems on a case by case basis. The universal and optimal nature of our bound, which follows from recent results developed independently by us [4, 5] and others [6, 7] (and refines earlier results of [8]) is somewhat surprising.

More can be said when the Hamiltonian is convex. In fact, we provide a non-asymptotic quantitative concentration result for energy in a canonical ensemble, which implies that the Gibbs measure is essentially the uniform measure in the energy shell

$$\{(x_1, x_2, ..., x_n) : |\mathcal{H}(x_1, x_2, ..., x_n) - \langle \mathcal{H} \rangle| \leq 2\sqrt{n}T\}$$

for large $n$. Note that typically, $\sqrt{n} \ll \langle \mathcal{H} \rangle$. This clearly sheds light on the equivalence of ensembles, which is one of the cornerstones of modern statistical mechanics. Gibbs noticed this point as early as in 1902 [9] and since then, there have been numerous works devoted to it. Rigorous studies include the early work based on local
central limit theorems (see, e.g., [10, 11]) and more recent work based on Large Deviation Principles (see, e.g., [12, 13, 14]). However, all of these results are asymptotic in nature, meaning that they are true in the thermodynamic limit. Our concentration result gives a non-asymptotic and explicit quantification of the equivalence assuming only convexity of the Hamiltonian.

As another application, we show rigorously that as long as the Hamiltonians are convex, there is no latent heat in the thermodynamic limit. The question of whether there is a liquid gas transition in a classical continuum system has attracted great attention. The answer is yes and many rigorous efforts have been made to find specific model systems where one can prove existence of such a transition [15, 16, 17]. Typically, such a transition will involve release or absorption of latent heat. Our result suggests that if the Hamiltonians are convex, then there is no such transition.

2. REMARKS

1. All our results can be extended to the case where the Hamiltonian is not convex on all of the state space $\mathbb{R}^n$, but is symmetric (in the sense of being invariant under permutation of the arguments) and is convex on each Weyl chamber (or equivalently, on the region $\{x \in \mathbb{R}^n : x_1 \leq x_2 \leq \ldots \leq x_n\}$). This significantly expands the applicability of the results to physical models in popular use, including for example the one-dimensional hardcore gravitational gas. Details, including specific applications, can be found in [18].

2. There exist extensions of the underlying concentration results to so-called convex measures defined by Borell, developed in [19]. It is unclear whether these have physical significance or applicability.

3. We mention in passing that the underlying concentration results have many applications outside of physics, for example in convex geometry and probability theory. Details may be found in [20, 21].

3. ACKNOWLEDGMENTS

M. Madiman was supported by NSF grants DMS-1409504 (CAREER) and CCF-1346564.

4. REFERENCES


SHANNON, TURING AND HATS: INFORMATION THEORY INCOMPLETENESS

Fabien Mathieu

Nokia Bell Labs, France

ABSTRACT

When Claude Shannon introduced his theory of information in 1948, he was mainly targeting potentially infinite sequences of symbols, e.g. finite sequences whose length \( n \) were tending to infinity. Parameters like information rate, error correction, compressibility and predictability are defined for finite \( n \) with interesting properties when \( n \) tends to infinity. Here we consider that the sequence of events is truly infinite (past and future). Manipulate truly infinite sequences is not exactly the same as manipulating potentially infinite sequences. Prediction can be related to infinite hat puzzles, so that if we introduce the axiom of choice in information theory over infinite sequences, then the number of prediction errors on an infinite sequence is finite. This implies that the prediction error rate is actually zero. This is a rather surprising result since the infinite sequence could have been generated by a memoryless source. Moreover, if the infinite sequence is a computable sequence toward the past, then we get the zero error rate result without the axiom of choice. This is also a surprising result since a computable sequence is not necessarily computable toward the future, since a Turing machine is not necessarily reversible.
NEW APPROACHES TO HYPERGRAPH CLUSTERING AND COMMUNITY DETECTION

Olgica Milenkovic
University of Illinois, Urbana-Champaign, USA

ABSTRACT
Motivated by emerging problems in network analysis, learning hierarchical rankings and image segmentation, we introduce a new paradigm in hypergraph clustering and community detection termed hypergraph correlation clustering. Hypergraph correlation clustering is an agnostic combinatorial learning method which allows for capturing communities with a priori unknown structures based only on the knowledge of hypergraph edge cardinalities. The method has statistical counterparts which can be applied to hypergraph generalizations of stochastic block models and for other community detection paradigms.
DATA COMPRESSORS BEYOND DATA COMPRESSION
Boris Ryabko1 and Nadezhda Savina2
1Institute of Computational Technologies of SB RAS
Novosibirsk, Russian Federation.
2 Siberian University of Consumer Cooperatives,
Novosibirsk, Russian Federation.

ABSTRACT
Nowadays data compressors are applied to many problems of text analysis, but many such applications are developed outside of the framework of mathematical statistics. In this paper we overcome this obstacle and show how several methods of classical mathematical statistics can be developed based on applications of the data compressors.

1. INTRODUCTION
Nowadays data compressors are widely used to compress texts for their storage or transmission. In the last 20 years, it was recognized that data compressors can be used for many purposes which are far from file compaction. In particular, it was shown that methods of data compression can be used for prediction and hypothesis testing for time series, see [9] and a review there. Later, several authors applied data compressors to problems which are close, in spirit, to homogeneity testing, estimation of correlation and covariance, classifications, clustering and some others; see [7, 1, 2, 8, 10]. The main idea of their approach can be understood from the following example. Suppose that there are three sequences of letters $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_k$, $z_1, z_2, \ldots, z_m$ and a certain data-compressor $\phi$. The sequences $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_k$ obey different probability distributions, whereas $z_1, z_2, \ldots, z_m$ obeys one of them. The goal is to determine this distribution. (It is the well-known “three samples problem”. ) If $x_1, x_2, \ldots, x_n$ and $z_1, z_2, \ldots, z_m$ obey the same probability distribution, then, the sequence $z_1, z_2, \ldots, z_m$ will be compressed better after $x_1, x_2, \ldots, x_n$ than after $y_1, y_2, \ldots, y_k$. More precisely, if one compresses sequences $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_k$ and combined ones as follows

$$\phi(x_1, x_2, \ldots, x_n, z_1, z_2, \ldots, z_m), \phi(y_1, y_2, \ldots, y_k, z_1, z_2, \ldots, z_m),$$

the difference $|\phi(x_1, x_2, \ldots, x_n, z_1, z_2, \ldots, z_m)| - |\phi(x_1, x_2, \ldots, x_n)|$ will be less than $|\phi(y_1, y_2, \ldots, y_k, z_1, z_2, \ldots, z_m)| - |\phi(y_1, y_2, \ldots, y_k)|$ (Here $|U|$ is the length of $U$.) For instance, let $x_1, x_2, \ldots, x_n$, $z_1, z_2, \ldots, z_m$ be texts in English, whereas $y_1, y_2, \ldots, y_k$ is in German. Then the English text $z_1, z_2, \ldots, z_m$ will be compressed better after the text in the same language ($x_1, x_2, \ldots, x_n$) than after the text in German ($y_1, y_2, \ldots, y_k$), i.e. the first difference will be less than the second one.

