Seventh Joint Workshop on Coding and Communications

booklet of abstracts

Seventh Joint Workshop on Coding and Communications (JWCC) 2014

November 13–15, 2014

Hotel Casa Fuster, Barcelona, Spain

Booklet of Abstracts

Table of Contents

Session 1 Chair: T. Weissman Fri 09:30 – 11:00

Session 2 Chair: A. Lapidoth Fri 11:30 – 13:00

Session 3 Chair: H. Boche Fri 15:00 – 16:30

Session 4 Chair: G. D. Forney, Jr Fri 17:00 – 18:00

Session 5 Chair: J. Huber Sat 10:00 – 11:00

Session 6 Chair: A. Goldsmith Sat 11:30 – 13:00

Session 7 Chair: G. Kramer Sat 15:00 – 16:30

Session 8 Chair: P. Narayan Sat 17:00 – 18:00

On the Information Theory of Caching Networks

Giuseppe Caire

Abstract

Caching is a well-known general principle for which, if some information message is likely to be requested by a user (or network node) in the future, this can be pre-stored in the node itself or in its “vicinity” (in some topological sense) at a favorable time, such that when the request comes, it can be satisfied with low latency and/or without causing congestion in the network. Building on the fact that media content requests are highly predictable, caching has allowed the implementation of content distribution networks (CDNs), which are at the basis of vastly popular video streaming services over the Internet, such as Netflix, iTune and Amazon Instant Video. More recently (in the past 2-3 years), caching has attracted also a significant attention in information theory. In this talk, we shall review in a tutorial fashion some popular “basic” information theoretic models for caching networks and the corresponding known results on their fundamental limits. Interestingly, the achievability results are constructive and have an algebraic network coding flavor, and come at a fixed multiplicative penalty factor (independent of the size of the network and size of the message library) from information theoretic outer-bounds.

Aspects of Random Linear Network Coding in Layered Networks

Michael Cyran[1], Birgit Schotsch[2], Johannes B. Huber[1], Robert F.H. Fischer[3]
[1]Lehrstuhl für Informationsübertragung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Institute of Communication Systems and Data Processing, RWTH Aachen University, Aachen, Germany
[3]Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany
cyran@LNT.de, schotsch@ind.rwth-aachen.de, huber@LNT.de, robert.fischer@uni-ulm.de

Abstract

Random linear network coding (RLNC) [1]–[5] is a method to maximize the information flow in a communication network by forming random linear combinations over some finite field $\mathbb{F}_q$ of the received information packets at each intermediate node. The network between one source node and one destination node acts as a linear map $\mathbb{F}_q^n \to \mathbb{F}_q^N$, which is represented by the *network channel matrix*. Since the linear factors, i.e., the coding coefficients at each intermediate node are drawn independently at random, there is no need of a central processor or of sharing side information between the nodes.

Currently, in RLNC there coexist different approaches for generating the linear combinations at intermediate nodes. Therefore, we classify and characterize the existing essentially two distinct variants and show their equivalence. In variant 1 each intermediate node calculates *one* linear combination and transmits it on its outgoing edges whereas in variant 2 the intermediate nodes compute *individual* linear combinations for each of their outgoing edges. Other variants can be seen as hybrids of these two variants. We show that each network which makes use of variant 2 (or of any hybrid variant) can be transformed into an equivalent network which applies variant 1, by splitting up intermediate nodes which transmit different messages on their outgoing edges into several single output edges.

Besides this classification into two encoding variants at node level, we introduce further structure in terms of layers into seemingly disparate and unstructured network topologies. By inserting redundant intermediate single input/single output nodes, arbitrary acyclic networks can be transformed into layered networks. Such layered networks constitute a special class of networks, where the intermediate nodes are arranged in L layers. Nodes in layer l only receive packets from nodes in layer $l-1$, i.e., there are no direct connections between non-adjacent layers. The three major advantages of the transformation of an arbitrary acyclic network into a layered network, which we denote *layering*, are:

I. The network channel matrix can be factorized into so called inter-layer matrices.
II. An inherent synchronization is provided, i.e., all paths that connect the source node and the destination node are equally long, i.e., each path has length L.
III. Layering simplifies or even enables an accurate analysis of RLNC systems.

We present two RLNC setups, which can be thoroughly analyzed because of their layered structure. At first, we examine the effects of joining or leaving nodes on the network channel matrix. This allows to derive a statistical network channel model for *slowly varying* networks which considers additive packet errors as well as changes in the network topology due to leaving or joining nodes [6]. In the second setup we derive an upper bound on the outage probability of two-layer network channel matrices, i.e., the probability that the network channel matrix does not have full column rank [7], [8]. This upper bound is particularly important for networks whose network channel matrix is sparse, i.e., in cases where the well known results for dense matrices do not apply. We numerically evaluate the proposed bound, compare it with corresponding Monte Carlo simulations and thereby show that this upper bound almost coincides with the simulations. Finally, we discuss the generalization of the bound to multi-layer networks.

References

[1] R. Ahlswede, N. Cai, S.-Y. Li, R.W. Yeung, “Network Information Flow,” *IEEE Transactions on Information Theory*, Vol. 46, No. 4, pp. 1204–1216, July 2000.
[2] S.-Y. Li, R. Yeung, N. Cai, “Linear Network Coding,” *IEEE Transactions on Information Theory*, Vol. 49, No. 2, pp. 371–381, February 2003.
[3] R. Koetter, M. Médard, “An Algebraic Approach to Network Coding,” *IEEE/ACM Transactions on Networking*, Vol. 11, No. 5, pp. 782–795, October 2003.
[4] T. Ho, M. Médard, R. Koetter, D. Karger, M. Effros, J. Shi, B. Leong. “A Random Linear Network Coding Approach to Multicast”. *IEEE Transactions on Information Theory*, Vol. 52, No. 10, pp. 4413–4430, October 2006.
[5] D. Silva, F.R. Kschischang, R. Koetter. “Communication Over Finite-Field Matrix Channels”. *IEEE Transactions on Information Theory*, Vol. 56, No. 3, pp. 1296–1305, March 2010.
[6] S. Puchinger, M. Cyran, R.F.H. Fischer, M. Bossert, J.B. Huber, “Error Correction for Differential Linear Network Coding in Slowly-Varying Networks,” *accepted for presentation at ITG Conference on Systems, Communication and Coding (SCC) 2015*.
[7] B. Schotsch, M. Cyran, J.B. Huber, R.F.H. Fischer, P. Vary, “An Upper Bound on the Outage Probability of Random Linear Network Codes with Known Incidence Matrices,” *accepted for presentation at ITG Conference on Systems, Communication and Coding (SCC) 2015*.
[8] B.E. Schotsch, “Rateless Coding in the Finite Length Regime,” Ph.D. dissertation, IND, RWTH Aachen University, Aachen, Germany, 2014.

Interactive Multiterminal Communication

Prakash Narayan
Dept. of Electrical and Computer Engineering
and Institute for Systems Research
University of Maryland
College Park, MD 20742, USA
Email: prakash@umd.edu

***Abstract*—Information theoretic models for multiuser source and channel coding usually take the communication among multiple terminals to be “simple" or autonomous. On the other hand, studies of multiparty function computation, especially in computer science, emphasize the useful role of interactive communication. We shall describe basic structural properties of interactive communication. “Single-shot" bounds will be presented for the amount of common randomness, i.e., shared information, that can be generated among the terminals using such communication. A few simple consequences with applications will be discussed. This talk is based on joint works with Imre Csiszár, Sirin Nitinawarat, Himanshu Tyagi and Shun Watanabe.**

Acknowledgment

This work was supported by the U.S. National Science Foundation under Grants CCF0830697 and CCF1117546.

{Detection and Estimation} and Information Theory

Sergio Verdú
Princeton University
Princeton, NJ 08544, USA
verdu@princeton.edu

***Abstract*—Teaching a course on *Detection and Estimation* from the perspective of information theory is a fun thing to do. This abstract collects some of its nuggets.**[1]

I. Notation

- In the statistics and signal processing literatures it is common to use X to denote the observable, which is the input to the statistical inference algorithm. Instead, we abide by the widespread usage in the information theory literature where noisy observations are denoted by Y.
- Relative information:
$$\imath_{P\|Q}(x) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \tag{1}$$
The cdf of the relative information is denoted by
$$\mathbb{F}_{X\|Y}(\alpha) = \mathbb{P}\left[\imath_{X\|Y}(X) \leq \alpha\right] \tag{2}$$
- Information density:
$$\imath_{V;Y}(\theta; a) = \log \frac{\mathrm{d}P_{Y|V=\theta}}{\mathrm{d}P_Y}(a) = \imath_{P_{Y|V=\theta}\|P_Y}(a) \tag{3}$$
- $h(\cdot)$ is the binary entropy; $d(\cdot\|\cdot)$ is the binary relative entropy.
- Markov chain: $V -\square- Z -\square- Y$ means that V and Y are conditionally independent given Z.
- $P_Y \to P_{Z|Y} \to P_Z$ means that the second marginal of the joint distribution $P_Y P_{Z|Y}$ is denoted by P_Z.

II. Sufficient statistics

Given sets $(\mathcal{Y}, \mathcal{Z}, \Theta)$ (we omit reference to the corresponding σ-fields for brevity) and

- $\{P_{Y|V=\theta}, \theta \in \Theta\}$: a collection of distributions on $\mathcal{Y}$,
- a random transformation $P_{Z|Y} \colon \mathcal{Y} \to \mathcal{Z}$,

Z is a sufficient statistic of Y for V if $P_{Y|Z,V=\theta}$ does not depend on θ, where $P_{YZ|V=\theta} = P_{Y|V=\theta} P_{Z|Y}$.

The classical notion of sufficient statistic was introduced by Ronald Fisher [8] as a deterministic function of the data such that "no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated." Therefore, we would expect that information theory would have something to say about it. For example, Cover and Thomas [5] give the necessary and condition for sufficient statistic as
$$I(V;Y) = I(V;Z) \tag{4}$$
which is equivalent to $V -\square- Z -\square- Y$. Unfortunately, the scope of this is limited to the Bayesian setting where a distribution P_V on Θ is specified. Even in the non-Bayesian setting, information theory has something to say. To that end, we define information densities with an auxiliary distribution P_Y that dominates the collection $\{P_{Y|V=\theta}, \theta \in \Theta\}$ (and does not depend on θ), and we denote $P_Y \to P_{Z|Y} \to P_Z$. Then, Z is a sufficient statistic of Y for V if
$$\imath_{V;Y}(\theta; y) - \imath_{V;Z}(\theta; z) \text{ is invariant to } \theta. \tag{5}$$
except possibly for (y, z) in an event of zero measure $P_{YZ|V=\theta}$ for all θ. A corollary (useful for the conventional setting of deterministic sufficient statistics) is that if $\imath_{V;Y}(\theta; y)$ depends on y only through $f(y)$, then $Z = f(Y)$ is a sufficient statistic of Y for V, a statement which is equivalent to the factorization theorem of Halmos and Savage [12]. The case $|\Theta| = 2$ deserves particular attention.

Theorem 1. *Assume $P_Y \ll Q_Y$. Z is a sufficient statistic of Y for $\{P_Y, Q_Y\}$ if and only if*
$$\imath_{P_Y\|Q_Y}(Y) = \imath_{P_Z\|Q_Z}(Z) \tag{6}$$
where $P_Y \to P_{Z|Y} \to P_Z$, $Q_Y \to P_{Z|Y} \to Q_Z$ and (6) *holds with probability one according to $P_Y P_{Z|Y}$ and $Q_Y P_{Z|Y}$.*

Furthermore, if $D(P_Z \,\|\, Q_Z) < \infty$, then
$$D(P_Y \,\|\, Q_Y) = D(P_Z \,\|\, Q_Z) \tag{7}$$
is necessary and sufficient for Z to be a sufficient statistic.

III. Binary Hypothesis Testing

In some applications (including those that originally motivated the study of hypothesis testing in statistics), the hypotheses play asymmetrical roles, with one of them representing the truth of a postulated scientific theory, the presence of an illness, anomaly, target, etc. This asymmetry led to the adoption of the following terminology:

H_0:	null hypothesis.
H_1:	alternative hypothesis.
$\pi_{0\|1}$:	type II probability; probability of missed detection; $1 - \pi_{0\|1}$ is also referred to as the probability of detection, power, or sensitivity of the test.
$\pi_{1\|0}$:	type I probability; false-alarm probability; size; specificity. The maximum allowable pre-specified value of $\pi_{1\|0}$ is called the significance level.

For a communications/information theory audience it is ill-advised to stick to that classical terminology since such

[1] Supported by the NSF Science and Technology Center for Science of Information under Grant CCF-0939370

asymmetry between the nature of the hypotheses is alien to most of the applications that the student of the course will encounter.

The fundamental tradeoff is characterized by the set of achievable error probability pairs:

$$\mathcal{C}(P_0, P_1) = \bigcup_{\phi\colon \mathcal{Y}\to\{0,1\}} \{(\mathbb{E}[\phi(Y_0)], 1 - \mathbb{E}[\phi(Y_1)])\} \tag{8}$$

with $Y_0 \sim P_0$ and $Y_1 \sim P_1$. The lower boundary of (8) is

$$\begin{aligned} \alpha_\nu(P_1, P_0) &= \min\{y \in [0,1]\colon (\nu, y) \in \mathcal{C}(P_0, P_1)\} & (9)\\ &= \min_{\phi\colon \pi_{1|0} \le \nu} \pi_{0|1} & (10)\end{aligned}$$

The role of relative entropy and Rényi divergence in asymptotic results such as the Chernoff(-Stein) Lemma is well-known in the information theory community, particularly since the work by Blahut [3]. I find that giving the proof of Sanov's large deviations theorem in the finite-alphabet setting using the method of types is a nice illustration of the power of that method. But in keeping with our philosophy most of the real work is done providing bounds for the non-asymptotic single shot version.

Converse results include (for simplicity we assume henceforth that $P_0 \ll\gg P_1$):

Theorem 2. *The error probabilities* $(\pi_{1|0}, \pi_{0|1})$ *of any test must satisfy*

$$\begin{aligned} d(\pi_{0|1} \| 1 - \pi_{1|0}) &\le D(P_1 \| P_0) & (11)\\ d(\pi_{1|0} \| 1 - \pi_{0|1}) &\le D(P_0 \| P_1) & (12)\end{aligned}$$

Theorem 3. *The error probabilities* $(\pi_{1|0}, \pi_{0|1})$ *of any test must satisfy for all* $\tau \in \mathbb{R}$*:*

$$\pi_{0|1} + \exp(\tau)\pi_{1|0} \ge \max\{\mathbb{F}_{P_1\|P_0}(\tau), \exp(\tau)\mathbb{F}_{P_0\|P_1}(-\tau)\}$$

The following result is inspired by an idea due to Shannon, Gallager and Berlekamp [22].

Theorem 4. *Suppose that the positive scalars* $(\theta_0, \theta_1, \tau_0, \tau_1)$ *are such that*

$$(\theta_0 \exp(-\tau_0), \theta_1 \exp(-\tau_1)) \in \mathcal{C}(P_0, P_1). \tag{13}$$

Then, they must satisfy

$$\theta_0 + \theta_1 \ge \mathbb{P}\left[\imath_{P_\alpha\|P_0}(Y_\alpha) \le \tau_0, \imath_{P_\alpha\|P_1}(Y_\alpha) \le \tau_1\right] \tag{14}$$

where $Y_\alpha \sim P_\alpha$*, for all* $\alpha \in [0,1]$*, and the tilted distribution is defined through*

$$\begin{aligned} \imath_{P_\lambda\|R}(a) = \lambda\, \imath_{P_1\|R}(a) &+ (1-\lambda)\imath_{P_0\|R}(a)\\ &+ (1-\lambda) D_\lambda(P_1 \,\|\, P_0). & (15)\end{aligned}$$

where $D_\lambda(P_1 \,\|\, P_0)$ *is the Rényi divergence.*

Achievability results include:

Theorem 5. *For all* $\nu \in (0,1)$*,*

1)

$$\alpha_\nu(P_1, P_0) \le \mathbb{F}_{P_1\|P_0}\left(\log\frac{1}{\nu}\right) \tag{16}$$

2) *If* τ *is such that* $\mathbb{F}_{P_0\|P_1}(\tau) \le \nu$*, then*

$$\alpha_\nu(P_1, P_0) \le \exp(-\tau)\left(1 - \mathbb{F}_{P_0\|P_1}(\tau)\right) \tag{17}$$

Theorem 6. *Let* $P_1 \ll P_0$*. For all* $\alpha \in (0,1)$ *there exists a deterministic test such that*

$$\begin{aligned} \pi_{1|0} &\le \exp\left(-D(P_\alpha \| P_0)\right) & (18)\\ \pi_{0|1} &\le \exp\left(-D(P_\alpha \| P_1)\right) & (19)\end{aligned}$$

IV. M-ary Hypothesis Testing

While in the statistical inference literature, testing among more than two hypotheses receives scant attention (except as a bounding technique in estimation), it has fundamental importance in communications and information theory, since the role of a receiver in a communication system, particularly one that uses error correcting codes, is to guess which message was transmitted out of a finite number of alternatives.

Suppose that we are given the random transformation $P_{Y|V}\colon \mathcal{M} = \{1, \ldots, M\} \to \mathcal{Y}$ and the task is to guess V based on the observation $y \in \mathcal{Y}$. This is an M-ary hypothesis testing problem where the observation is generated under one of M distributions:

$$\begin{aligned} \mathsf{H}_1\colon&\ y \sim P_1 = P_{Y|V=1}\\ &\ \vdots\\ \mathsf{H}_M\colon&\ y \sim P_M = P_{Y|V=M}\end{aligned}$$

For brevity, here I only consider the Bayesian case in which there is a prior distribution P_V. Optimizing over the test, the minimal error probability is a function of $(P_1, \ldots, P_M)$ and the prior distribution P_V, which we denote, as $\varepsilon_{V|Y}$. We give upper and lower bounds in terms of information measures. Tebbe and Dwyer [23] and Gallager [10, p. 521] gave the upper bound

$$\varepsilon_{V|Y} \le \frac{1}{2} H(V|Y) \text{ bits} \tag{20}$$

which can be further tightened by the following result.

Theorem 7. *Let* k *be a positive integer. If* $\log k \le H(V|Y) \le \log(k+1)$*, then*

$$\varepsilon_{V|Y} \le \frac{H(V|Y) + (k^2-1)\log(k+1) - k^2 \log k}{k(k+1)\log\frac{k+1}{k}} \tag{21}$$

The idea behind the following upper bound is prominent in the analysis of the fundamental limits of data transmission, and in particular Shannon's original channel coding achievability approach.

Theorem 8.

$$\varepsilon_{V|Y} \le \inf_{\gamma > 0} \left\{\mathbb{P}[\imath_{V;Y}(V;Y) \le \imath_V(V) + \gamma] + \exp(-\gamma)\right\}. \tag{22}$$

with the information $\imath_V(a) = \log\frac{1}{P_V(a)}$.

As far as lower bounds are concerned, Fano's inequality yields

Theorem 9. *If V takes M possible values, then*

$$\varepsilon_{V|Y} \geq \varphi_M^{-1}(H(V|Y)) \tag{23}$$

with φ_M^{-1} denoting the inverse function of $\varphi_M \colon [0, 1 - \frac{1}{M}] \to [0, \log M]$:

$$\varphi_M(t) = t \log(M-1) + h(t). \tag{24}$$

from which we can get as a corollary that

$$\varepsilon_{V|Y} \geq \frac{\log \frac{M}{2}}{\log(M-1)} - \frac{1}{M^2 \log(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} D\left(P_i \| P_j\right),$$

which is reminiscent of Birgé's lower bound [1] on the minimax error probability:

$$d\left(\bar{\varepsilon} \,\Big\|\, 1 - \frac{\bar{\varepsilon}}{M-1}\right) \leq \frac{1}{M-1} \min_j \sum_{i \neq j} D(P_i \| P_j) \tag{25}$$

Another bound based on $H(V|Y)$ is due to none other than Shannon [21].

Theorem 10.

$$\varepsilon_{V|Y} \geq \frac{1}{6} \frac{H(V|Y)}{\log M + \log\log M - \log H(V|Y)} \tag{26}$$

The following is due to Poor and Verdú [19]

Theorem 11.

$$\begin{aligned}\varepsilon_{V|Y} &\geq \sup_{0 \leq \alpha \leq 1} (1-\alpha)\, \mathbb{P}\left[P_{V|Y}(V|Y) \leq \alpha\right] & (27)\\ &= \sup_{\gamma > 0} (1 - \exp(-\gamma))\, \mathbb{P}\left[\imath_{V;Y}(V;Y) \leq \imath_V(V) - \gamma\right] & (28)\end{aligned}$$

Weakening Theorem 11 we arrive at the following pleasing companion to Theorem 8:

Theorem 12.

$$\varepsilon_{V|Y} \geq \sup_{\gamma > 0} \{\mathbb{P}\left[\imath_{V;Y}(V;Y) \leq \imath_V(V) - \gamma\right] - \exp(-\gamma)\} \tag{29}$$

The following lemma (which modulo conceptually minor variations is the "meta-converse" due to Polyanskiy *et al.* [18, Theorem 26]) offers a way to lower bound the error probability of an M-ary hypothesis testing problem by means of the analysis of an auxiliary binary hypothesis testing problem.

Lemma 1. *Fix*

- *P_V on $\mathcal{M}$ and $P_{Y|V} \colon \mathcal{M} \to \mathcal{Y}$,*
- *Q_V on $\mathcal{M}$ and $Q_{Y|V} \colon \mathcal{M} \to \mathcal{Y}$,*
- *test $P_{\widehat{V}|Y} \colon \mathcal{Y} \to \mathcal{M}$.*

Denote the average error probability attained by the M-ary test under $P_{Y|V}P_V$ (resp. $Q_{Y|V}Q_V$) by ϵ (resp. ϵ'). Then,

$$\epsilon \geq \alpha_{1-\epsilon'}(P_{VY}, Q_{VY}) \tag{30}$$

Originating in Barcelona [26], the next result shows that, in fact, Lemma 1 is tight by appropriate choice of the auxiliary distribution.

Theorem 13. *Fix $P_{Y|V} \colon \mathcal{M} \to \mathcal{Y}$, and P_V on $\mathcal{M}$ such that $P_V(m) > 0$ for all $m \in \mathcal{M}$. Then, the minimal error probability is equal to*

$$\varepsilon_{V|Y} = \alpha_{\frac{1}{M}}\left(P_V P_{Y|V}, P_U \times P_{\overline{Y}}\right) \tag{31}$$

where

- *P_U is equiprobable on $\mathcal{M}$;*
- *$P_{\overline{Y}}$ is defined through*

  $$\imath_{\overline{Y}\|Y}(y) = \log \kappa - \min_{m \in \mathcal{M}} \imath_{V|Y}(m|y), \tag{32}$$

  with $P_V \to P_{Y|V} \to P_Y$ and $\kappa > 1$ chosen so that $P_{\overline{Y}}$ is a probability measure.

One of the uses of Theorem 13 is to sharpen Theorem 12 to actually yield an exact result.

Theorem 14. *Denote $\imath_{V;\widehat{Y}}(v;y) = \log \frac{\mathrm{d}P_{Y|V=v}}{\mathrm{d}P_{\widehat{Y}}}(y)$. Then,*

$$\varepsilon_{V|Y} = \max_{\gamma > 0}\left\{\max_{P_{\widehat{Y}}} \mathbb{P}\left[\imath_{V;\widehat{Y}}(V;Y) \leq \imath_V(V) - \gamma\right] - \exp(-\gamma)\right\} \tag{33}$$

V. Minimum Mean-Square Error Estimation

A. Non-Bayesian

In the setting of non-Bayesian minimum-variance unbiased estimation where we have

- Unknown real-valued parameter: $\theta \in \Theta \subset \mathbb{R}$
- Set of observations: $\mathcal{Y}$
- Model for the data: $P_{Y|V=\theta} \colon \Theta \to \mathcal{Y}$
- Loss function: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

it is advisable to bring up the following elegant result due to Blackwell [2], Kolmogorov [14], and Rao [20].

Theorem 15. *Given $P_{Y|V} \colon \Theta \to \mathcal{Y}$, suppose that*

- *$P_{\widehat{V}|Y} \colon \mathcal{Y} \to \Theta$ is an unbiased estimator of V;*
- *$P_{Z|Y} \colon \mathcal{Y} \to \mathcal{Z}$ is a sufficient statistic of Y for V.*

Then,

1) *the estimator $P_{\overline{V}|Y}$ defined by*

   $$\overline{V} = \mathbb{E}[\widehat{V}|Z, V = \theta] \tag{34}$$

   is an unbiased estimator of V;
2) *for all $\theta \in \Theta$*

   $$\mathbb{E}[(V - \overline{V})^2|V = \theta] \leq \mathbb{E}[(V - \widehat{V})^2|V = \theta] \tag{35}$$

(In the literature, it is common to find $\overline{V} = \mathbb{E}[\widehat{V}|Z]$ in lieu of (34) but that quantity is not defined in the non-Bayesian setting. Note that (34) is indeed a function of the data only since Z is a sufficient statistic.) In addition, it is convenient to introduce the notion of complete sufficient statistic and give the Lehmann-Scheffé theorem [16], [17] which guarantees that any function of a complete sufficient statistic yielding an unbiased estimator is a uniform minimum-variance unbiased estimator.

The Crámer-Rao bound, also known as the *information inequality*, follows from the Cauchy-Schwartz inequality. I do not know of an information theoretic proof, but in a departure from convention, I prefer to give an information theoretic definition of Fisher's information:

Definition 1. *Given $P_{Y|X} \colon \mathbb{R} \to \mathcal{Y}$, and $\theta \in \mathbb{R}$, suppose that*

$$\lim_{\alpha \to \theta} \frac{1}{\alpha - \theta} D\left(P_{Y|X=\alpha} \| P_{Y|X=\theta}\right) = 0 \tag{36}$$

Then, the Fisher information of $P_{Y|X}$ at θ is

$$\begin{aligned} J(\theta, P_{Y|X}) &= \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} D\left(P_{Y|X=\alpha} \| P_{Y|X=\theta}\right) |_{\alpha \leftarrow \theta} && (37) \\ &= \lim_{\alpha \to \theta} \frac{2}{(\alpha - \theta)^2} D\left(P_{Y|X=\alpha} \| P_{Y|X=\theta}\right) && (38) \end{aligned}$$

where the relative entropy is in nats. When the parameter θ has a probability distribution, we define the conditional Fisher information as

$$J(Y|X) = \mathbb{E}\left[J(X, P_{Y|X})\right]. \tag{39}$$

As is well-known since [15], under appropriate technical sufficient conditions that ensure the validity of swapping of expectation-differentiation and Taylor series expansion, we can express (37) in its conventional, less insightful, form:

$$J(\theta, P_{Y|X}) = -\mathbb{E}\left[\frac{\partial^2}{\partial \alpha^2} \imath_{P_{Y|X=\alpha} \| P_{Y|X=\theta}}(Y_\theta)\right] |_{\alpha \leftarrow \theta} \tag{40}$$

where $Y_\theta \sim P_{Y|X=\theta}$ and the relative information is in nats.

The original concept of Fisher information [9] of a density function f_Z, simply corresponds to the special case of Definition 1 with $P_{Y|X=\theta} = P_{Z+\theta}$, yielding

$$J(Z) = -\mathbb{E}\left[\nabla^2 \log_e f_Z(Z)\right] \tag{41}$$

which is not invariant to reversible transformations.

B. Bayesian

The estimation topic that benefits the electrical engineering graduate student the most is minimum mean-square error estimation in the Bayesian setting, in which we are given

- a priori distribution P_X on $\mathbb{R}$ for the unknown real-valued random variable to be estimated, known as the *estimand*;
- a model for the data $P_{Y|X} \colon \mathbb{R} \to \mathcal{Y}$.

The *minimum mean-square error* for estimating X given Y is

$$\mathsf{mmse}(X|Y) = \min \mathbb{E}[(X - \widehat{X})^2] \tag{42}$$

where the minimum is over all Borel-measurable random transformations $P_{\widehat{X}|Y} \colon \mathcal{Y} \to \mathbb{R}$, and $X - \square - Y - \square - \widehat{X}$, and is achieved by the conditional mean $\mathbb{E}[X|Y]$.

Theorem 16. .

1) *If $f \colon \mathbb{R} \to \mathbb{R}$ is an injective mapping, then*

$$\mathsf{mmse}(X|f(X)) = 0. \tag{43}$$

2) *If $\mathbb{E}[X^2] < \infty$, then*

$$\mathsf{mmse}(X|Y) \le \sigma_X^2, \tag{44}$$

achieved with equality if X and Y are independent.

3) *If $\mathbb{E}[X^2] < \infty$, then*

$$\mathsf{mmse}(X|Y) = \mathbb{E}[X^2] - \mathbb{E}[\mathbb{E}^2[X|Y]]. \tag{45}$$

4)

$$\mathbb{E}[X|Y = y] = \mathbb{E}[X \exp(\imath_{X;Y}(X; y))] \tag{46}$$

5) *Suppose $X - \square - Y - \square - Z$. Then,*

$$\mathsf{mmse}(X|Y) \le \mathsf{mmse}(X|Z) \tag{47}$$

with equality if Z is a sufficient statistic of Y for X.