This method was used for diagnostic of the authorship of literary and musical texts, for estimation of closeness of DNA sequences, construction of phylogenetic trees and many other problems ([7, 1, 2, 8, 10, 11, 4]). Many papers (see [8] and review there) were devoted to the measurement of the interdependence between sequences (or the association, similarity, closeness, etc.). It is important to note, that their approaches are outside of the framework of mathematical statistics and, in particular, do not give a possibility to reason about consistency of estimates, tests, classifiers, clustering, etc.

2. THE MAIN RESULTS
In the report we give a compression-based solution for the following problems:

i) Homogeneity test, where there are several sequences $x_1, x_2, \ldots, x_{n_1}$, $\ldots$, $x_1, x_2, \ldots, x_{n_s}$, $y_1, y_2, \ldots, y_{m_1}$, $\ldots$, $y_1, y_2, \ldots, y_{m_s}$ generated either by a single source or by two different ones, and two corresponding hypotheses. We also consider the more general case where there are more than two different sets of sequences.

ii) Classification problems, where there are samples $x_1, x_2, \ldots, x_{n_1}$, $\ldots$, $x_1, x_2, \ldots, x_{n_s}$, $y_1, y_2, \ldots, y_{m_1}$, $\ldots$, $y_1, y_2, \ldots, y_{m_s}$ generated two different (but unknown) sources and $z_1, z_2, \ldots, z_l$ is generated by one of the two. The goal is to determine which of them generated $z_1, z_2, \ldots, z_l$.

iii) Estimation of a so-called measurement of interdependence, or the association.

A distinction of the suggested method from other approaches is that it belongs to the framework of mathematical statistics.

3. REFERENCES


CONDITIONAL MUTUAL INFORMATION ESTIMATION
AND ITS APPLICATION TO CONDITIONAL INDEPENDENCE DETECTION

Joe Suzuki

Division of Mathematical Science, Osaka University, Toyonaka, Osaka 560-8531, Japan
j-suzuki@sigmath.es.osaka-u.ac.jp

ABSTRACT

This paper considers to estimate (conditional) mutual information for general random variable not limited to discrete variables, and proves under a regular its condition consistency and asymptotic correctness of (conditional) independence testing.

1. INTRODUCTION

We consider to estimate mutual information \( I(X, Y) \) given \( n \) pairs of examples \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).

Suppose that \( X \) and \( Y \) take finite numbers of values. Then, some might estimate \( I(X, Y) \) as

\[
I_n = \sum_{x} \sum_{y} \frac{c_{XY}(x, y)}{n} \log \frac{c_{XY}(x, y)}{c_X(x) c_Y(y) / n} \tag{1}
\]

given the frequencies \( c_X(x), c_Y(y) \), and \( c_{XY}(x, y) \) of \( X \), \( Y \), and \( (X, Y) \), respectively. On the ground that \( I_n \) converges to \( I(X, Y) \) with probability one as \( n \to \infty \). However, we note that even when \( X \) and \( Y \) are independent, \( I_n > 0 \) with a positive probability as \( n \to \infty \), whereas

\( I(X, Y) = 0 \iff X \perp Y \),

where we write \( X \perp Y \) when \( X \) and \( Y \) are independent. More precisely, we have \( I_n > 0 \) unless \( c_X(x)c_Y(y) = nc_{XY}(x, y) \) for all \( x, y \), which is unlikely even for large \( n \).

The author proposed an alternative estimation of mutual information [5]. The alternative estimator is

\[
J_n := I_n - \frac{l}{2n} \log n \tag{2}
\]

with \( l = (\alpha - 1)(\beta - 1) \), where \( \alpha \) and \( \beta \) are the numbers of values that \( X \) and \( Y \) take, respectively. The values of description length of \( X, Y \), and \( (X, Y) \) are, up to constants,

\[
\sum_{x} c_X(x) \log \frac{c_X(x)}{n} + \frac{\alpha - 1}{2} \log n , \tag{3}
\]

\[
\sum_{y} c_Y(y) \log \frac{c_Y(y)}{n} + \frac{\beta - 1}{2} \log n , \tag{4}
\]

and

\[
-\sum_{x} \sum_{y} c_{XY}(x, y) \log \frac{c_{XY}(x, y)}{n} + \frac{\alpha \beta - 1}{2} \log n , \tag{5}
\]

respectively, so that (2) is obtained from (1) and \( \{ (3) + (4) - (5) \} / n \). From consistency of MDL [4], we have \( J_n \leq 0 \) with probability one as \( n \to \infty \). If we redefine \( J_n \) by

\[
J_n := \max \{ I_n - \frac{l}{2n} \log n, 0 \} \tag{6}
\]

then

\( J_n = 0 \iff X \perp Y \) \tag{7}

with probability one as well as

\[
J_n \to I(X, Y) \tag{8}
\]

as \( n \to \infty \). Figure 1 illustrates the difference between \( I_n \) and \( J_n \).

The main issue of this paper is an extension of the estimation to the general case that contains continuous and countable infinite variables. For example, if \( X \) and \( Y \) follow the Gaussian distribution, the problem reduces to the estimation of correlation between them:

\[
I(X, Y) = -\frac{1}{2} \log \{1 - \rho(X, Y)^2\} ,
\]

where \( \rho(X, Y) \) is the correlation coefficient of \( X \) and \( Y \). Apparently, this does not help for the general case. In particular, the estimation for discrete variables is not covered by the estimation of correlation.

We consider to estimate mutual information \( I(X, Y) \) without distinguishing whether \( X \) and \( Y \) are discrete or continuous. To this end, we prepare a sequence of contingency tables each of which expresses the frequencies of quantized values of \( X \), \( Y \), and \( (X, Y) \) such that each of rows \( X \) and each of columns \( Y \) contain almost equal sample sizes (the difference is at most one, see Figure 2), respectively. As the sequence proceeds, the contingency table is monotonically refined so that each cell contains fewer samples (Figure 3).

If two samples with the same value of \( X \) is divided into different quantized value of \( X \), the border crossing them will be adjusted as in Figure 5. If two borders coincide, we regard them as one border. In this sense, if
In Figure 1, the values of $I_n$ and $J_n$ when the mutual information values are zero (Top) and positive (Bottom). 500 pairs of binary sequences of length 200 are generated so that $X$ and $Y$ occur equiprobably. The true mutual information values are zero and 0.368 nats with $P(X \neq Y) = 0.5$ and $P(X \neq Y) = 0.1$, respectively.

\begin{align*}
1 & 2 3 4 5 6 7 8 \\
1 & 100 73 37 23 10 7 0 0 \\
2 & 20 20 56 49 41 30 12 2 \\
3 & 5 12 26 42 49 49 45 22 \\
4 & 0 0 6 11 25 39 68 101
\end{align*}

Figure 2. The space is divided into $4 \times 8$ for $n = 1000$ in which each column of $X$ and each row of $Y$ contain 125 and 250 samples, respectively. However, the frequencies of $(X, Y)$ differ.

\begin{align*}
& \begin{array}{cccc}
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\end{array} & \begin{array}{cccc}
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\end{array} & \begin{array}{cccc}
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\end{array} & \begin{array}{cccc}
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\text{Y} & \text{Y} & \text{Y} & \text{Y} \\
\text{X} & \text{X} & \text{X} & \text{X} \\
\end{array}
\end{align*}

Figure 3. The two-dimensional Euclid space of $X$ and $Y$ is sequentially divided and a nested quantization sequence is obtained.

\begin{align*}
& 13 \rightarrow 13 \rightarrow 13 \rightarrow 13 \\
& 24 \rightarrow 24 \rightarrow 24 \rightarrow 24
\end{align*}

Figure 4. The Chow-Liu Algorithm when $I(1, 2) > I(1, 3) > I(2, 3) > I(1, 4) > I(3, 4) > I(2, 4)$.

\begin{align*}
\begin{array}{cccc}
& z_1 & z_2 & z_3 \\
& 1 & 2 & 2 \\
& 2 & 2 & 3 \\
& 3 & 1 & 2 \\
\end{array}
\end{align*}

Figure 5. When the quantization tries to divide $z_2$ and $z_3$ and finds $z_2 = z_3 = 2$ (Left), because the pair $(z_1, z_2)$ is closer than $(z_4, z_5)$ from $(z_2, z_3), (z_1, z_2)$ is chosen as the border instead of $(z_2, z_3)$. On the other hand, when $z_3 = z_4 = 2$ (Right), the pair $(z_4, z_5)$ is chosen as the border.

\begin{align*}
J^{(1)}_n & \hspace{1cm} J^{(2)}_n \\
& \hspace{1cm} J^{(3)}_n \\
& \hspace{1cm} J^{(m_n)}_n
\end{align*}

Figure 6. Generate $m_n$ quantizations and choose the maximum value $J_n = \max_{k=1}^{m_n} J^{(k)}_n$ of $m_n$ mutual information estimations $J^{(k)}_n, k = 1, \cdots, n$.

$X$ takes $\alpha$ values, at most $\alpha - 1$ borders will be generated for $X$. The same rule is applied to $Y$ as well as to $X$.

We assume that the length $m_n$ of the contingency table sequence is at most $n$, and note that from each contingency table $k = 1, 2, \ldots, m_n$, the estimation $J^{(k)}_n$ of mutual information is obtained by regarding that $X$ and $Y$ take $\alpha^{(k)}$ and $\beta^{(k)}$ value, respectively (Figure 6). The final estimation will be

$$J_n := \max \{0, J^{(1)}_n, J^{(2)}_n, \ldots, J^{(m_n)}_n\}$$

Even under the extended conditions, (7) has been proved in [8].

The algorithm has been already published [7, 8] with applications of independence testing [7] and genome analysis [8]. The procedures can be executed via CRAN package BNSL [9]. Using the Chow-Liu algorithm (Figure 4) [2] in which a forest is constructed by connecting a pair of variables in the ascending order of the mutual information values unless any loop is generated, the extended mutual information procedure constructed gene regular network with 1000 genes [8] that minimize the description length.

The same idea is applied to estimating the conditional mutual information $I(X, Y | Z)$ of variables $X$ and $Y$ given another variable $Z$. Suppose that $X$, $Y$, and $Z$ take $\alpha$, $\beta$, and $\gamma$ values. The maximum likelihood estimator of conditional mutual information is given by

$$I'_n = \sum_{x, y, z} \frac{c_{xyz}(x, y, z)}{n} \log \frac{c_{xyz}(x, y, z)}{c_{x}(x)\cdot c_{y}(y)\cdot c_{z}(z)}$$

and the author proposed the estimator [6]

$$J'_n := I'_n - \frac{p'}{2n} \log n$$

(9)
with \( l' := (\alpha - 1)(\beta - 1)\gamma \), where we write \( X \perp \perp Y | Z \) when \( X \) and \( Y \) are conditionally independent given \( Z \), and proved the two properties.

\[
J'_n = 0 \iff X \perp \perp Y | Z
\]

with probability one as well as

\[
J'_n \to I(X, Y | Z)
\]

as \( n \to \infty \).

For the extension to general variables that can take continuous variables, we can execute it similarly to the mutual information estimation: prepare a sequence of three dimensional contingency tables. As the sequence proceeds, each cell is monotonically refined and contains fewer samples.

We assume that the length \( m_n \) of the contingency table sequence is at most \( n \), and note that from each contingency table \( k = 1, 2, \ldots, m_n \), the estimation \( J'_n \) of conditional mutual information is obtained by regarding that \( X \) and \( Y \) take \( n^{(k)} \) and \( \beta^{(k)} \) value, respectively (Figure 6). The final estimation will be

\[
J'_n := \max \{0, J^{(1)}_n, J^{(2)}_n, \ldots, J^{(m_n)}_n\}
\]

Even under the extended conditions, (10) has been proved in [8].

While (7) and (10) have been proved in [8] for the general case, this paper proves the unsolved statements (8) and (11). To this end, we derive tighter bounds on the error probabilities \( P(J_n > 0) \) and \( P(J'_n > 0) \) for \( X \perp \perp Y \) and \( X \perp \perp Y | Z \), respectively. We will find that the new bounds make proving (7) and (10) easier as well, and briefly re-prove them using the bounds.

Seeking the maximum mutual information estimation over the contingency tables \( k = 1, 2, \ldots, m_n \) is used in maximal information coefficient (MIC, [3]) that maximizes \( I_{n}^{(k)} \) rather than \( J_n^{(k)} \). The MIC does not guaranteed any property as (7) and (8). In this paper, we mathematically prove the merits of the estimation \( J_n \) and \( J'_n \).

2. RESULTS

In this section, we prove important properties on the estimators \( J_n \) and \( J'_n \) of mutual information and conditional mutual information given \( n \) examples. We utilize the following propositions for proving the theorems (the proofs will appear in the near future).

**Proposition 1** Suppose \( X \) and \( Y \) are independent, then

\[
P(J_n > 0) \leq \left( \frac{\epsilon \log n}{n} \right)^l
\]

**Proposition 2** Suppose \( X \) and \( Y \) are conditionally independent given \( Z \), then

\[
P(J'_n > 0) \leq \left( \frac{\epsilon \log n}{n} \right)^{l'}
\]

For \( k = 1, 2, \ldots, m_n \), let \( X^{(k)} \) and \( Y^{(k)} \) be the variables \( X \) and \( Y \) in the \( k \)th contingency table, \( \alpha^{(k)} \) and \( \beta^{(k)} \) the numbers of values that \( X^{(k)} \) and \( Y^{(k)} \) take, respectively. Also, for \( k = 1, 2, \ldots, m_n \), let \( I^{(k)} := (\alpha^{(k)} - 1)(\beta^{(k)} - 1) \), and \( I(X, Y) := I(X^{(k)}, Y^{(k)}) \).