6) *[24] Assume that X is a continuous random variable with density function f_X, whose Fisher information is $J(X)$ and whose mean is finite. Then,*

$$\mathsf{mmse}(X|Y) \ge \frac{1}{J(X) + J(Y|X)}. \tag{48}$$

7) *[29] $\mathsf{mmse}(X|Y)$ is a concave functional of P_{XY}.*

Of great special interest is the model

$$Y = \sqrt{\gamma} X + N \tag{49}$$

where $N \sim \mathcal{N}(0, 1)$, independent of X. Then, we abbreviate the notation as

$$\mathsf{mmse}(X, \gamma) = \mathsf{mmse}(X|\sqrt{\gamma} X + N) \tag{50}$$

Beyond those in Theorem 16, $\mathsf{mmse}(X, \gamma)$ satisfies the following properties.

Theorem 17. .

1)

$$\mathsf{mmse}(aX + b, \gamma) = a^2\, \mathsf{mmse}(X, a^2 \gamma) \tag{51}$$

2)

$$\mathsf{mmse}(X, \gamma) \le \frac{\sigma_X^2}{1 + \sigma_X^2 \gamma} \tag{52}$$

achieved with equality if $X \sim \mathcal{N}(\mu, \sigma_X^2)$.

3) *Even if $\mathbb{E}[X]$ does not exist,*

$$\mathsf{mmse}(X, \gamma) \le \frac{1}{\gamma} \tag{53}$$

4) *[7] The conditional mean estimator is given by*

$$\mathbb{E}[X|\sqrt{\gamma}\, X + N = y] = -\frac{1}{\sqrt{\gamma}} \nabla \imath_{X;Y}(0; y) \tag{54}$$

5) *[13] The conditional mean-squared error is given by*

$$\mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\, |Y = y\right] = -\frac{1}{\gamma} \nabla^2 \imath_{X;Y}(0; y) \tag{55}$$

6) *[4]*

$$\gamma\, \mathsf{mmse}(X, \gamma) = 1 - J(\sqrt{\gamma} X + N) \tag{56}$$

7) *If X has a density function, then*

$$\mathsf{mmse}(X, \gamma) \ge \frac{1}{J(X) + \gamma} \tag{57}$$

8) *[27][2] If X_1 and X_2 are independent and $\alpha \in [0, 2\pi]$,*

$$\begin{aligned}&\mathsf{mmse}((\cos\alpha)X_1 + (\sin\alpha)X_2, \gamma) && (58)\\ &\geq (\cos^2\alpha)\,\mathsf{mmse}(X_1, \gamma) + (\sin^2\alpha)\,\mathsf{mmse}(X_2, \gamma)\end{aligned}$$

9) *If $N_1, N_2 \sim \mathcal{N}(0,1)$ and (N_1, N_2, X) are independent, then*

$$\begin{aligned}&\mathsf{mmse}\left(X|\sqrt{\gamma_1}\,X + N_1, \sqrt{\gamma_2}\,X + N_2\right)\\ &= \mathsf{mmse}\left(X, \sqrt{\gamma_1 + \gamma_2}\right) && (59)\end{aligned}$$

10) *[29]* $\mathsf{mmse}(X, \gamma)$ *is strictly concave in P_X.*

11) *[25] Let $X_1, \ldots, X_n$ be independent and let $(\lambda_1, \ldots, \lambda_n)$ be a probability distribution. Then,*

$$\mathsf{mmse}\left(\sum_{i=1}^{n} X_i, \gamma\right) \geq \sum_{i=1}^{n} \lambda_i\ \mathsf{mmse}\left(\frac{\overline{X}_{\setminus i}}{\sqrt{(n-1)\lambda_i}}, \gamma\right)$$

where $\overline{X}_{\setminus i} = -X_i + \sum_{j=1}^{n} X_j$.

12) *[28] The bound in* (53) *is tight asymptotically if X is absolutely continuous:*

$$\lim_{\gamma\to\infty} \gamma\,\mathsf{mmse}(X, \gamma) = 1 \qquad (60)$$

13) *[28] If X is discrete, then the MMSE dimension satisfies*

$$\lim_{\gamma\to\infty} \gamma\,\mathsf{mmse}(X, \gamma) = 0 \qquad (61)$$

14) *[11], [29] Let X be a real-valued random variable independent of $N \sim \mathcal{N}(0,1)$. Then, for any* $\mathsf{snr} \geq 0$, *in nats,*

$$I(X; \sqrt{\mathsf{snr}}X + N) = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\,\mathrm{d}\gamma \qquad (62)$$

15) *[11] Let X be a discrete random variable taking values on $\mathcal{X}$. Then, in nats,*

$$H(X) = \frac{1}{2}\int_0^{\infty} \mathsf{mmse}(g(X), \gamma)\,\mathrm{d}\gamma \qquad (63)$$

for any injective function $g\colon \mathcal{X} \to \mathbb{R}$.

In addition, it is useful to consider the vector, discrete-time and continuous-time versions of (49) and give the non-causal Wiener filter and the Kalman filters as examples.

Finally, we mention the universal relationship between the causal and noncausal continuous-time MMSE achieved by arbitrary input processes [11] for which only an information theoretic proof is known. The full proof is beyond the scope of the course as the verification of Duncan's formula [6] requires stochastic calculus.

[2] This property is the gateway to the simplest proof of Shannon's entropy-power inequality.

References

[1] L. Birgé, "A new lower bound for multiple hypothesis testing," *IEEE Trans. Information Theory*, vol. 51, no. 4, pp. 1611–1615, Apr. 2005.

[2] D. Blackwell, "Conditional expectation and unbiased sequential estimation," *Annals of Math. Statistics*, vol. 18, no. 1, pp. 105–110, 1947.

[3] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Information Theory*, vol. 20, pp. 405–417, July 1974.

[4] L. D. Brown, "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *The Annals of Mathematical Statistics*, vol. 42, no. 3, pp. 855–903, 1971.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[6] T. E. Duncan, "On the calculation of mutual information," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 215–220, 1970.

[7] R. Esposito, "On a relation between detection and estimation in decision theory," *Information and Control*, vol. 12, no. 2, pp. 116–120, 1968.

[8] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Trans. of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922.

[9] ——, "Theory of statistical estimation," *Mathematical Proceedings of the Cambridge Mathematical Society*, vol. 22, pp. 700–725, 1925.

[10] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[11] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. on Information Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[12] P. R. Halmos and L. J. Savage, "Application of the Radon-Nikodym theorem to the theory of sufficient statistics," *The Annals of Mathematical Statistics*, vol. 20, number 2, pp. 225–241, 1949.

[13] C. Hatsell and L. Nolte, "Some geometric properties of the likelihood ratio," *IEEE Trans. Information Theory*, vol. 17, pp. 616–618, 1971.

[14] A. N. Kolmogorov, "Unbiased estimates," *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, vol. 14, no. 4, pp. 303–326, 1950.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[16] E. L. Lehmann and H. Scheffé, "Completeness, similar regions, and unbiased estimation: Part I," *Sankhyā: the Indian Journal of Statistics*, pp. 305–340, 1950.

[17] ——, "Completeness, similar regions, and unbiased estimation: Part II," *Sankhyā: The Indian Journal of Statistics*, pp. 219–236, 1955.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[19] H. V. Poor and S. Verdú, "A lower bound on the probability of error in multihypothesis testing," *IEEE Trans. on Information Theory*, vol. 41, no. 6, pp. 1992–1994, 1995.

[20] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, no. 3, pp. 81–91, 1945.

[21] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Research and Development*, vol. 2, no. 4, pp. 289–293, Oct. 1958.

[22] C. E. Shannon, R. G. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels, I," *Information and Control*, vol. 10, pp. 65–103, 1967.

[23] D. L. Tebbe and S. J. Dwyer III, "Uncertainty and probability of error," *IEEE Trans. Information Theory*, vol. IT-14, pp. 516–518, May 1968.

[24] H. L. V. Trees, *Detection, estimation and modulation theory. 1, Detection, estimation and linear modulation theory*. John Wiley, 1968.

[25] A. Tulino and S. Verdú, "Monotonic decrease of the non-gaussianness of the sum of independent random variables: A simple proof," *IEEE Trans. Information Theory*, vol. 52, pp. 4295–4297, Sep. 2006.

[26] G. Vázquez-Vilar, A. Tauste-Campo, A. Guillén-Fábregas, and A. Martinez, "The meta-converse bound is tight," in *Proc. 2013 IEEE International Symposium on Information Theory*, July 2013, pp. 1730–1733.

[27] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Information Theory*, vol. 52, pp. 2165–2166, May 2006.

[28] Y. Wu and S. Verdú, "MMSE dimension," *IEEE Trans. Information Theory*, vol. 57, no. 8, pp. 4857–4879, Aug. 2011.

[29] ——, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Information Theory*, vol. 58, no. 3, pp. 1289–1301, Mar. 2012.

Why and How to Estimate Mutual Information?

Tsachy Weissman
Stanford University
tsachy@stanford.edu

Mutual information emerged in Shannon's 1948 masterpiece as the answer to the most fundamental questions of compression and communication. Since that time, however, it has been widely used and estimated in a variety of other disciplines. The two parts of this talk will respectively address two questions: why should we care about estimating mutual information, and how should we go about estimating it?

The first part will present a recent set of results establishing the status of mutual information as a "canonical" measure of relevance. Specifically, we show that, when measuring relevance by the extent to which one variable is helpful in estimating the other, the only loss function for estimation satisfying a natural invariance requirement is the logarithmic loss, and mutual information is the resulting dependence measure. Other objects with mutual information at their core inherit analogous justifications. A notable example is directed information, which emerges as the only measure of the 'degree to which one process is helpful in predicting the other' to satisfy the natural invariance requirement.

The second part of the talk will showcase a new approach to the estimation of mutual information between random objects with distributions residing in high-dimensional spaces (e.g., large alphabets), as is the case in increasingly many applications. We will discuss the shortcomings of traditional estimators, and suggest a new one achieving essentially optimum worst-case performance under L2 risk (i.e., achieves the minimax rates). We will exhibit a couple of examples illustrating the benefits afforded by this estimator in practice.

The talk is based on humbling recent collaborations with Jiantao Jiao, Kartik Venkat, Thomas Courtade, Yanjun Han, and Albert No reported on in [1], [2], [3], [4], and [5] (available on arXiv).

References

[1] J. Jiao, T. Courtade, K. Venkat and T. Weissman, "Justification of Logarithmic Loss via the Benefit of Side Information," submitted.

[2] J. Jiao, T. Courtade, A. No, K. Venkat and T. Weissman, "Information Measures: the Curious Case of the Binary Alphabet," to appear in *IEEE Trans. Inform. Theory*.

[3] J. Jiao, K. Venkat, Y. Han and T. Weissman, "Order-Optimal Estimation of Functionals of Discrete Distributions ," submitted.

[4] J. Jiao, K. Venkat and T. Weissman, "Non-Asymptotic Theory for the Plug-In Rule in Functional Estimation," submitted.

[5] J. Jiao, K. Venkat, Y. Han and T. Weissman, "Beyond Maximum Likelihood: from Theory to Practice," submitted.

Error Probability and Hypothesis Testing

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

Consider two random variables V and Y, where V takes values in a finite set $\mathcal{V}$ of cardinality $|\mathcal{V}| = M$, and Y is arbitrary. The joint distribution of these two random variables is described by P_{VY}. The problem of estimating V from an observation of Y is an M-ary hypothesis-testing problem. Since the joint distribution P_{VY} defines a prior distribution P_V over the alternatives, the problem is naturally cast within the Bayesian framework.

An M-ary hypothesis test is defined by a (possibly random) transformation $\mathcal{Y} \to \mathcal{V}$ described by the conditional distribution $P_{\hat{V}|Y}$. The average error probability of a test $P_{\hat{V}|Y}$ can be expressed as

$$\bar{\epsilon}(P_{\hat{V}|Y}) \triangleq \Pr\{\hat{V} \neq V\} \tag{1}$$

$$= \sum_{v,y} P_{VY}(v,y)\left(1 - P_{\hat{V}|Y}(v|y)\right). \tag{2}$$

Minimizing over all possible conditional distributions $P_{\hat{V}|Y}$ gives the smallest average error probability, namely

$$\bar{\epsilon} \triangleq \min_{P_{\hat{V}|Y}} \bar{\epsilon}(P_{\hat{V}|Y}). \tag{3}$$

The minimum in (3) is achieved by the test choosing the hypothesis v with largest maximum a posteriori (MAP) metric given the observation y, i.e.,

$$\bar{\epsilon} = 1 - \sum_{y} \max_{v'} P_{VY}(v', y). \tag{4}$$

For a binary hypothesis test between distributions P and Q, let $\alpha_\beta(P, Q)$ be the minimum type-I error for a maximum type-II error $\beta \in [0, 1]$.

Theorem 1: The average error probability of a M-ary hypothesis-testing problem satisfies

$$\bar{\epsilon} = \max_{Q_Y} \alpha_{\frac{1}{M}}\left(P_{VY}, Q_V^\star \times Q_Y\right) \tag{5}$$

$$= \max_{Q_Y} \sup_{\gamma \geq 0} \left\{ \Pr\left[\frac{P_{Y|V}(Y|V)P_V(V)}{Q_Y(Y)} \leq \gamma\right] - \gamma \right\} \tag{6}$$

where $Q_V^\star(v) \triangleq \frac{1}{M}$ for all v. Moreover, a maximizing distribution Q_Y in both expressions is

$$Q_Y^\star(y) \triangleq \frac{1}{\mu} \max_{v'} P_{VY}(v', y) \tag{7}$$

where $\mu \triangleq \sum_y \max_{v'} P_{VY}(v', y)$ is a normalizing constant.

This result shows that the error probability of Bayesian M-ary hypothesis testing can be expressed as the best type-I error probability of a binary hypothesis test discriminating between the original distribution P_{VY} and an alternative distribution $Q_V^\star \times Q_Y^\star$ with type-II-error constraint equal to $\frac{1}{M}$. We conclude that the MAP criterion, that minimizes the average error probability of an M-ary hypothesis test, can be alternatively used to solve a binary hypothesis-testing problem upon appropriately defining the alternative distribution. In the channel coding setting (5) coincides with the metaconverse bound [1, Th. 26] for the choice $Q_{VY} = Q_V^\star \times Q_Y$. Theorem 1 thus shows that the metaconverse is tight after optimization over the auxiliary distribution Q_{VY}.

Theorem 1 also provides an alternative characterization based on information-spectrum measures. In particular, the probability term in (6) corresponds to the tail probability of the information density $\log \frac{P_{VY}(v,y)}{P_Y(y)}$ being below a threshold $\log \frac{\gamma Q_Y(y)}{P_Y(y)}$ which, in general, is allowed to depend on the observation y. By choosing $Q_Y = P_Y$ in (6) we recover the Verdú-Han lower bound in the channel [2, Th. 4] and joint source-channel coding settings [3, Lem. 3.2]. By setting $Q_Y = Q_Y^\star$ and $\gamma = \mu$, the identity (6) can be interpreted as the error probability of an M-ary hypothesis test that, for each v, compares the posterior likelihood $P_{VY}(v, y)$ with a threshold equal to $\max_{v'} P_{VY}(v', y)$, i. e., this test emulates the MAP test yielding the exact error probability.