2.1. Asymptotic Correctness of Independence Testing

**Theorem 1** Let \( X \) and \( Y \) be any random variables. \( X \perp \perp Y \) if and only if the event \( J'_n > 0 \) finitely occurs almost surely.

Proof. For (\( \Leftarrow \)), if \( X \perp \perp Y \), then we have \( I(X, Y) > 0 \) for at least one \( k \). From (8), \( J'_n \) converges to \( I(X, Y) > 0 \) with probability one as \( n \to \infty \). Thus, \( J'_n > 0 \) occurs infinitely many times with probability one, which establishes (\( \Leftarrow \)).

Let \( K \) be the maximum \( k \) such that \( I^{(k)} \leq 4 \). Then, from (7), for each \( k \), the event \( J'_n > 0 \) finitely occurs almost surely. Since \( K \) does not depend on \( n \), the event \( \bigcup_{k=1}^{K} (J'_n > 0) \) finitely occurs almost surely. From Proposition 1, we have

\[
P(\bigcup_{k=1}^{K} (J'_n > 0)) \leq m_n P(J'_n(K+1) > 0) \leq n(\frac{\epsilon \log n}{n})^{(K+1)/2}
\]

which means

\[
\sum_{n=1}^{\infty} P(\bigcup_{k=1}^{m_n} (J'_n(K) > 0)) < \infty
\]

From Borel-Cantelli’s Lemma [1], the event \( \bigcup_{k=1}^{m_n} (J'_n(K) > 0) \) finitely occurs almost surely as well, which means that \( \bigcup_{k=1}^{m_n} (J'_n > 0) \) finitely occurs almost surely. This establishes (\( \Leftarrow \)).

**Theorem 2** Let \( X, Y, \) and \( Z \) be any random variables. \( X \perp \perp Y | Z \) if and only if the event \( J'_n > 0 \) finitely occurs almost surely.

Proof. An almost similar proof to Theorem 1 will suffice.

2.2. Consistency

**Theorem 3** If there exists \( K \geq 1 \) such that for each \( r = 1, 2, \ldots, \)

\[
I(X^{(K+r)}, Y^{(K+r)}) = I(X^{(K)}, Y^{(K)})
\]

Then, the estimation \( J_n \) converges to the mutual information \( I(X, Y) \).

Since \( I^{(k)}(X, Y) \) is upper bounded by \( I(X, Y) \) and monotonically increasing with \( k \), we have

\[
\lim_{k \to \infty} I^{(k)}(X, Y) = I(X, Y)
\]

so that the assumption may be reasonable.
Proof. Note for each \( r = 1, 2, \ldots \)
\[
I^{(r)}(X, Y) = I^{K+r}(X, Y)
\]
\[
\iff P(X(K), Y(K)) = P(X(K+r), Y(K+r))
\]
\[
\iff \begin{cases} 
X^{(K+r)} \perp \perp Y^{(K+r)} | \{X^{(K)}, Y^{(K)}\} \\
X^{(K)} \perp \perp Y^{(K)} | \{X^{(K)}, Z^{(K)}\} \\
X^{(K)} \perp \perp Y^{(K)} | \{Y^{(K)}, Z^{(K+r)}\}
\end{cases}
\]
and its graphical model is expressed by Figure 7.

Let \( J_n^A, J_n^B, \) and \( J_n^C \) be the conditional mutual information estimations in the form of (9) for
\[
I(X^{(K+r)}, Y^{(K+r)}) | X^{(K)}, Y^{(K)}
\]
\[
I(X^{(K+r)}, Y^{(K+r)}) | X^{(K)}, Y^{(K)}
\]
and \( I_n^A, I_n^B, \) and \( I_n^C \) be the associated maximum likelihood estimations. More precisely, \( J_n^A = I_n^A = \frac{A}{2n} \log n, \)
\[
J_n^B = I_n^B = \frac{B}{2n} \log n, \quad \text{and} \quad J_n^C = I_n^C = \frac{C}{2n} \log n,
\]
where
\[
I_n^A = (\alpha(K+r) - \alpha(K))(\beta(K+r) - \beta(K))
\]
\[
I_n^B = (\alpha(K+r) - \alpha(K))(\beta(K) - 1), \quad \text{and} \quad I_n^C = (\alpha(K) - 1)(\beta(K+r) - \beta(K)).
\]

From inspection, one checks
\[
I_n^{(K+r)} - I_n^{(K)} = I_n^A + I_n^B + I_n^C,
\]
and
\[
I_n^A + I_n^B + I_n^C = (\alpha(K+r) - 1)(\beta(K+r) - 1) - (\alpha(K) - 1)(\beta(K) - 1),
\]
so that we have
\[
J_n^{(K+r)} - J_n^{(K)} = J_n^A + J_n^B + J_n^C.
\]
which means
\[
P(J_n^{(K)} < J_n^{(K+r)}) \leq \frac{P(J_n^A > 0) + P(J_n^B > 0) + P(J_n^C > 0)}{n^{2 \log n}} \leq \left( \frac{e \log n}{n^2} \right)^A + \left( \frac{e \log n}{n^2} \right)^B + \left( \frac{e \log n}{n^2} \right)^C.
\]
Similar to the discussion in the proof of Theorem 1, for \( r = 1, 2, \ldots \) the event \( J_n^{(K)} < J_n^{(K+r)} \) occurs at most finite time with probability one as \( n \to \infty \).

On the other hand, \( J_n^{(K)} \) converges to \( I^{(K)}(X, Y) \) with probability one as \( n \to \infty \), for each \( k = 1, 2, \ldots \). This completes the proof.

**Theorem 4** If there exists \( K \geq 1 \) such that for each \( r = 1, 2, \ldots \),
\[
I(X^{(K+r)}, Y^{(K+r)} | Z^{(K+r)}) = I(X^{(K)}, Y^{(K)} | Z^{(K)})
\]
then, the estimation \( J_n^{(K)} \) converges to the mutual information \( I(X, Y | Z) \).

Proof. The derivation is similar to Theorem 3. Simply replace the three conditional independence statements by
\[
\begin{cases}
X^{(K+r)} \perp \perp Y^{(K+r)} | \{X^{(K)}, Y^{(K)}, Z^{(K+r)}\} \\
X^{(K+r)} \perp \perp Y^{(K)} | \{X^{(K)}, Z^{(K+r)}\} \\
X^{(K)} \perp \perp Y^{(K)} | \{Y^{(K)}, Z^{(K+r)}\}
\end{cases}
\]

Figure 7. Given \( X^{(K)} \) and \( Y^{(K)} \), the pair \( X^{(K+r)} \) and \( Y^{(K+r)} \) are independent; given \( X^{(K)} \), the pair \( X^{(K+r)} \) and \( Y^{(K)} \) are independent; and given \( Y^{(K)} \), the pair \( X^{(K)} \) and \( Y^{(K+r)} \) are independent.

**Acknowledgements**

This work was supported by JSPS Core-to-Core Program, A. Advanced Research Networks ('establishing International Research Network of Mathematical Oncology’, Prof. Takashi Suzuki, Osaka University).