Acknowledgment

This is joint work with Gonzalo Vazquez-Vilar, Adrià Tauste Campo and Alfonso Martinez. This work has been funded in part by European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme under grant agreements 303633 and 329837 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03.

References

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[3] T. S. Han, "Joint source-channel coding revisited: Information-spectrum approach," *arxiv preprint arXiv:0712.2959v1*, 2007.

Subspace methods: An old technique and some recent developments

Ezio Biglieri
Department of Communication and Information Technologies
Universitat Pompeu Fabra, Barcelona, Spain

We examine the problem of identifying the active antennas in a receiver-training mode of space-modulated MIMO [3], where only N out of N_t antennas are allowed to transmit. The problem is formally identical to that of identifying the active users in synchronous CDMA multiuser detection. Since direct use of Random-Set Theory [2] may be too complex, we examine the use of subspace methods [13], [14], as advocated for example in [15].

These methods involve evaluation of the covariance matrix $\mathbf{R}$ of the observed signal, and the estimation of its eigenvalues and eigenvectors. Once these quantities are made available, signal subspace and noise subspace can be separated. The number N of active antenna is derived as the rank of the signal subspace, while the identity of the active antennas require deriving the covariance of the transmitted signal.

We examine in particular three subproblems to be solved in order to apply subspace methods: ① Reliable estimation of the covariance matrix, ② Reliable estimation of its eigenvalues and eigenvectors, and ③ Separation of the signal subspace from the noise subspace.

① The sample covariance matrix $\mathbf{R}_n$ is known to converge to the true matrix $\mathbf{R}$ almost surely as the number n of samples involved grows to infinity. Now, what is the minimum sample size n that guarantees approximation with a given accuracy? Random-matrix theory can be used to solve this problem. Yet, while asymptotic random-matrix theory offers remarkably accurate predictions as n grows to infinity, their sharpness at infinity is often counterweighted by lack of understanding of what happens in finite dimensions. Recent results involving nonasymptotic random-matrix theory [8], [9], [11], [12] will be reviewed in relation to problem ①.

② The standard estimate of eigenvalues and eigenvectors of $\mathbf{R}$ consists of the eigenvalues and eigenvectors of $\mathbf{R}_n$. These estimators are designed to yield good estimates when the sample size n is sufficiently large with respect to the observation dimension N_r. Now, when n is comparable in magnitude to N_r, better estimators can be obtained using a variety of approaches. Recent results by Xavier Mestre [6], [7] are reviewed, showing that estimators can be designed which outperform their standard counterparts in the small-sample regime.

③ Determining the dimension of the signal space, and hence N, requires the evaluation of the multiplicity of the smallest eigenvalue of $\mathbf{R}$. Now, if $\mathbf{R}_n$ is used in lieu of $\mathbf{R}$, the smallest eigenvalues may not be equal, but rather they tend to cluster around their true value. The determination of this dimension can be obtained by using model-order selection theory, based on Akaike results [1], [4], [5], [10].

Acknowledgment

This work was supported by Project TEC2012-34642.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. AC-19, pp. 716–723, 1974.

[2] E. Biglieri, E. Grossi, and M. Lops, "Random-set theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 7, no. 4, pp. 317–462, 2012.

[3] M. Di Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities, and implementation," *Proc. IEEE*, vol. 102, no. 1, pp. 56–103, January 2014.

[4] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[5] S. Huet, "Model selection for estimating the non zero components of a Gaussian vector," *ESAIM: Probability and Statistics*, vol. 10, pp. 164–183, March 2006.

[6] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates," *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 5113–5129, November 2008.

[7] X. Mestre, "On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices," *IEEE Trans. Signal Processing*, vol. 56, no. 11, pp. 5353–5368, November 2008.

[8] M. Rudelson, *Lecture Notes on Non-Asymptotic Theory of Random Matrices*. AMS Short Course on Random Matrices, San Diego, CA, January 2013.

[9] M. Rudelson, and R. Vershynin, "Non-asymptotic theory of random matrices: extreme singular values," *Proc. International Congress of Mathematicians*, Hyderabad, India, 2010.

[10] P. Stoica and Y. Selén, "Model-order selection: A review of information criterion rules," *IEEE Signal Processing Magazine*, pp. 36–47, July 2004.

[11] R. Vershynin, "Estimation of covariance matrices," *Workshop on Probability and Geometry in High Dimensions*, Paris, France, May 17–21, 2010.

[12] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," Chapter 5 of: *Compressed Sensing, Theory and Applications*, edited by Y. Eldar and G. Kutyniok, pp.210–268. Cambridge University Press, 2012.

[13] M. Wax, T.-J. Shan, and T. Kailath, "Spatio-temporal spectral analysis by eigenstructure methods," *IEEE Trans. Acoustics, Speech, and Signal processing*, vol. ASSP-32, no. 4, pp. 817–827, August 1984.

[14] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, April 1985.

[15] W.-C. Wu and K.-C. Chen, "Identification of active users in synchronous CDMA multiuser detection," *IEEE J. Select. Araeas Commun.*, vol. 16, no. 9, pp. 1723–1735, December 1998.

ML and CS Processing for DOA Estimation of Bird Sources

Kung Yao
UCLA

***Abstract*—For some years, we have been conducting research on the algorithms and applications of acoustical beamforming for source localization, enhancement, and separation of multiple bird sources with colleagues in the UCLA Evolutionary Biology and Complexity Dept. The ultimate goal of this NSF sponsored research is to permit humans to understand the language, grammar, and meanings of bird songs. Our specific efforts have been in the design, analysis, and fabrication of multiple acoustical arrays to serve as the front-end of data collecting systems to operate in diverse jungle and realistic field conditions in providing data for verifying various language modeling conjectures of bird songs. Spectra of radar and communication systems (including cellular applications) are narrow-band, while human speech and bird waveforms are wide-bands. By using sub-band methodologies, narrow-band beamforming techniques can be modified for our purposes. After DFT/FFT transformations, our beamformer uses ML estimation methods to perform far-field direction-of-arrival (DOA) estimation, as well as multiple source separation and enhancement. In this talk, we will summarize various algorithms and practical arrays we have built and their applications. Since the number of possible source DOA's is sparse relative to the azimuth and elevation search space, we provide some preliminary results on possibly using recent Compressive Sensing methodology to tackle this problem. Desirable RIP property in the selection of the randomizing operation in a CS scheme need not easily be met with the small number of sensors in our array. The Johnson–Lindenstrauss Lemma on transforming the sparse condition in a high dimensional space to a low dimensional space and still preserving the sparse condition may be relevant for us. Computational complexity of the ℓ_1 optimization needed in a CS operation also poses a challenge if we want real-time applications.**

Pattern Detection Filtering with Spiking Signals

Hans-Andrea Loeliger
ETH Zürich
ISI (D-ITET), Sternwartstr. 7
CH-8092 Zürich, Switzerland
Email: loeliger@isi.ee.ethz.ch

Abstract

A new model of self-timed pulse-based computation is proposed. The basic unit of computation is a feature detection filter, which looks for some feature in its multichannel-input signal and produces some sort of a score signal (or likelihood signal); whenever the score signal exceeds some threshold, an output pulse is generated. A layered network of such feature detection filters can be used for multiscale signal parsing. The feasibility of the proposed approach is demonstrated with a network that understands Morse code.

References

[1] H.-A. Loeliger, S. Neff, and Ch. Reller, "Self-synchronizing signal parsing with spiking feature-detection filters," *Proc. 52nd Annual Allerton Conference on Communication, Control, and Computing,* Monticello, Illinois, USA, Oct. 1–3, 2014.

A unified graphical approach to random coding for wireless networks

Stefano Rini
National Chiao Tung University (NCTU)
Department of Electrical and Computer Engineering
1001 University Road Hsinchu, Taiwan 300
Email: stefano@nctu.edu.tw

Andrea Goldsmith
Stanford University
Department of Electrical Engineering
Stanford, CA 94305
Email: andrea@ee.stanford.edu

***Abstract*—A unified approach to the derivation of rate regions for one-hop wireless networks is presented. The transmission scheme, derived for any memoryless, single-hop, k-user channel with or without common information, is obtained in two steps. The first step is user virtualization: each user is divided into multiple virtual sub-users using rate-splitting to preserve the rates of the original messages. This results in an enhanced channel with a possibly larger number users, for which more coding possibilities are available. Moreover, user virtualization provides a simple mechanism to encode common messages to any subset of users. Following user virtualization, the message of each user in the enhanced model is coded using any combination of coded time-sharing, superposition coding and binning. The interdependencies between the coding strategies at each node leads to exponential growth in the number of possible coding schemes, which has precluded general achievable schemes to date. A novel graph-based approach is used to represent all coding strategies, thereby circumventing this problem; nodes in the graph represent codewords while edges represent coding operations. This graph is used to construct a graphical Markov model which represents the statistical relationship among codewords based on the mapping between virtual users and actual users. Using this statistical representation of the overall codebook distribution, the error probability of the code is shown to vanish via a unified analysis based on the packing and covering lemmas. The rate bounds that define the achievable rate region are obtained by linking the error analysis to the properties of the graphical Markov model.**

Given any single-hop network, the largest achievable rate region under random coding is obtained by considering all possible rate-splitting strategies and taking the union over all possible ways to superimpose or bin the associated codewords via the graphical model. The achievable rates obtained based on this unified method encompass the best random coding achievable rates for all memoryless single-hop networks known to date, including broadcast, multiple access, interference, and cognitive radio channels, as well as new results for topologies not previously studied. We demonstrate the technique for several single-hop network topologies to either obtain the first known achievable rate regions for these topologies or to improve upon known achievable regions. Upper bounds and extensions to multi-hop networks will also be discussed.

it 4 5G

Angel Lozano
Universitat Pompeu Fabra (UPF)
C/ Roc Boronat 138
08018 Barcelona, Spain
Email: angel.lozano@upf.edu

***Abstract*—This presentation deals with 5th-generation wireless networks, and with the role that information theory may play on their design.**

I. Extended Abstract

With the 4th generation of wireless systems still being rolled out, challenges keep mounting: traffic demand doubles yearly and user needs and expectations keep growing at a very fast pace. Against this backdrop, the debate is open as to what shall come next in the evolution of wireless systems—what is loosely referred to as the 5th generation—and this debate stage can be a very fertile time for ideas to take root. This presentation intends to contribute to the ongoing discussion, and specifically on the role that information theory might play on the design of 5G wireless networks. The presentation is organized around several reflections that touch on this subject, but that also have broader conceptual implications for research.

Avoiding mediocrity

Emre Telatar
EPFL – I&C – LTHI
CH-1015 Lausanne, Switzerland
Email: emre.telatar@epfl.ch

***Abstract*—It is natural to view Arıkan's 2-by-2 polar construction as an operation that transforms a sequence of channels to a new sequence of channels, by combining the odd indexed members of the source sequence with their even indexed partner. The successive applications of this operation (with appropriate shuffling in between the applications) thus results in a 'sequence of sequence-of-channels'. We will show that the limiting states of this procedure are those sequences of channels for which mediocre channels forever avoid meeting each other. This observation yields a simple proof that Arıkan's procedure, when seeded with a sequence of identical channels (the classical stationary scenario) necessarily results in polarization. With a bit of help from combinatorics, one can further show that the same procedure when started with any sequence of channels also necessarily results in polarization.**

On a Reductionist View of Information Theory

Michelle Effros
California Institute of Technology
Dept. of Electrical Engineering, 136-93
Pasadena, CA
Email: effros@caltech.edu

Michael Langberg
SUNY Buffalo
214 Davis Hall
Buffalo, NY 14260
Email: mikel@buffalo.edu

***Abstract*—The network information theory literature contains many beautiful results describing capacity and source coding bounds for a variety of network types and topologies. While the literature develops and employs a host of common tools and strategies, to some extent, each new network model engenders its own new theory. This work explores "reduction" as a tool for network information theory research. The focus is on the use of reduction for demonstrating new relationships between seemingly disparate problems and the potential of this approach for deriving a new kind of unifying theory for the field.**

I. Introduction

Reduction is a proof strategy for transforming one problem into another. Problem A reduces to a distinct problem B if it can be shown that the availability of a solution for B would enable the construction of a solution for A. The power of the reductive strategy is that it enables the derivation of relationships between disparate problems even when solutions to both problems are unavailable. That is, one need not have a solution for problem B in order to demonstrate the above-described reduction; instead, it suffices to demonstrate that if a solution to B were available, then it could be used to derive a solution to A.

While reduction proofs are more commonly applied in the context of computational complexity theory and cryptography, a host of recent results demonstrates the power of reduction in information theory. Here the emphasis is not on algorithmic complexity but instead simply on showing that one problem can be solved by the solution of another. For example, [1] proves that 0-error code design for any multiple multicast network coding instance can be solved through the solution of 0-error code design for a corresponding multiple unicast network coding instance. Related later work [2] proves that the capacity region of any memoryless network with multiple multicast demands can be derived by finding the capacity region of a related memoryless network with multiple unicast demands. These results are derived despite the fact that both 0-error code design for multiple unicast network coding and capacity derivation for memoryless multiple unicast networks remain open problems.

Many of the information theoretic results derived to date focus on code design. For example, linear code design for network coding networks reduces to linear code design for index coding networks [3]. Code design (without the restriction to linear codes) for network coding networks reduces to code design for index coding networks [4]. Code design for k-unicast network coding networks reduces to code design for 2-unicast network coding networks in the 0-error coding regime [5].

In some but not all cases, reductions in code design have been used to derive corresponding reductions in the problem of solving network capacity regions. For example, solving the linear capacity region (i.e., the set of rate vectors asymptotically achievable by linear codes) for multiple multicast network coding networks reduces to solving the linear capacity region for multiple unicast index coding networks [6]. Linear capacity calculation for memoryless networks under multiple multicast demands reduces to linear capacity calculation for memoryless networks under multiple unicast demands [7]. Capacity calculation for networks of memoryless point-to-point channels reduces to capacity calculation for network coding networks [8]. Capacity calculation for memoryless networks with finite, known delays at each receiver [9] reduces to capacity calculation for memoryless networks without delays.

II. Potential

Rather than focusing on any specific reductive argument or result, this talk is intended to explore the potential of reductive strategies for uncovering important new connections between existing information theoretic problems and for inspiring new questions that may enrich the field.