**3. REFERENCES**


THE INFORMATION THEORY OF DEEP LEARNING

Naftali Tishby

The Hebrew University of Jerusalem, Israel

ABSTRACT

Based on analytic and numerical studies, I will describe novel theoretical understanding of Deep Neural Network (DNN) based on the analysis of the representations of the layers in the Information Plane - the mutual information between each layer and the input layer and the desired output of the layer. I will show that the standard Stochastic Gradient Descent (SGD) training of DNNs has two distinct phases: (1) fast drift with fitting the training data and thus increase the mutual information between the layers and the desired tables; (2) slow diffusion which compress the representation of the input layer and reduce the mutual information between the layers and the input. The compression phase, which - as we prove - dramatically improves the generalization power of the network, is in fact the more important aspect of SGD. It also provides a new understanding of the benefit of the hidden layers, the lack of overfitting, and the features the layers represent.

Based on joint works with Ravid Schwartz-Ziv and Noga Zaslavsky
RATE-DISTORTION DIMENSION AND BAYESIAN LEARNING COEFFICIENT

Kazuho Watanabe

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

ABSTRACT

Rate-distortion theory analyzes fundamental limits of lossy compression methods. In this study, we discuss the connection of rate-distortion theory to Bayesian learning. We formulate a rate-distortion problem by the distortion measure defined by the pointwise regret of the model parameter, and show that the generalized Bayesian posterior distribution appears as the optimal solution to the problem. We also consider another rate-distortion problem by replacing the empirical expectation with the true expectation, and show the relationships between the asymptotic behaviors of the rate-distortion functions and the learning coefficient of Bayesian learning, which characterizes the model’s generalization ability.

1. INTRODUCTION

Rate-distortion (RD) theory studies and characterizes the performance of lossy compression systems by the RD function [1]. There are studies interpreting learning and prediction problems by rate-distortion theoretic views. In particular, clustering methods can be considered as vector quantizers and naturally related to RD theory [2]. RD theoretic interpretations have been obtained also for problems such as classification [3] and sequence prediction [4].

In this study, we provide an RD theoretic view of Bayesian learning, and show that its optimal solution is given by the generalized posterior distribution, which has been used for purposes such as the posterior consistency [5, 6] and model selection [7, 8, 9]. We also formulate an RD problem with a fixed prior as in the practical scenario of Bayesian learning, and discuss its connection to the asymptotic theory of Bayesian learning. The generalization ability of a learning machine in Bayesian learning is characterized by a constant called the learning coefficient [10]. There is an interpretation of the learning coefficient as a volume dimension [11, 10]. In RD theory, the asymptotic behavior of an RD function in the small distortion limit is characterized by the RD dimension [12]. We provide an upper bound for the RD dimension of the RD problem with a fixed prior, which is related to the learning coefficient and yet another quantity called the singular fluctuation. We also formulate another RD problem by replacing the empirical expectation of data samples with the true expectation, and show that the RD dimension of this problem is upper bounded by twice the learning coefficient.

2. RATE-DISTORTION FUNCTION

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

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

where

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

is the mutual information between \( X \) and \( Y \), 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 \( d(1,13) \).

If there exists the conditional distribution \( q_s^* \) that achieves the minimum of the following free energy functional parameterized by \( s \geq 0 \),

\[
F(q(y)) = E_p \left[ - \log \int \exp(-sd(X,y))q(y)dy \right],
\]

then, the rate-distortion function is parametrically given by

\[
R(D_s) = I(q_s^*(y|x)),
\]

\[
D_s = E_{q_s^*(y|x)p(x)}[d(X,Y)],
\]

where

\[
q_s^*(y|x) = \frac{\exp(-sd(x,y))q_s^*(y)}{\int \exp(-sd(x,y))q_s^*(y)dy}.
\]

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

\[
D_{\text{max}} = \inf_y E_p[d(X,y)],
\]

and \( R(D) = 0 \) for \( D \geq D_{\text{max}} \) [1].
3. BAYESIAN LEARNING

Given training data \( x = \{ x_1, \ldots, x_n; x_i \in X_0, \forall i \} \) in the subset of \( X_0 \) of a finite dimensional Euclidean space, we assume the statistical model \( \{ p(x|\theta); \theta \in \Theta \subset R^m \} \) with an \( m \)-dimensional parameter \( \theta \). We further assume that the data are i.i.d., \( p(x) = \prod_{i=1}^n p(x_i) \) and \( p(x|\theta) = \prod_{i=1}^n p(x_i|\theta) \) and that the maximum likelihood estimator (MLE), \( \hat{\theta}(x) \), for the data \( x \) exists for all \( x \).

3.1. Rate-Distortion Theoretic Framework

Let us consider learning \( \theta \) from \( x \) through the following RD problem. The training data set \( x \in X_0^n \) is an instance of the source output, \( X = X_0^n \), and the parameter \( \theta \) is viewed as the reconstruction of \( x \), \( Y = \Theta \). The distortion measure between them is defined by

\[
d(x, \theta) = \log \frac{p(x|\hat{\theta}(x))}{p(x|\theta)}, \tag{3}
\]

the regret of \( \theta \) for the data set \( x \).

Assume that there exists

\[
\theta_0 = \arg\min_{\theta} \text{KL}(p(x)||p(x|\theta)),
\]

the minimizer of the Kullback-Leibler divergence from the source. Then, the minimum distortion at zero rate is

\[
D_{\text{max}} = E_X[\log p(X|\theta(X))] - E_X[\log p(X|\theta_0)], \tag{4}
\]

where the second term reduces to the source entropy, \( nh(p) = nE_X[-\log p(X)] \) if the source is realizable by the model, \( p(x) = p(x|\theta_0) \). Here \( E_X \) and \( E_X^\theta \) denote the expectations with respect to the source densities \( p(x) \) and \( p(x|\theta) \), respectively. \( D_{\text{max}} \) in (4) is the (negative) training error of the MLE normalized by the second term.

For the distortion levels smaller than \( D_{\text{max}} \), if there exists

\[
q^*_s(\theta) = \arg\min_{q(\theta)} E_X \left[ -\log \int p(X|\theta)^s q(\theta) d\theta \right], \tag{5}
\]

for \( s > 0 \), then the RD function is given by

\[
R(D,s) = E_X \left[ -\log \int p(X|\theta)^s q^*_s(\theta) d\theta \right] + sE_X E_{\theta_0} \left[ \log p(X|\theta_0) \right], \tag{6}
\]

\[
D_s = E_X E_{\theta_0} \left[ \log \frac{p(X|\theta(x))}{p(X|\theta_0)} \right], \tag{7}
\]

where \( E_{\theta_0} \) represents the expectation with respect to \( q^*_s(\theta|x) \propto p(x|\theta)^s q^*_s(\theta) \).