One key area for advance is in understanding the relationships between information theoretic problems. For example, it would be useful to understand which problems are equivalent from the perspective of code design. Reductions are useful for uncovering such relationships: If A reduces to B and B reduces to A, then A and B are equivalent under the given reduction type. Reductive strategies are also useful for deriving hierarchical relationships between problems: If A reduces to B and B reduces to C, then A reduces to C. Combining equivalences and hierarchical relationships can yield rich taxonomies of problems. Even the short list of reductive results derived to date hints at such a taxonomy and suggests new questions for investigation. For example, existing results start the process of deriving independent (yet almost identical) taxonomies for the code design and capacity calculation problems. It would be useful to understand how the taxonomies are related. Does every reduction in code design imply a corresponding reduction in capacity calculation?

Reductive arguments can also be used to identify canonical problems whose solution would solve all problems in some larger class. Two types of canonical problems for information theory are discussed in [10]. An information theoretic problem A is said to be "hard" with respect to some class S of information theoretic problems (written "A is S-hard") if solution of problem A would solve all problems in class S. Problem A is said to be "complete" for S (written "A is S-complete") if A is S-hard and A is in S. Completeness results may serve to focus our attention on a particularly important subset of the space of problems. Hardness results demonstrate the existence of a problem outside the class whose solution would solve all problems in the class.

The previously discussed results highlight multiple unicast index coding as a complete network coding problem from the perspective of network code design and linear capacity calculation. These results also raise a number of interesting questions. For example: Are there additional special instances of network coding (of perhaps a completely different nature) that are complete for the class of network coding problems? Do there exist other natural classes such as network coding for which there exist complete problems?

Prior results also hint at the existence of "hard" problems for information theory. For example, the "edge removal" problem [11], [12] asks whether removing a single edge of capacity δ can ever change the capacity of a network coding network by more than δ in every dimension. For all but a few special cases, the edge removal problem is unsolved even in the case where δ approaches 0. In [13], [14], the edge removal problem is tied to the "network coding zero-vs-ε error" question, which asks how much (if at all) the capacity region of a network coding instance can change if we insist on zero- rather than asymptotically negligible error probability. Specifically, in [13], [14] it is established that if the removal of a single edge with vanishing capacity (in the block length) has a vanishing effect on the achievable capacity, then zero-error and vanishing error are equivalent for network coding capacity. The opposite direction of this connection, tying the network coding zero-vs-ε error problem to the edge removal problem, appears in [14]. In [15], the edge removal problem is tied to the problem of source dependence in networks. Here it is shown that the edge removal problem captures the difference between communicating over network coding instances in which the sources are independent (as commonly assumed) and instances in which the sources may be dependent. In [2] connections between edge removal and the completeness of the index coding problem with respect to *capacity* reductions are established. These rich connections demonstrate the potential impact of solving the edge removal problem on a wide variety of other open questions. They also raise a number of new questions: Are there other natural problems that are connected to the edge removal problem? Are there other canonical problems in the context of information theory that have such broad and diverse connections?

While the preceding discussion treats solutions of one problem through solution of another, reduction can also be used for bounding the solution of one problem through the derivation of solutions or bounds for another problem. The "network equivalence" paradigm of [8], [16] illustrates such an approach. Specifically, network equivalence seeks to study the capacity region of networks built from independent memoryless channels by designing *bounding models* for each channel in the network. Channel A is an upper bounding model for channel B if replacing channel B by channel A in *any* network yields a new network whose capacity region is a superset of the capacity region of the original network. Similarly, replacing channels by their lower bounding models yields a new network whose capacity region is a subset of that of the original network. Several channel models and their bounding counterparts were studied in [8], [16], including point-to-point channels, broadcast channels, and multiple access channels.

III. Summary

Existing recent studies in the context of information theory hint at an emerging classification of communication problems through the lens of reductions. This talk is intended to survey existing reductive arguments in the context of information theory and to outline some of the intriguing possibilities for future studies.

References

[1] R. Dougherty and K. Zeger. Nonreversability and equivalent constructions of multiple-unicast networks. *IEEE Transactions on Information Theory*, 52(11):5067–5077, 2006.

[2] M. F. Wong, M. Langberg, and M. Effros. On an equivalence between multicast and multiple unicast. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2011. IEEE.

[3] S. Y. El Rouayheb, A. Sprintson, and C. Georghiades. On the relation between the index coding and the network coding problems. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1823–1827, Toronto, July 2008. IEEE.

[4] M. Effros, S. El Rouayheb, and M. Langberg. An equivalence between network coding and index coding. In *Proceedings of the IEEE International Symposium on Information Theory*, Istanbul, Turkey, July 2013. IEEE.

[5] S. Kamath, D. Tse, and C.-C. Wang. Two-unicast is hard. In *Proceedings of the IEEE International Symposium on Information Theory*, Honolulu, 2014.

[6] H. Maleki, V. Cadambe, and S. Jafar. Index coding: an interference alignment perspective. In *Proceedings of the IEEE International Symposium on Information Theory*, Cambridge, MA, July 2012. IEEE.

[7] M. F. Wong, M. Langberg, and M. Effros. Linear capacity equivalence between multiple multicast and multiple unicast. In *Proceedings of the IEEE International Symposium on Information Theory*, Honolulu, HI, July 2014. IEEE.

[8] R. Koetter, M. Effros, and M. Médard. A theory of network equivalence – Part I: Point-to-point channels. *IEEE Transactions on Information Theory*, 57:972–995, February 2011.

[9] M. Effros. On dependence and delay: Capacity bounds for wireless networks. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, Paris, France, April 2012. IEEE.

[10] M. Effros and M. Langberg. Is there a canonical network for network information theory? In *Proceedings of the IEEE Information Theory Workshop*, Hobart, Tasmania, November 2014.

[11] T. Ho, M. Effros, and S. Jalali. On equivalence between network topologies. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2010. IEEE.

[12] S. Jalali, M. Effros, and T. Ho. On the impact of a single edge on the network coding capacity. In *Information Theory and Applications Workshop*, pages 1–5, San Diego, California, February 2011. IEEE.

[13] T. Chan and A. Grant. On capacity regions of non-multicast networks. In *Proceedings of the IEEE International Symposium on Information Theory*, Austin, Texas, June 2010. IEEE.

[14] M. Langberg and M. Effros. Network coding: Is zero error always possible? In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2011. IEEE.

[15] M. Langberg and M. Effros. Source coding for dependent sources. In *Proceedings of the IEEE Information Theory Workshop*, Lausanne, September 2012. IEEE.

[16] R. Koetter, M. Effros, and M. Médard. A theory of network equivalence – Part II: Multiterminal channels. *IEEE Transactions on Information Theory*, 60(7):3709–3732, 2014.

Information Age

Anthony Ephremides

Abstract

A neglected figure of merit in network system performance is the so-called "Age of Information". This is not the same as message delay. Rather, it reflects the time lapse between the instant of the generation of the most recent available value of a process at the receiver and the current time. This is an especially useful performance index for applications where an on-going process is being monitored. Also, there are similar issues in surveillance applications and machine-to-machine communications. We provide an overview of what has been accomplished so far in this exciting new direction and outline the challenges that lie ahead.

Recent Results on Secrecy and Stealth

Gerhard Kramer
Lehrstuhl für Nachrichtentechnik
Technische Universität München
Email: gerhard.kramer@tum.de

ABSTRACT

The goal of the talk is to give basic insight into wiretap channels. For this purpose, we first review Shannon's channel coding theorem, Wyner's common information, and Han and Verdú's resolvability. We then consider Wyner's wonderful wiretap channel and list several secrecy measures. Yet another security measure is introduced that is based on informational divergence and that includes both strong secrecy and stealth (or covert) communication. The new measure leads to a capacity region that we relate to Wyner's secrecy region. The converse follows by a short proof that uses a telescoping identity. The coding theorem is established by using a simple proof whose key step is applying Jensen's inequality to the logarithm, as well as a few standard typicality arguments. An operational meaning for stealth follows in the usual way by using binary hypothesis testing.

The talk is based on joint work with Jie Hou from TUM.

Lattices and Linear Codes

Joseph J. Boutros and Nicola di Pietro
Texas A&M University at Qatar
c/o Qatar Foundation, Education City
23874 Doha, Qatar
Email: {joseph.boutros, nicola.dipietro}@qatar.tamu.edu

***Abstract*—We consider real lattices built from error-correcting codes. After recalling the definition of a lattice, we review Construction A over a prime field. Construction A combined with non-binary low-density parity-check (LDPC) codes gives rise to the family of LDA lattices. Generalized low-density (GLD) codes are another interesting type of codes on graphs yielding the ensemble of GLD lattices. This manuscript ends with the analysis of the goodness of LDA and GLD lattices when used for communication over a Gaussian channel.**

I. Introduction

Lattice are mathematical structures with specific algebraic and topological properties. We consider the simplest form of lattices, i.e. lattices in real Euclidean spaces equipped with the standard scalar product. In communication theory, lattices can play different roles in the processing and the transmission of information. They are suitable for vector quantization of analog sources, for channel coding as coded modulations, and also for joint source-channel coding. In the recent literature, lattices are found to be good tools in network coding and secure coding at the physical layer. More information can be found in [1] and references therein.

A lattice is a $\mathbb{Z}$-module of the Euclidean vector space $\mathbb{R}^N$. Concretely, it is simply a discrete additive subgroup of $\mathbb{R}^N$, according to the following definition:

Definition 1: Given M and N two natural numbers, $M \leq N$, and given a set of M linearly independent vectors $\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_M} \in \mathbb{R}^N$, an *$M$-dimensional lattice* Λ is defined as the set of all integer linear combinations of the $\mathbf{b_i}$'s:

$$\Lambda = \left\{ \mathbf{x} \in \mathbb{R}^N : \mathbf{x} = \sum_{i=1}^{M} z_i \mathbf{b_i},\ z_i \in \mathbb{Z} \right\}. \qquad (1)$$

The set $\{\mathbf{b_i}\}_{i=1}^{M}$ is a basis of Λ and M is the rank of Λ. We consider full rank lattices ($M = N$) in this paper. The $N \times N$ matrix G whose rows are the $\mathbf{b_i}$'s is a generator matrix.

Given a generator matrix G, the lattice Λ is

$$\Lambda = \left\{ \mathbf{x} \in \mathbb{R}^N : \mathbf{x} = \mathbf{z}G,\ \mathbf{z} \in \mathbb{Z}^N \right\} = \mathbb{Z}^N G. \qquad (2)$$

We define the fundamental volume of the lattice as $\mathrm{Vol}(\Lambda) = |\det(G)|$. Let $H = G^{-1}$. Another definition of Λ is

Definition 2:

$$\Lambda = \left\{ \mathbf{x} \in \mathbb{R}^N : \mathbf{x}H \text{ is an integer vector} \right\}. \qquad (3)$$

This second definition of a lattice appeared in modern coding theory with a sparse matrix H. The sparsity of H is essential for the iterative decoding of Λ in high dimensions via message passing over its factor graph [2].

In the literature on mathematics and coding theory, the pioneering work by Leech and Sloane [3] opened new ways for building lattices out of error-correcting codes. More than four decades later, Construction A proposed by Leech and Sloane is considered to be the most promising lattice construction both for theoretical and practical reasons. Other algebraic constructions are found in the bible of sphere packings and lattices [4]. Construction A is described in the next section. Section III elucidates the construction of LDA [5] [6] [7] and GLD [8] [9] lattices considered to be among the most powerful families of lattices nowadays due to their reduced-complexity decoding and their excellent error rate performance.

II. Construction A over $\mathbb{F}_p$

Let $C[N,K]_p$ be a linear code of dimension K and length N defined over $\mathbb{F}_p$, where $\mathbb{F}_p$ is a prime finite field [10]. Elements of $\mathbb{F}_p$ can be embedded in $\mathbb{Z}$ by two standard mappings. The first mapping is natural, it identifies elements of $\mathbb{F}_p$ with the most common coset leaders of $\mathbb{Z}/p\mathbb{Z}$ which is the set of integers $\{0, 1, \ldots, p-1\}$. The second mapping is centered on 0 and corresponds to the set of integers $\{-(p-1)/2, \ldots, -1, 0, 1, \ldots, (p-1)/2\}$, for p odd prime, which is more convenient to study the distribution of Euclidean distances. The lattice $p\mathbb{Z}^N$ has p^N cosets in $\mathbb{Z}^N$. A subset of size p^K cosets is selected among the p^N cosets via the code C. This yields a lattice, a coset code in Forney's terminology with the formula [11]:

$$\Lambda = C[N,K]_p + p\mathbb{Z}^N. \qquad (4)$$

The ring $\mathbb{Z}$ can be replaced by other rings such as the ring of Gaussian integers $\mathbb{Z}[i]$ and the ring of Eisenstein integers $\mathbb{Z}[\omega]$. Construction A can also be used to build Λ from the ring of integers $O_{\mathbb{K}}$ of a number field $\mathbb{K}$ by $\Lambda = C[N,K]_p + I^N$, where I is a prime ideal of $O_{\mathbb{K}}$ and $C[N,K]_p$ should be correctly embedded in $O_{\mathbb{K}}$ [4].

Construction A defined by (4) should be thought of as drawing p^K points representing the codewords of C inside the cube $[0,p[^N$ and then paving the whole space $\mathbb{R}^N$ by translating the cube by multiples of p in all directions. The volume of Λ is $\mathrm{Vol}(\Lambda) = p^{N-K}$ and its theta series coincides with the theta series of C inside the ball of radius $(p-1)/2$ centered on the origin.

III. LDA and GLD Constructions

Let Λ be a lattice built via formula (4). If $C[N,K]_p$ is an LDPC code [12] defined over $\mathbb{F}_p$, then Λ is called an LDA lattice. The family of LDA lattices was studied in [5] [6] [7]. Assume that the variable node degree and the check node degree are chosen such that the LDPC graph has good expansion properties. Under such a condition, the LDA ensemble defined by random coefficients in $\mathbb{F}_p$ associated to LDPC check nodes is capable of attaining Poltyrev limit for infinite constellations [6] and Shannon capacity for finite constellations [7].

If $C[N,K]_p$ is a GLD code defined over $\mathbb{F}_p$, then Λ is called a GLD lattice. See [13] [14] for a definition of GLD codes. We will focus in the remaining part of this section on the newly proposed family of GLD lattices [8]. We start by introducing notations to reach a general definition of GLD lattices that is not necessarily related to Construction A.

Consider two lattices Λ_1 and Λ_2 of same rank N. Our objective is to create a better lattice Λ by taking their intersection

$$\Lambda = \Lambda_1 \cap \Lambda_2 \tag{5}$$

The intersection implies a superposition of the constraints on a lattice point x given by the two matrices $H_1 = G_1^{-1}$ and $H_2 = G_2^{-1}$, where G_1 and G_2 are generator matrices for Λ_1 and Λ_2 respectively.

$$\Lambda = \left\{ \mathbf{x} \in \mathbb{R}^N : \mathbf{x}H \text{ is integer, where } H = [H_1 H_2] \right\}. \tag{6}$$

Do we get an improvement from (5) or (6)? Think about the figure of merits of Λ, e.g. the Hermite constant defined as the ratio of the minimum Euclidean distance of Λ by its normalized fundamental volume

$$\frac{d^2_{Emin}(\Lambda)}{\mathrm{Vol}(\Lambda)^{2/N}}. \tag{7}$$

The intersection results in an increase of both $d^2_{Emin}(\Lambda)$ and $\mathrm{Vol}(\Lambda)$. It is not sure whether we get a better figure of merit. Other issues encountered while making the intersection are

- For large dimensions N, how can we build $\Lambda = \Lambda_1 \cap \Lambda_2$ if we do not know how to build the two lattices Λ_1 and Λ_2 in advance?
- $\Lambda_1 \cap \Lambda_2$ may be equal to $a\mathbb{Z}^N$, i.e. it is an uncoded modulation. A bi-dimensional illustration is given in Figure 1.
- $\Lambda_1 \cap \Lambda_2$ may have a lower rank. A bi-dimensional illustration is given in Figure 2.
- $\Lambda_1 \cap \Lambda_2$ may be reduced to $\{0\}$. A bi-dimensional illustration is given in Figure 3.

In order to solve the issues listed above, let Λ_0 be an elementary lattice of small dimension n. Let $L = N/n$, L large enough. Define $\Lambda_1 = \Lambda_0^{\oplus L}$, i.e. the direct sum of L copies of Λ_0. Finally take $\Lambda_2 = \pi(\Lambda_1)$, where π is a permutation of $\{1, 2, \ldots, N\}$. Our definition of a GLD lattice is

Definition 3:

$$\Lambda = \Lambda_1 \cap \Lambda_2 = \Lambda_0^{\oplus L} \cap \pi(\Lambda_0^{\oplus L}). \tag{8}$$

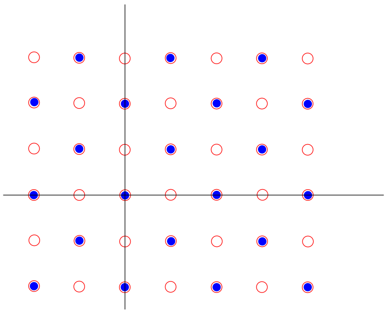

Figure 1. $D_2 \cap \mathbb{Z}^2 = D_2 = \sqrt{2}Rot(\mathbb{Z}^2)$ (uncoded rotated modulation).

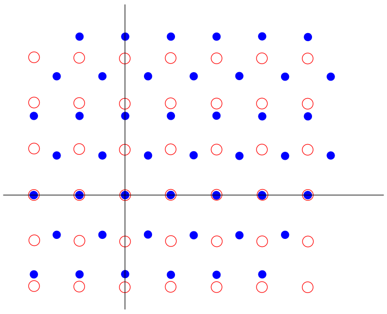

Figure 2. $A_2 \cap \mathbb{Z}^2 = \mathbb{Z}$ (rank was reduced).

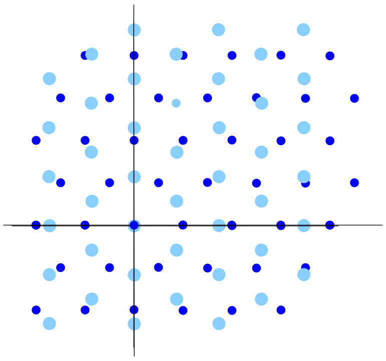

Figure 3. $A_2 \cap \pi(A_2) = \{0\}$ (collapse to 0).

For Λ_1, its matrix H_1 has L copies of H_0. For Λ_2, H_2 has the same rows as H_1 but their order is defined by π (row permutation).

$$H_1 = \begin{bmatrix} H_0 & 0 & \ldots & 0 \\ 0 & H_0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & H_0 \end{bmatrix} \tag{9}$$

and $H = [H_1 H_2] = [H_1 \ \pi(H_1)]$ is $N \times 2N$. A good choice for the elementary lattice Λ_0 is to use Construction A. $\Lambda_0 = C_0[n,k]_p + p\mathbb{Z}^n$. We get an integer GLD lattice

$$\begin{aligned}\Lambda &= (C_0 + p\mathbb{Z}^n)^{\oplus L} \cap \pi((C_0 + p\mathbb{Z}^n)^{\oplus L}) \\ &= \left(C_0^{\oplus L} \cap \pi(C_0^{\oplus L})\right) + p\mathbb{Z}^N \\ &= C_{\mathrm{GLD}}[N,K]_p + p\mathbb{Z}^N. \qquad (10)\end{aligned}$$

Proposition 1: The integer GLD lattice satisfies $p\mathbb{Z}^N \subseteq \Lambda \subseteq \mathbb{Z}^N$, Λ has rank N, $\mathrm{Vol}(\Lambda) = p^{N-K} = p^{2N(1-k/n)}$, and $\min(p^2, d_H(C_{GLD})) \leq d^2_{Emin}(\Lambda) \leq p^2$ where $d_H(C_{GLD})$ is the minimum Hamming distance of C_{GLD}.
Non-binary GLD codes $C_{\mathrm{GLD}}[N,K]_p$ are asymptotically good as shown in [9]. This property of their minimum Hamming weight has a direct effect on their gap to capacity on unconstrained Gaussian channels.

The performance of a GLD lattice under iterative decoding in dimensions $N = 1000$ and $N = 32000$ is plotted in Figure 4. It shows the symbol error rate (error probability on lattice coordinates) versus the gap in decibels to Poltyrev limit on a Gaussian channel. The finite field size is $p = 11$.

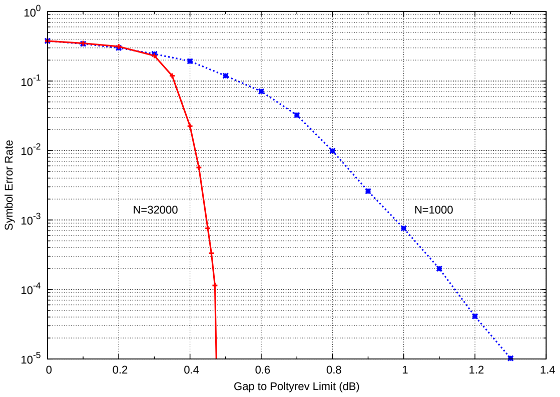

Figure 4. Numerical results for a GLD lattice, $\Lambda_0 = [4,3,2]_{11} + 11\mathbb{Z}^4$.

Let Λ_f be a lattice and Λ a sublattice both built from Construction A with codes of rate R_f and R respectively. A finite Voronoi constellation is built from the quotient group Λ_f/Λ and its size is [11]

$$|\Lambda_f/\Lambda| = p^{N(R_f-R)}. \qquad (11)$$

In the next section, we discuss the goodness of finite LDA and GLD lattice constellations.

IV. Goodness of LDA and GLD Constellations

A finite Voronoi constellation carved from the LDA ensemble is proven to achieve Shannon capacity [7]. An overview and discussion of the proof is given below. The same strategies can be applied to GLD ensembles with similar results. See [9] as an intermediate step for the proof of GLD goodness on the unconstrained Gaussian channel. In this section, we try to give a general description of the exact proof in [7], by the means of a heuristic argument that does not take into account all the probabilistic and asymptotic aspects of the rigorous demonstration. Our result is based on the following facts:

- The points of the Voronoi constellation typically have the same norm and lie very close to the surface of a sphere of a given radius (see Lemma 4.3 in [7]).
- The AWGN noise is typically orthogonal to the sent vector, in the sense that, if $\mathbf{x}$ is our transmitted constellation point and $\mathbf{w}$ is the noise, then $\mathbf{x}\mathbf{w}^T$ is relatively small in norm (cf. Lemma 4.4 in [7]).
- The effective noise due to MMSE scaling and the sent point are not decorrelated. Consequently, it is not possible to show that lattice decoding works with very high probability independently of the sent point. Nevertheless, Theorem 4.1 in [7] is based on the fact that the number of points for which this does not happen is not big enough to perturb the average error probability of the family.
- For a certain channel output (MMSE-scaled, in this case), we look for lattice points inside a sphere centered at it and with a typical radius to be specified later. There will be no decoding error if the only point in this *decoding sphere* is the transmitted one.

Consider that when we use the adverb "typically", we mean "with probability tending to 1 when N tends to infinity". The accurate proof is treated in all detail in [7], but let us try to understand the geometric sense of the elements that we have just listed. So, suppose that the channel input is a point $\mathbf{x}$ whose norm is fixed to be $||\mathbf{x}|| = \sqrt{NP}$, for some $P > 0$. Suppose also that $\mathbf{x}\mathbf{w}^T = 0$; if $\mathbf{y} = \mathbf{x} + \mathbf{w}$ is the channel output, then $||\mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{w}||^2$. Now, let us multiply $\mathbf{y}$ by a scalar value $\alpha = P/(\sigma^2 + P)$ (Wiener coefficient) such that the distance between $\mathbf{x}$ and $\alpha\mathbf{y}$ is minimized. This lets us guess that MMSE lattice decoding helps in bringing the decoder input closer to the sent point.

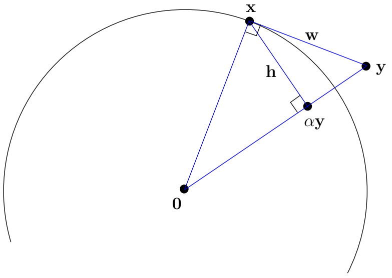

Figure 5. Geometric interpretation of MMSE scaling.

The receiver passes $\alpha\mathbf{y}$ to the lattice decoder and there will be no decoding error if there is no other lattice point closer to $\alpha\mathbf{y}$ than $\mathbf{x}$. We will show that this typically happens if:

1) SNR > 1.
2) $P \approx p^{2(1-R)}/2\pi e$;
3) $||\alpha\mathbf{y} - \mathbf{x}||^2$ asymptotically does not exceed $Np^{2(1-R_f)}/2\pi e$.

Notice that the previous bound concretely means that our constellation tolerates an "effective" noise after MMSE scaling whose variance per dimension is at most as strong as the noise corresponding to Poltyrev capacity. We intuitively understand that this can be the good condition, admitting that no issue comes from the fact that the "effective" noise and the sent point $\mathbf{x}$ are not decorrelated (here, we have no dither to guarantee that).

The condition on the signal-to-noise ratio can be simply understood with the following argument: let us call $\mathbf{h} = \alpha\mathbf{y} - \mathbf{x}$ and suppose that it takes the maximum value according to the third condition above here, $||\mathbf{h}||^2 = Np^{2(1-R_f)}/2\pi e = N\sigma_{\text{dec}}^2$. We use the index "dec" to indicate that the quantity corresponds to the (upper bound of the) "decodable" effective noise. If we want good decoding, we need $\alpha\mathbf{y}$ to be closer to $\mathbf{x}$ than to $\mathbf{0}$, because the latter always belongs to the lattice constellation; in other terms, it is necessary that $||\alpha\mathbf{y}||^2 > ||\mathbf{h}||^2$. Again, a Euclidean geometry argument based on Figure 5 shows that (always supposing that $\mathbf{x}\mathbf{w}^T = 0$)

$$N\sigma_{\text{dec}}^2 = ||\mathbf{h}||^2 = \frac{||\mathbf{x}||^2||\mathbf{w}||^2}{||\mathbf{y}||^2} = \frac{N^2P\sigma^2}{NP+N\sigma^2} = \frac{NP\sigma^2}{P+\sigma^2}, \tag{12}$$

while

$$||\alpha\mathbf{y}||^2 = \frac{P^2(NP+N\sigma^2)}{(P+\sigma^2)^2} = \frac{NP^2}{P+\sigma^2}.$$

Then, $||\alpha\mathbf{y}||^2 > ||\mathbf{h}||^2$ becomes

$$\frac{NP^2}{P+\sigma^2} > \frac{NP\sigma^2}{P+\sigma^2},$$

that is $P > \sigma^2$ or, equivalently, $\text{SNR} > 1$.

Taking $||\mathbf{h}||^2 = N\sigma_{\text{dec}}^2$ corresponds to a maximum rate for the constellation that equals capacity, as can be understood from the following calculation (that, again, is based on the approximations done till now and has only a demonstrative purpose): from (12) we can derive that

$$\sigma^2 = \frac{P\sigma_{\text{dec}}^2}{P-\sigma_{\text{dec}}^2}.$$

This implies that

$$\text{SNR} = \frac{P}{\sigma^2} = \frac{P}{\sigma_{\text{dec}}^2} - 1.$$

To conclude, recall that we make the hypothesis that $P = p^{2(1-R)}/2\pi e$; this, together with (11) leads to

$$\begin{aligned}\frac{1}{2}\log_2(1+\text{SNR}) &= \frac{1}{2}\log_2\left(\frac{P}{\sigma_{\text{dec}}^2}\right) \qquad (13)\\ &\approx \frac{1}{2}\log_2(p^{2(R_f-R)})\\ &= \frac{1}{N}\log_2(p^{N(R_f-R)}), \qquad (14)\end{aligned}$$

which is the rate of our constellation.

Originality of our proof and lattice decoding of $\alpha\mathbf{y}$: what we have explained till now gives an intuitive description of the typical geometry that characterizes the AWG noise and the random Voronoi constellations of Construction A nested lattices. Now we drop a hint on the original idea behind our proof that allows to avoid dithering. The main argument is the following: if $\alpha\mathbf{y}$ is the real point that the receiver passes to the lattice decoder, we would like to ensure that the only lattice point inside the *decoding sphere* $B_{\alpha\mathbf{y},N}(\sqrt{N}\sigma_{\text{dec}})$ is the sent point. It was shown in [7] that the probability of a decoding failure (without MMSE) is asymptotically negligible for the unconstrained channel, independently on the sent point $\mathbf{x}$. This last feature makes the big difference with the case of MMSE lattice decoding. Indeed, the average argument that we apply will lead to the estimation of (a more elaborated version of) the following sum:

$$\sum_{\mathbf{z}\in B_{\alpha\mathbf{y},N}(\sqrt{N}\sigma_{\text{dec}})} \mathcal{P}\{\mathbf{z}\in\Lambda_f \mid \mathbf{x}\in\Lambda_f\}. \tag{15}$$

The multiplication by α adds some correlation between $\mathbf{x}$ and $\alpha\mathbf{y}$. One can interpret Erez and Zamir's dithering technique as a method of eliminating this correlation [1].

We do not use dither and consequently there will be some points $\mathbf{x}$'s for which the previous sum takes a "big" value. Our analysis shows that the proportion of this kind of $\mathbf{x}$'s in the constellation is very small and the total error probability still goes to 0 when N tends to infinity.

Acknowledgment

The research work presented in this paper on LDA and GLD lattices is supported by QNRF, a member of Qatar Foundation, under NPRP projects 5-597-2-241 and 6-784-2-329.