3.2. Learning Coefficient and Rate-Distortion Dimension

If we fix a prior distribution \( q(\theta) \), as in the usual Bayesian learning, then we have the following upper bound of \( R(D) \),

\[
\overline{R}(D_s) \equiv E_X \left[ \log \frac{p(X|\hat{\theta}(X))^s}{p(X|\theta)^s q(\theta) d\theta} \right] - sD_s, \tag{8}
\]

\[
= E_X \left[ -\log Z_s(X) + sE_{\theta_0} [\log p(X|\theta)] \right], \tag{9}
\]

where \( D_s \) is the average Gibbs training error relative to the MLE,

\[
Z_s(x) = \int p(x|\theta)^s q(\theta) d\theta
\]

is the generalized evidence, and \( E_{\theta_0} \) denotes the expectation with respect to the generalized posterior distribution,

\[
q_s(\theta|x) = \frac{p(x|\theta)^s q(\theta)}{Z_s(x)}. \tag{10}
\]

The generalized posterior distribution has been used for purposes such as guaranteeing posterior consistency [5, 6] and model selection [7, 8, 9]. Note here that the slope parameter \( s \) shows the negative slope of the tangent of \( \overline{R}(D) \), that is, \( \overline{R}(D_s) = -s \) holds.

For the following discussions, we assume the fundamental conditions (I) and (II) of [10]. Under these conditions, it was proved that the generalization performance of Bayesian learning is characterized by the coefficient \( \lambda \) which is referred to as the learning coefficient, and is also the largest pole of the zeta function,

\[
\int KL(p(x|\theta_0)||p(x|\theta))^z q(\theta) d\theta
\]

of \( z \in C \) [10]. More specifically, the following asymptotic expansion of the (generalized) log-evidence holds [10, Corollary 6.1]

\[
E_X \left[ -\log Z_s(X) \right] = nsh(p) + \lambda \log n + O(\log \log n).
\]

It was also proved that the Bayes generalization error has the following asymptotic expansion,

\[
E_X \left[ KL(p(x|\theta_0)||p(x|X)) \right] \propto \frac{\lambda}{n} + o \left( \frac{1}{n} \right),
\]

where \( p(x|X) = E_{\theta_0}[p(x|\theta)] \) is the Bayes predictive distribution obtained from the posterior \( q_1(\theta|x) \) with \( s = 1 \) for the given data set \( x \) [10, Theorem 6.8 and Corollary 6.3].

The learning coefficient depends on the model, prior and source. If the model is regular, \( 2\lambda = m \), the dimension of the parameter, whereas if it is non-regular, \( 2\lambda \leq m \) holds in general.

Since the distortion measure (3) is a squared metric in the case of the Gaussian model, if we apply the same definition of the RD dimension [12] to other models, the RD dimension of \( R(D) \) is defined by

\[
\dim_R \equiv \lim_{D \to 0} \frac{R(D)}{-\frac{1}{2} \log D}. \tag{10}
\]

In addition to the above fundamental conditions, we assume the following.
**Assumption 1** For the distortion $D_s$ in (8), it holds that,
\[
\tilde{D}_s \equiv \lim_{n \to \infty} sD_s \leq C, \quad \text{and} \quad \lim_{s \to \infty} \tilde{D}_s = C \quad (11)
\]
for a constant $C > 0$.

A sufficient condition for (11) is that the Gibbs training error $E_0[-\log p(x|\theta)]$ converges to that of the MLE, $-\log p(x|\hat{\theta}(x))$ as $s \to \infty$ for all $x$ in the order of $1/s$.

Then, we obtain the following theorem.

**Theorem 1**
\[
\lim_{D \to 0} \frac{1}{s} \log D \geq \frac{1}{s} \log \frac{1}{s} + \sum_{i=1}^{s} \nu(i) \quad (12)
\]
for the singular fluctuation,
\[
\nu(s) = \lim_{n \to \infty} \frac{s}{n} E_X [V_{\theta_0}[\log p(X|\theta)]]
\]
where $V_{\theta_0}$ denotes the variance with respect to $q_{\theta_0}(\theta|X)$.

**3.3. True Expectation RD Problem**

By replacing the empirical expectation in the RD problems defined in the previous sections with the true expectation with respect to $p(x|\theta_0)$, we can define the true expectation RD problem where the distortion measure is
\[
d(\theta_0, \theta) = nKL(p(x|\theta_0)||p(x|\theta)),
\]
and the source is identified with the true prior distribution $p(\theta_0)$ on $\Theta$. The optimal reconstruction distribution for this problem is
\[
q_{\theta}(\theta) = \frac{1}{Z_{\theta}^* \theta_0} \exp(-nKL(p(x|\theta_0)||p(x|\theta)))q(\theta),
\]
where $Z_{\theta}^* \theta_0 = \int \exp(-nKL(p(x|\theta_0)||p(x|\theta)))q(\theta)d\theta$ for the optimal or a fixed prior $q(\theta)$. The same problem was discussed by [5] without the expectation with respect to $p(\theta_0)$.

Let $R^{*}(D)$ and $\overline{R}^{*}(D)$ be the RD function and its upper bound obtained for the optimal priors $\{q_{\theta}^{*}(\theta)\}$ and a fixed prior $q(\theta)$ as the upper bound (7) in the case of the empirical expectation problem. Then, we have the following upper bound for the RD dimension of the true expectation problem.

**Theorem 2** If the support of the true prior $p(\theta_0)$ is a subset of $\Theta_0 \equiv \{\theta \in \Theta: KL(p(x)||p(x|\theta)) = 0\}$, then for any natural number $n$,
\[
\lim_{D \to 0} \frac{R^{*}(D)}{\log D} - \frac{n}{2} \leq \frac{\overline{R}^{*}(D)}{\log D} - \frac{n}{2} = 2\lambda,
\]
where $\lambda$ is the learning coefficient, and
\[
F^{*}(n) = E_{\theta_0}[ -\log Z_{\Theta}^* (\theta_0)].
\]
The proof is similar to the derivation of the asymptotic expansion of $F(n)$ in (12)[10]. Note that $F(n) \leq F^{*}(n)$ holds and the difference between them is a constant.

**3.4. Shannon-Bregman Lower Bound**

If the model $p(x|\theta)$ is a member of the exponential family, it is represented by the distribution on a finite number of sufficient statistics. Let $T = T(x)$ be the $m$-dimensional sufficient statistic determined from data $x$. Then the RD problem in Sect. 3 is reduced to that of the sufficient statistics since $I(X; \theta) = I(T; \theta)$ and the distortion measure (3) can also be expressed in terms of the value of the sufficient statistic $t$. For the exponential family model, a lower bound to the RD function, named Shannon-Bregman lower bound, was derived as follows [14].