References

[1] R. Zamir, *Lattice Coding for Signals and Networks*, Cambridge, 2014.
[2] N. Sommer, M. Feder, and O. Shalvi, "Low-density lattice codes," *IEEE Trans. on Inf. Theory*, vol. 54, no. 4, pp. 1561-1585, April 2008.
[3] J. Leech and N.J.A. Sloane, "Sphere packing and error-correcting codes,“ Canadian Journal of Mathematics, no. 23, pp. 718-745, 1971.
[4] J.H. Conway and N.J.A. Sloane. Sphere Packings, Lattices and Groups. Springer-Verlag, New York, 3rd edition, 1999.
[5] N. di Pietro, J.J. Boutros, G. Zémor, and L. Brunel, "Integer low-density lattices based on Construction A," *Proc. of the 2012 IEEE Information Theory Workshop*, pp. 422-426, Lausanne, Sept. 2012.
[6] N. di Pietro, G. Zémor, and J.J. Boutros "New results on Construction A lattices based on very sparse parity-check matrices," *Proc. of the 2013 IEEE Intern. Symp. on Inf. Theory (ISIT)*, pp. 1675-1679, July 2013.
[7] N. di Pietro, "On infinite and finite lattice constellations for the additive white gaussian noise channel," PhD thesis, Univ. Bordeaux, Jan. 2014. PDF downloadable at www.josephboutros.org/coding/diPietro_thesis.pdf.
[8] J.J. Boutros, N. di Pietro, and N. Basha, "Generalized low-density (GLD) lattices," *2014 IEEE Information Theory Workshop*, Hobart, Nov. 2014.
[9] N. di Pietro, N. Basha, and J.J. Boutros "Non-Binary GLD Codes and their Lattices," *2015 IEEE Information Theory Workshop*, Jerusalem, April 2015.
[10] F.J. Williams and N.J.A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam, 1977.
[11] G. D. Forney, "Coset codes I: introduction and geometrical classification,", *IEEE Trans. on Inf. Theory*, vol. 34, no. 5, pp. 1123-1151, 1988.
[12] T. Richardson and R. Urbanke. Modern Coding Theory. Cambridge University Press, New York, 2008.
[13] O. Pothier, L. Brunel, and J.J. Boutros, "A low complexity FEC scheme based on the intersection of interleaved block codes," *IEEE Veh. Tech. Conf.*, vol. 1, pp. 274-278, Houston, May 1999.
[14] J.J. Boutros, O. Pothier, and G. Zémor, "Generalized low density (Tanner) codes," *IEEE Intern. Conf. on Comm. (ICC)*, vol. 1, pp. 441-445, Vancouver, June 1999.

Codes on Graphs: An Overview

G. David Forney, Jr.
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
Email: forney@mit.edu

***Abstract*—We give an overview of the history of the field of "codes on graphs," including a survey of recent developments.**

I. Introduction

The subject of "codes on graphs" is concerned with the representation of codes by efficient graphical models. Such a model may be used to specify a decoding algorithm, whose complexity is governed by the complexity of the model. It may also give insight into structural properties of the code.

In the style of behavioral system theory [25], these models involve three kinds of elements: external (symbol) variables, representing code symbols; internal (state) variables, which may be freely specified by the model designer; and constraint codes, each of which involves a subset of the external and internal variables. The *behavior* $\mathfrak{B}$ of the model is the set of all symbol/state variable configurations that satisfy all constraints. The *code* $\mathcal{C}$ represented by the model is the set of all symbol variable configurations that appear in some valid symbol/state configuration in $\mathfrak{B}$.

For example, Figure 1 shows three graphical models for the well-known $(24, 12, 8)$ Golay code: a conventional trellis (state-space) model, represented by a simple chain graph [19]; a cycle-free representation on a simple "cubic" graph [22]; and a tail-biting trellis model on a single-cycle graph [1]. All parameters are the best possible (smallest) for the model type.

Linear codes are often used for Hamming-space coding, whereas group codes are often used for Euclidean-space coding (*e.g.,* lattice codes, trellis codes). Consequently, we generally consider linear or group models; *i.e.,* the variable alphabets are vector spaces or groups, and the constraint codes are linear or group codes. Interestingly, almost all results in this field apply equally to linear or group models, and are most easily and insightfully proved by elementary group-theoretic techniques.

II. History

The earliest "codes on graphs" were Gallager's low-density parity-check (LDPC) codes [14] (although Gallager never drew a graph). Gallager also invented the iterative sum-product ('"belief propagation") decoding algorithm. These remarkable achievements were largely forgotten for many decades. Twenty years later, Tanner [20] restarted the field with actual graphs (bipartite "Tanner graphs") and many important results, including the optimality of the sum-product algorithm on cycle-free graphs; however, his work was also forgotten.

The first "codes on graphs" of practical interest were linear convolutional codes. In the 1960s, it was recognized that linear convolutional codes could be understood as discrete-time linear systems over finite fields, and analyzed using discrete-time linear systems theory [4]. A unique feature of convolutional codes was their depiction using a trellis diagram, due to the discrete finite alphabets used in coding [5].

In the 1980s, the emergence of Euclidean-space codes (*e.g.,* lattice and trellis codes) led to increased interest in block and trellis codes over groups as "coset codes" [2], [6]. It was found that many of the results of discrete-time linear systems theory could be rederived using only their group properties [12].

The 1990s saw much research towards finding efficient trellis representations of linear block codes, well summarized in [21]. The main result is that a code with a given symbol ordering has an essentially unique minimal trellis representation; the main problem is therefore to find the best ordering for a given code. The same result holds for group codes [12].

A more important practical development in the 1990s was the invention of turbo codes, which used linear convolutional codes, and the rediscovery of Gallager's LDPC codes, which were both shown to be capacity-approaching in a practical sense. The important thesis of Wiberg [23], [24], entitled "Codes and decoding on general graphs," reawakened the field of "codes on graphs" by rediscovering many of the results of Tanner, and generalizing them to systems with state variables, thereby developing a unified theory of LDPC codes, turbo codes, trellis codes, and their decoding algorithms.

One offshoot of this work was increased interest in tail-biting trellis codes, which may be understood as trellis codes on a circular time axis. Wiberg showed that the maximum state space size for a tail-biting trellis model could be as small as the square root of the best maximum state space size for a conventional trellis model. For example, for the $(24, 12, 8)$ Golay code, the best conventional trellis model, shown in Figure 1(a), has a state space size of 256 at its center, whereas the tail-biting trellis model shown in Figure 1(c) has all state space sizes equal to 16. A general theory of linear tail-biting trellis representations was developed in [16].

These developments triggered the development of "factor graphs" as a general method of representing "sum-of-products" expressions (*e.g.,* partition functions) by graphical models [17]. If all "factors" are indicator functions of codes, then a factor graph is isomorphic to a graphical model of a code, but the factor graph formalism is much more general.

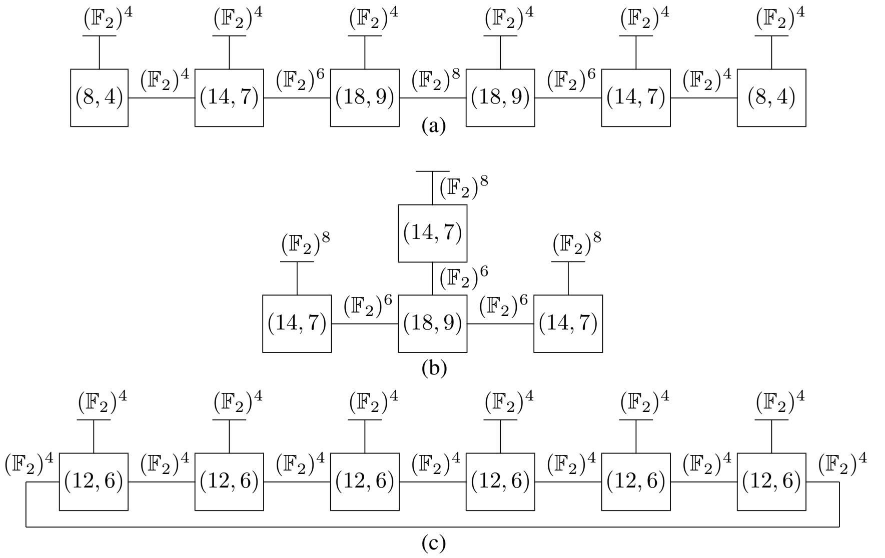

Fig. 1. Graphical models for $(24, 12, 8)$ Golay code: (a) conventional trellis; (b) cycle-free; (c) tail-biting trellis.

III. More recent developments

In [7], the foundations of the subject were revisited. It was shown that any realization $\mathcal{R}$ defined by sets of symbol variables, state variables and constraints could be made into a *normal realization* by trivial modifications. Furthermore, a normal realization has a natural representation by a *normal graph*, in which constraints are represented by vertices, state variables by edges, and symbol variables by half-edges. (The three examples of Figure 1 are normal graphs.) Loeliger [18] subsequently integrated these conventions into “normal factor graphs."

An important consequence of this reformulation is *normal realization duality*: if $\mathcal{R}$ is a normal realization that realizes a linear or group code $\mathcal{C}$, then a dual normal realization $\mathcal{R}^\circ$ may be defined by a set of simple local substitutions (*i.e.,* replace each constraint $\mathcal{C}_i$ by its dual constraint $(\mathcal{C}_i)^\perp$, each variable by its dual variable, and each edge by a sign-inverting edge); then $\mathcal{R}^\circ$ realizes the dual code $\mathcal{C}^\perp$. For example, in each of the binary linear realizations of Figure 1, each of the constraint codes is self-dual, so $\mathcal{R}^\circ = \mathcal{R}$, implying $\mathcal{C}^\perp = \mathcal{C}$. This general theorem has myriad applications, including showing how to decode a graph by Fourier-transforming variables and decoding the dual graph (of which a particular instance is the “tanh rule" of turbo decoding).

In [7], the *cut-set bound* of Wiberg [23], [24] was rederived for normal realizations. It follows that for normal realizations on cycle-free graphs, there is a canonical minimal realization that is unique up to isomorphism; furthermore, for general graphs, the cut-set bound constrains the product of the sizes of state spaces in a cut set. The three realizations of Figure 1 meet all possible cut-set bounds with equality.

In [8], similar cut-set bounds are developed for constraint codes in cycle-free realizations, rather than for state spaces. It is shown that the minimum maximum constraint code size can always be achieved by a “cubic" realization in which no constraint code (vertex) has degree greater than 3. Figure 1(b) is an example of such a “cubic" realization in which the maximum constraint code size is 2^9, which is believed to be the best possible for the Golay code, and in addition the maximum state space size is only 2^6.

More recently, [11] studied the system-theoretic properties of trimness, properness, observability and controllability of linear and group realizations. *Trimness* means that in any constraint code, each state alphabet consists only of the values that occur in some allowable configuration, and thus is obviously desirable. *Properness* means that in any constraint code, knowledge of all input variables and all but one state variable determines the final state variable; it has been much less appreciated. It is shown that trimness and properness are dual properties, so both should be valued equally. Moreover, it is shown that if any constraint code is not trim and proper, then the realization is reducible. Finally, a finite cycle-free realization is minimal if and only if every constraint code is trim and proper.

A realization $\mathcal{R}$ is called *observable* if there is precisely one configuration in $\mathfrak{B}$ for each codeword in $\mathcal{C}$, which seems obviously desirable. $\mathcal{R}$ is called *controllable* if it has independent constraints; *e.g.,* in an LDPC realization, if the parity checks are independent. Controllability is not so obviously desirable; indeed, practical LDPC codes sometimes use redundant parity checks. Nonetheless, it is shown in [11] that a linear or group realization $\mathcal{R}$ is observable if and only if its dual realization $\mathcal{R}^\circ$ is controllable.

A trim and proper (*i.e.,* minimal) cycle-free realization is observable and controllable; thus unobservable or uncontrollable behavior must be supported on cycles in a trim and proper normal realization. Any unobservable or uncontrollable realization may be locally reduced, eventually to a trim, proper, observable and controllable realization.

In [9], many of these results are simplified and generalized. It is shown that the single external state space of a trim and proper leaf fragment is uniquely determined, up to isomorphism. This result leads to a simple proof of the "minimal ⇔ trim + proper" theorem of [11]. It is also observed that a cyclic realization may be partitioned into a number of cycle-free leaf fragments, which act as static interface nodes, and a leafless cyclic *2-core*, which is the essential dynamical core of the realization, as illustrated in Figure 2. A new proof is given of the result of [11] that any unobservability or uncontrollability of a trim and proper realization $\mathcal{R}$ must reside within its 2-core.

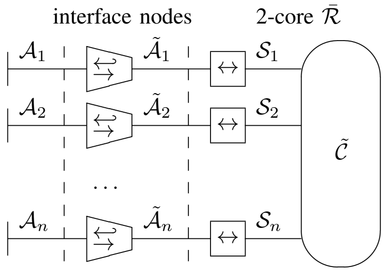

Fig. 2. Partition of a graphical realization $\mathcal{R}$ into its 2-core $\bar{\mathcal{R}}$ and n cycle-free leaf fragments.

Very recently, Conti and Boston [3] have given a simpler and more insightful proof of the Koetter-Vardy Factorization Theorem for linear tail-biting trellis realizations. In [10], this result has been generalized to group tail-biting trellis realizations, and it has been shown that there is unique factorization into "controller granules" as in [12], except that the granules arise from the behavior $\mathfrak{B}$ rather than the code $\mathcal{C}$.

IV. Future work

An immediate objective is to dualize the results of [10], using a nice dual observer granule decomposition. However, as discussed in [13], such a dualization is not straightforward, even for minimal conventional trellis realizations.

It would also be nice to extend the results of [10] to non-trellis realizations. However, it is known (see [11, Appendix A]) that unique factorization does not generally hold for non-trellis linear and group realizations, even simple cycle-free realizations. New ideas will therefore be needed.

Finally, an ultimate goal for our research is to redevelop all of the principal results of classical discrete-time linear systems theory using a purely group-theoretic approach. However, the classical theory generally assumes an infinite time axis. A possible approach would be to regard a time-invariant or periodically time-varying linear or group system on an infinite time axis as the "limit" of a sequence of covers of a linear or group tail-biting trellis realization on a sequence of finite time axes of increasing length. Such an approach would hopefully be purely algebraic, and thus avoid the subtle topological issues discussed in [13].

Acknowledgments

I have greatly enjoyed collaborations with many colleagues in this field, particularly D. Conti, H. Gluesing-Luerssen, R. Koetter, F. R. Kschischang, H.-A. Loeliger, Y. Mao, M. D. Trott, A. Vardy, and P. O. Vontobel.

References

[1] A. R. Calderbank, G. D. Forney, Jr. and A. Vardy, "Minimal tail-biting trellises: The Golay code and more," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1435–1455, July 1999.

[2] A. R. Calderbank and N. J. A. Sloane, "New trellis codes based on lattices and cosets," *IEEE Trans. Inf. Theory*, vol. IT-33, pp. 177–195, 1987.