Let
\[
p_{\text{Bayes}}^{(s)}(t) = \phi_{s}(t) \int p(t|\theta)^s q(\theta)d\theta,
\]
where $\phi_{s}(t)$ is a function satisfying
\[
\int p(t|\theta)^s \phi_{s}(t)dt = 1
\]
for all $\theta$. If such a function $\phi_{s}(t)$ exists, we have the following lower bound to $R(D)$,
\[
R(D) = \sup_{s \geq 0} \inf_{q(\theta)} \left\{ KL(p||p_{\text{Bayes}}^{(s)}) + E_{T}[\log \phi_{s}(T)] + h(p(t)) - sE_{T}[\log p(T|\theta(T))] - sD \right\}
\]
\[
\geq \sup_{s \geq 0} \left\{ E_{T}[\log \phi_{s}(T)] + h(p(t)) - sE_{T}[\log p(T|\theta(T))] - sD \right\}.
\]
The lower bound is tight at the slope parameter $s$ if and only if
\[
p(t) = p_{\text{Bayes}}^{(s)}(t) \quad (\forall t),
\]
that is, the source is expressed as a Bayes mixture by some prior $q(\theta)$.

**4. EXAMPLES**

There are corresponding results in RD theory for the Gaussian model with only the mean parameter [12] and the exponential model [16]. In these examples, the Shannon-Bregman lower bound is tight for all $D$, and
hence the RD-function is explicitly obtained. Furthermore, \( \dim_R \) is equal to twice the learning coefficient, which is also the dimension of the parameter space.

In the case of the Bernoulli model, the optimal reconstruction distribution is given by a discrete distribution for each distortion level, which can be computed by an iterative algorithm similar to the expectation-maximization algorithm [17, 14]. The RD dimension of the RD function is zero whereas \( 2\lambda = 1 \) if the prior is continuous.

As an example of the simplest latent variable models, we consider the binomial mixture model,

\[
p(x|\theta) = a \text{Bin}_N(x|b) + (1 - a) \text{Bin}_N(x|0.5),
\]

where \( \theta = (a, b) \in \Theta = [0, 1] \times [0, 1] \) and \( \text{Bin}_N(x|b) = \binom{N}{x} b^x (1 - b)^{N-x} \) is the binomial probability for \( x \in \{0, 1, \ldots, N\} \). The distortion measure (3) is

\[
d(x, \theta) = \sum_{i=1}^{n} \log \frac{1 - \hat{a} + \hat{a}(2b)^x_i (2 - 2b)^{N-x_i}}{1 - a + a(2b)^x_i (2 - 2b)^{N-x_i}},
\]

where the maximum likelihood estimate \( \hat{b}(x) = (\hat{a}, \hat{b}) \) was computed by maximizing the log-likelihood \( \sum_{i=1}^{n} \log (1 - a + a(2b)^x_i (2 - 2b)^{N-x_i}) \) over 1000 \times 1000 grid points on \([0, 1] \times [0, 1] \) for each \( x \). We can numerically verify the relationship between the learning coefficient and the rate-distortion dimension for this model.

5. CONCLUSION

In this study, we formulated an RD theoretic view of Bayesian learning. We also related the learning coefficient to the RD dimensions of the empirical and true expectation RD problems. We examined these relationships in regular statistical models analytically and in the binomial mixture model numerically. A method for computing the learning coefficient based on these relationships is to be further explored.

6. ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI 25120014, 15K16050, and 16H02825.

7. REFERENCES


DIFFERENCE BETWEEN BAYES CROSS VALIDATION AND WAIC FOR CONDITIONAL INDEPENDENT SAMPLES

SUMIO WATANABE¹

¹ Department of Mathematical and Computing Science, Tokyo Institute of Technology
Mailbox G5-19, Nagatsuta Midoriku Yokohama 226-8502, JAPAN, swatanab@c.titech.ac.jp

ABSTRACT

In Bayesian estimation, if a sample consists of independent random variables, then the leave-one-out cross validation (LOOCV) and the widely applicable information criterion (WAIC) are asymptotically equivalent as estimators of the generalization loss. However, if otherwise, then such equivalence does not hold in general. In this paper, we show that, if a sample consists of conditionally independent random variables, then the average of WAIC is asymptotically equal to that of the generalization loss, whereas that of LOOCV is not. Consequently, WAIC is the better estimator of the generalization loss than LOOCV for conditional independent samples.

1. INTRODUCTION

In statistical inference, the generalization loss is defined by the average minus log loss function of the estimated probability density function, which is a measure of the accuracy of statistical estimation. There are two general measures by which the generalization loss is estimated: one is the leave-one-out cross validation (LOOCV) and the other is the information criterion. There are many researches on cross validation and information criteria [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

In the maximum likelihood method, if a true distribution is realizable by and regular for a statistical model, then both LOOCV and Akaike information criterion (AIC) can estimate the generalization loss. In the Bayesian method, even if a true distribution is unrealizable by and singular for a statistical model, then both LOOCV and the widely applicable information criterion (WAIC) can also estimate the generalization loss.

However, if such equivalence requires independence of random variables in a sample. In fact, if a sample contains dependent variables, then LOOCV can not estimate the generalization loss. In this paper, we study Bayesian estimation in a case when a sample is consists of conditionally independent random variables, and show that WAIC can estimate the generalization loss whereas LOOCV not.

2. MAIN RESULTS

Let \( x_i \in \mathbb{R}^M \) for \( i = 1, 2, ..., n \) and \( x^n = (x_1, x_2, ..., x_n) \in (\mathbb{R}^M)^n \) is fixed or may be dependent. It is assumed that random variables \( Y^n = (Y_1, Y_2, ..., Y_n) \) are independently subject to a true conditional distribution

\[
\prod_{i=1}^{n} q(y_i|x_i).
\]

Then \( (Y_1, Y_2, ..., Y_n) \) are independent, however,

\[
((x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n))
\]

are dependent. This conditional independent condition allows the following cases.

1. The set \( x^n \) consists of fixed points.
2. The set \( x^n \) is a time sequence.
3. The point \( x_i \) depends on other \( \{x_j; j \neq i\} \).

If a sample consists of independent variables, then prediction and estimation is mathematically equivalent, however, if otherwise, prediction and estimation are different. In general conditional independent cases, prediction for a new point \( x_{n+1} \) has no meaning.

We define a statistical model and a prior by

\[
p(y|x, w), \quad \varphi(w),
\]

where \( w \in \mathbb{R}^d \) is a parameter. For a given sample \((x^n, Y^n)\), a posterior density is defined by

\[
p(w|x^n, Y^n) = \frac{1}{Z_n} \varphi(w) \prod_{i=1}^{n} p(Y_i|x_i, w),
\]

where \( Z_n \) is a normalizing constant,

\[
Z_n = \int \varphi(w) \prod_{i=1}^{n} p(Y_i|x_i, w)dw.
\]

Also we define a Bayesian estimated conditional density by

\[
p(y|x, x^n, Y^n) = \mathbb{E}_w[p(y|x, w)],
\]

where \( \mathbb{E}_w[\cdot] \) is the expectation value using the posterior density. The generalization loss \( G_n \), the training loss \( T_n \),
the cross validation loss $C_n$, and and WAIC $W_n$ are respectively defined by

$$G_n = - \frac{1}{n} \sum_{i=1}^{n} \int q(y|x_i) \log p(y|x_i, x^n, Y^n) dy,$$

$$T_n = - \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|x_i, x^n),$$

$$C_n = - \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|x_i, x^n \setminus x_i, Y^n \setminus Y_i),$$

$$W_n = T_n + \frac{1}{n} \sum_{i=1}^{n} \mathbb{V}_w[\log p(Y_i|x_i, w)].$$

In conditional independent cases, the generalization loss cannot be estimated by the cross validation loss, in other words,

$$\mathbb{E}[G_{n-1}] \neq \mathbb{E}[C_n].$$

If $x^n$ is dependent, then we can not use the relation between the generalization loss and the cross validation loss. However, the following theorem shows that, even in such cases, the generalization loss can be estimated by using WAIC.

**Theorem 1.** (Main Theorem) Assume that

$$L(w) = - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Y[\log p(Y|x_i, w)],$$

is an analytic function of $w$. Then,

$$\mathbb{E}[G_n] = \mathbb{E}[W_n] + o\left(\frac{1}{n}\right).$$

**Proof.** Let $w_0$ be one of parameters which minimize $L(w)$. The average and empirical log losses are defined by

$$L_0 = - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Y[\log p(Y|x_i, w_0)],$$

$$L_n = - \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|x_i, w_0).$$

The log density ratio function is defined by

$$f(x, y, w) = \log(p(y|x, w_0)/p(y|x, w)).$$

By using the resolution theorem [11], there exists an analytic function [12] from a manifold to the set of parameter $w = g(u)$ which satisfies

$$f(x, y, g(u)) = a(x, y, u) u^k,$$

$$|g'(u)| \varphi(g(u)) = b(u) u^h,$$

on each local coordinate LC of the manifold, where $|g'(u)|$ is the absolute value of the Jacobian determinant and

$$u^k = u^k_1 u^k_2 \cdots u^k_d,$$

$$u^h = u^h_1 u^h_2 \cdots u^h_d.$$
Then
\[
\mathbb{E}[\mathbb{E}^* [\sqrt{\xi} \exp(\sqrt{\xi})]] = \mathbb{E} \left[ \frac{S[\sqrt{\xi} \exp(\sqrt{\xi})]}{S[\exp(\sqrt{\xi})]} \right]
\]
\[
= \sum_{j=1}^{\infty} \mathbb{E} \left[ \frac{\partial}{\partial y_j} \left( \frac{S[\sqrt{\xi} \exp(\sqrt{\xi})]}{S[\exp(\sqrt{\xi})]} \right) \right]
\]
\[
= \sum_{j=1}^{\infty} \mathbb{E} \left[ \frac{S[\sqrt{\xi} \exp(\sqrt{\xi})]}{S[\exp(\sqrt{\xi})]} \right]
\]
\[
- \mathbb{E} \left[ \frac{S[\sqrt{\xi} \exp(\sqrt{\xi})]^2}{S[\exp(\sqrt{\xi})]} \right].
\]
The last equation is equal to \( \mathbb{E}[V^*[\xi]] \), which completes Lemma.

3. AN EXPERIMENT

Let \( x, y, a \in \mathbb{R}^1 \) and \( s > 0 \). A statistical model
\[
p(y|x, a, s) = \sqrt{\frac{s}{2\pi}} \exp \left( -\frac{s}{2} (y - ax)^2 \right)
\]
with a prior
\[
\varphi(a, s) \propto s \exp \left( -\frac{0.01}{2} s(1 + a^2) \right)
\]
were studied in an experiment. The sample size was \( n = 10 \). \( x_1, x_2, \ldots, x_9 \) were set as \( x_i = 1 + i/10 \) and a leverage sample point \( x_{10} \) was chosen from \( \{2, 3, \ldots, 8\} \). \( Y_1, Y_2, \ldots, Y_{10} \) were independently taken from \( p(y|x_i, a_0, s_0) \), where \( a_0 = 0.2 \) and \( s_0 = 100 \), that is to say, the standard deviation of \( Y_i \) was 0.1. Figure 1 shows the expectation values and standard deviations of the generalization error \( G - S \), LOOCV error \( C_n - S_n \) and WAIC error \( W_n - S_n \), where \( S \) and \( S_n \) are average and empirical entropies of samples. The horizontal line of Figure 1 shows the place of the leverage sample point. If a leverage sample point is far from the other sample points, then LOOCV error did not estimate the generalization error.

4. DISCUSSION

In this section, we discuss mathematical difference between the cross validation and information criterion from the viewpoint of statistical estimator of the generalization loss.

The leave-one-out cross validation loss is defined by
\[
\text{cv}(X^n) = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(p(X_i|X^n \setminus X_i)).
\]
Hence, if \( X^n \) are independent, the equation
\[
\mathbb{E}[\text{cv}(X^n)] = \mathbb{E}[C_{n-1}]
\]
holds automatically, however, if independent condition is not ensured, it is not easy to generalize such relation. Moreover, the variance of \( \text{cv}(X^n) \) is also unknown.

On the other hand, an information criterion is defined by
\[
\text{ic}(X^n) = \text{Training Loss} + \text{complexity},
\]
where the complexity term is derived by the asymptotic property of the difference between training and generalization losses. Because such asymptotic property is based on the convergence of the empirical loss to a Gaussian process on the parameter space, it can be generalized onto dependent cases. This difference affects not only theoretical consideration but also practical applications. Information criteria can be used in conditional independent cases, whereas LOOCV not.

5. CONCLUSION

We proved that, if a sample consists of conditional independent random variables, then WAIC is the better estimator of the generalization loss than the leave-one-out cross validation.

Acknowledgment. This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 25120013 and 15K00331.

6. REFERENCES


ABSTRACT

In the conventional model selection framework it has been discussed how to estimate the best dimensionality of probabilistic models where the notion of dimensionality is predetermined. This paper rather introduces a new notion of dimensionality of probabilistic models from the view of descriptional complexity, which we call descriptional dimensionality (D-dim). D-dim is not necessarily integer-valued. It is a suitable measure to analyze what happens during a model transition period when the probabilistic model may change over time. I give a theoretical basis of D-dim, referring to its relation to minimum description length (MDL) principle. Further I relate D-dim to error probability exponents for model change detection. I also show how to apply D-dim to gradual model change detection.
ASYMPTOTIC DENSITY IN LOGIC AND COMPUTABILITY

Marek Zaionc
Jagiellonian University, Poland

ABSTRACT

This talk presents numerous results from the area of quantitative investigations in logic and computability. We present a quantitative analysis of random formulas or random lambda terms or random combinatory logic terms. Our main goal is to investigate likelihood of semantic properties of random computational objects.

For the given logical calculus (or type theory or combinatory logic term) we investigate the proportion of the number of distinguished formulas (or types or terms) of a certain length $n$ to the number of all formulas of such length. We are especially interested in asymptotic behavior of this fraction when $n$ tends to infinity.

For the given set $A$ of objects the limit $\mu(A)$ if exists, is an asymptotic probability of finding formula from the class $A$ among all formulas or may also be interpreted as the asymptotic density of the set $A$. Results obtained are having some philosophical flavor like for example:

- How big is the fraction of one logic being sub-logic of the bigger one based on the same language,
- Estimate chances that random formula is true or how big is the fraction of tautologies?
- What are chances that random program terminates?