[3] D. Conti and N. Boston, "On the algebraic structure of linear trellises," submitted to *IEEE Trans. Inf. Theory*, Jan. 2014. ArXiv: 1402.6404.

[4] G. D. Forney, Jr., "Convolutional codes I: Algebraic structure," *IEEE Trans. Inf. Theory*, vol. IT-16, pp. 720–738, Nov. 1970.

[5] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, March 1973.

[6] G. D. Forney, Jr., "Coset codes— Part I: Introduction and geometrical classification," *IEEE Trans. Inf. Theory*, vol. 34, pp. 1123–1151, Sept. 1988

[7] G. D. Forney, Jr., "Codes on graphs: Normal realizations," *IEEE Trans. Inf. Theory*, vol. 47, pp. 520–548, Feb. 2001.

[8] G. D. Forney, Jr., "Codes on graphs: Constraint complexity of cycle-free realizations of linear codes," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1597–1610, Jul. 2003.

[9] G. D. Forney, Jr., "Codes on graphs: Fundamentals," *IEEE Trans. Inf. Theory*, vol. 60, pp. 5809–5826, Oct. 2014.

[10] G. D. Forney, Jr., "Unique factorization and controllability of tail-biting trellis realizations via controller granule decompositions," submitted to *2015 IEEE Inf. Theory Workshop, Jerusalem*, Oct. 2014.

[11] G. D. Forney, Jr. and H. Gluesing-Luerssen, "Codes on graphs: Observability, controllability and local reducibility," *IEEE Trans. Inf. Theory*, vol. 59, pp. 223–238, Jan. 2013.

[12] G. D. Forney, Jr. and M. D. Trott, "The dynamics of group codes: State spaces, trellis diagrams and canonical encoders," *IEEE Trans. Inf. Theory*, vol. 39, pp. 1491–1513, Sept. 1993.

[13] G. D. Forney, Jr. and M. D. Trott, "The dynamics of group codes: Dual abelian group codes and systems," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2935–2965, Dec. 2004.

[14] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. IT-8, pp. 21–28, Jan. 1962.

[15] H. Gluesing-Luerssen and G. D. Forney, Jr., "Local irreducibility of tail-biting trellises," *IEEE Trans. Inf. Theory*, vol. 59, pp. 6597–6610, Oct. 2013.

[16] R. Koetter and A. Vardy, "The structure of tail-biting trellises: Minimality and basic principles," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2081–2105, Sept. 2003.

[17] F. R. Kschischang, B. J. Frey and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, pp. 498–519, Feb. 2001.

[18] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Sig. Proc. Mag.*, vol. 21, pp. 28–41, Jan. 2004.

[19] D. J. Muder, "Minimal trellises for block codes," *IEEE Trans. Inf. Theory*, vol. 34, pp. 1049–1053, Sept. 1988.

[20] R. M. Tanner, "A recursive approach to low-complexity codes," *IEEE Trans. Inf. Theory*, vol. IT–27, pp. 533–547, Sept. 1981.

[21] A. Vardy, "Trellis structure of codes," in *Handbook of Coding Theory* (V. S. Pless and W. C. Huffman, eds.). Amsterdam, The Netherlands: Elsevier, 1999.

[22] A. Vardy and Y. Be'ery, "More efficient soft-decision decoding of the Golay codes," *IEEE Trans. Inf. Theory*, vol. 37, pp. 667–672, May 1991.

[23] N. Wiberg, "Codes and decoding on general graphs," Ph.D. dissertation, Univ. Linköping, Linköping, Sweden, 1996.

[24] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *Euro. Trans. Telecomm.*, vol. 6, pp. 513–525, Sept./Oct. 1995.

[25] J. C. Willems, "Models for dynamics," in *Dynamics Reported* (U. Kirchgraber and H. O. Walther, eds.), vol. 2, pp. 171–269. New York: Wiley, 1989.

Positivity, Discontinuity and Finite Resources for Arbitrarily Varying Quantum Channels

Holger Boche
TU München

Abstract

This work is motivated by a quite general question: Under which circumstances are the capacities of information transmission systems continuous? The research is explicitly carried out on finite arbitrarily varying quantum channels (AVQCs).

We give an explicit example that answers the recent question whether the transmission of messages over AVQCs can benefit from assistance by distribution of randomness between the legitimate sender and receiver in the affimative. The specific class of channels introduced in that example is then extended to show that the unassisted capacity does have discontinuity points, while it is known that the randomness-assisted capacity is always continuous in the channel. We characterize the discontinuity points and prove that the unassisted capacity is always continuous around its positivity points.

After having established shared randomness as an important resource, we quantify the interplay between the distribution of finite amounts of randomness between the legitimate sender and receiver, the (nonzero) probability of a decoding error with respect to the average error criterion and the number of messages that can be sent over a finite number of channel uses.

We relate our results to the entanglement transmission capacities of finite AVQCs, where the role of shared randomness is not yet well understood, and give a new sufficient criterion for the entanglement transmission capacity with randomness assistance to vanish.

On Maximum Rényi Entropy Rate

Christoph Bunte and Amos Lapidoth
ETH Zurich
Switzerland
Email: {bunte,lapidoth}@isi.ee.ethz.ch

***Abstract*—We compute the supremum of the Rényi entropy rate over the class of stationary stochastic processes having autocovariance sequences that begin with $p+1$ given values. Our results are closely related to Burg's maximum entropy theorem on the supremum over the same class but of the Shannon entropy rate.**

I. Introduction

Motivated by spectral estimation, Burg found the maximum of the differential Shannon entropy rate over the class of stationary stochastic processes whose autocovariance sequences begin with $p+1$ given values [1], [2, Theorem 12.6.1]. Here we consider the same class, but we maximize a different objective function: the Rényi entropy rate.

To recall the definition of the Rényi rate of a stochastic process, we begin with the Rényi entropy of a random vector or of a (joint) density. The order-α Rényi entropy of a probability density function (PDF) f is defined as

$$h_\alpha(f) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} f(x)^\alpha \,\mathrm{d}x, \tag{1}$$

where α can be any positive number other than one. The integral on the RHS of (1) always exists, possibly taking on the value $+\infty$, in which case we define $h_\alpha(f) = +\infty$ if $0 < \alpha < 1$ and $h_\alpha(f) = -\infty$ if $\alpha > 1$. When a random variable (or random vector) X is of density f_X we sometimes write $h_\alpha(X)$ instead of $h_\alpha(f_X)$.

The order-α Rényi entropy rate (or "Rényi rate" for short) of a stochastic process (SP) $\{X_k\}$ is defined as

$$h_\alpha(\{X_k\}) = \lim_{n\to\infty} \frac{1}{n} h_\alpha\bigl(X_1^n\bigr),$$

whenever the limit exists. Here we use the notation X_i^j to denote the tuple $(X_i, \dots, X_j)$.

The Rényi entropy rate of finite-state Markov chains was computed by Rached, Alajaji, and Campbell [3] with extensions to countable state space in [4].[1] The Rényi entropy rate of stationary Gaussian processes was found by Golshani and Pasha in [5]. Extensions to other types of rate are explored in [6].

The Rényi entropy is closely related to the differential Shannon entropy:

$$h(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) \,\mathrm{d}x. \tag{2}$$

(The integral on the RHS of (2) need not exist. If it does not, then we say that $h(f)$ does not exist.) Under some mild technical conditions [7],

$$h_\alpha(f) \le h(f), \qquad \text{for } \alpha > 1; \tag{3}$$
$$h_\alpha(f) \ge h(f), \qquad \text{for } 0 < \alpha < 1; \tag{4}$$

and

$$\lim_{\alpha\to 1} h_\alpha(f) = h(f). \tag{5}$$

The entropy of a pair of independent random variables is the sum of the individual entropies. This is true for both differential Shannon entropy and Rényi entropy. But the two entropies behave differently when the random variables are dependent. While the differential Shannon entropy of a pair is always upper-bounded by the sum of the individual entropies, this need not hold for Rényi entropy: the Rényi entropy of a random vector can exceed the sum of the Rényi entropies of its components. Consequently, the random vector of highest Rényi entropy among all those whose components have some prespecified distribution need not have independent components. This is, of course, also true if the distributions of the components are not specified but only constrained.[2] Likewise, the supremum of the Rényi *rate* subject to constraints on the marginal distribution is not achieved by memoryless processes [8].

Here we focus on the supremum of the Rényi rate subject to autocovariance constraints. We show that the solution exhibits a dichotomy: when the order α is smaller than one, the supremum is infinite; and when it is greater than one the supremum is the same as if we were maximizing the Shannon rate (with the supremum thus being computable using Burg's theorem). Note, however, that the supremum—unlike the supremum in Burg's theorem—is not achieved by a Gauss-Markov process. It is, however, approachable by stochastic processes having the same autocovariance sequence as the Gauss-Markov process.

II. Preliminaries

Key to our results is the following proposition [8, Corollary 4]:

Proposition 1 (Rényi Rate under a Variance Constraint)**.**

1) For every $\alpha > 1$, every $\sigma > 0$, and every $\varepsilon > 0$ there exists a centered stationary SP $\{Y_k\}$ whose Rényi entropy

[1] In the discrete case the density in (1) is replaced by the probability mass function, and the integral is replaced by a sum.

[2] Nevertheless, the maximization of Rényi entropy subject to linear constraints does typically have a simple solution [8].

rate exceeds $\frac{1}{2}\log(2\pi e\sigma^2) - \varepsilon$ *and which satisfies*

$$\mathrm{E}[Y_k Y_{k'}] = \sigma^2 \,\mathrm{I}\{k = k'\}, \tag{6}$$

where $\mathrm{I}\{\text{statement}\}$ *is* 1 *when* statement *is true and* 0 *when it is not.*

2) *For every* $0 < \alpha < 1$*, every* $\sigma > 0$*, and every* $\mathsf{M} > 0$ *there exists a centered stationary SP* $\{Y_k\}$ *whose Rényi entropy rate exceeds* M *and which satisfies* (6).

To address some technical boundary issues we shall also need the following lemma.

Lemma 2. *Let* $f_1, \ldots, f_p$ *be probability density functions on* $\mathbb{R}^n$*, and let* $q_1, \ldots, q_p \geq 0$ *be nonnegative numbers that sum to one. Let* f *be the mixture density*

$$f(\mathbf{x}) = \sum_{\ell=1}^{p} q_\ell f_\ell(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n.$$

Then

$$h_\alpha(f) \geq \min_{1\leq \ell \leq p} h_\alpha(f_\ell).$$

Proof. For $0 < \alpha < 1$ this follows by the concavity of Rényi entropy. Consider now $\alpha > 1$:

$$\begin{aligned}
\log \int f^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \log \int \biggl(\sum_{\ell=1}^{p} q_\ell f_\ell(\mathbf{x})\biggr)^{\alpha} \mathrm{d}\mathbf{x} \\
&\leq \log \int \sum_{\ell=1}^{p} q_\ell f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \log \biggl(\sum_{\ell=1}^{p} q_\ell \int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\biggr) \\
&\leq \log \max_{1\leq\ell\leq p} \int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \max_{1\leq\ell\leq p} \log \int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x},
\end{aligned}$$

from which the claim follows because $1/(1-\alpha)$ is negative. □

III. Results

Given $\alpha_0, \ldots, \alpha_p \in \mathbb{R}$, we consider the family of all stochastic processes $X_1, X_2, \ldots$ for which

$$\mathrm{E}[X_i X_{i+k}] = \alpha_k, \quad \bigl(i \in \mathbb{N},\ k \in \{0, \ldots, p\}\bigr). \tag{7}$$

We assume that the $(p+1) \times (p+1)$ matrix whose Row-ℓ Column-m element is $\alpha_{|\ell-m|}$ is positive definite. This implies [9] that there exist constants $a_1, \ldots, a_p, \sigma^2$ and a $p \times p$ positive definite matrix K_p such that the following holds:[3] if the random p-vector $(W_{1-p}, \ldots, W_0)$ is of second-moment matrix K_p (not necessarily centered) and if $\{Z_i\}_{i=1}^\infty$ are independent of $(W_{1-p}, \ldots, W_0)$ with

$$\mathrm{E}[Z_i] = 0, \qquad i \in \mathbb{N}, \tag{8a}$$

$$\mathrm{E}[Z_i Z_j] = \sigma^2 \,\mathrm{I}\{i = j\}, \qquad i, j \in \mathbb{N}, \tag{8b}$$

then the process defined inductively via

$$X_i = \sum_{k=1}^{p} a_i X_{i-k} + Z_i, \quad i \in \mathbb{N} \tag{9}$$

with the initialization

$$(X_{1-p}, \ldots, X_0) = (W_{1-p}, \ldots, W_0) \tag{10}$$

satisfies the constraints (7).

By Burg's maximum entropy theorem [2, Theorem 12.6.1], of all stochastic processes satisfying (7) the one of highest (differential) Shannon entropy rate is the p-th order Gauss-Markov process. It is obtained when $(W_{1-p}, \ldots, W_0)$ is a centered Gaussian and $\{Z_i\}$ are IID $\sim \mathcal{N}(0, \sigma^2)$. Its Shannon entropy rate is

$$\lim_{n\to\infty} \frac{1}{n} h(X_1, \ldots, X_n) = \frac{1}{2}\log(2\pi e\sigma^2).$$

Our interest is in the maximum Rényi entropy rate.

Theorem 3. *The supremum of the order-*α *Rényi entropy rate over all stochastic processes satisfying* (7) *is* $+\infty$ *for* $0 < \alpha < 1$ *and is equal to the Shannon entropy rate of the* p*-th order Gauss-Markov process for* $\alpha > 1$.

Proof. We first consider the case where $\alpha > 1$. Let $a_1, \ldots, a_p, \sigma^2$ and K_p be as above, and let $\varepsilon > 0$ be arbitrarily small. By Proposition 1 there exists a stochastic process $\{Z_i\}$ such that (8) holds and such that

$$\lim_{n\to\infty} \frac{1}{n} h_\alpha(Z_1, \ldots, Z_n) \geq \frac{1}{2}\log(2\pi e\sigma^2) - \varepsilon. \tag{11}$$

The matrix K_p is positive definite, so by the spectral representation theorem we can find vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p \in \mathbb{R}^p$ and constants $q_1, \ldots, q_p > 0$ with $q_1 + \cdots + q_p = 1$ such that

$$\mathsf{K}_p = \sum_{\ell=1}^{p} q_\ell \mathbf{w}_\ell \mathbf{w}_\ell^{\mathsf{T}}. \tag{12}$$

(The vectors are eigenvectors of K_p, and the constants $q_1, \ldots, q_p$ are the scaled eigenvalues of K_p.) Draw the random vector $\mathbf{W}$ independently of $\{Z_i\}$ with

$$\Pr[\mathbf{W} = \mathbf{w}_\ell] = q_\ell,$$

so that, by (12),

$$\mathrm{E}[\mathbf{W}\mathbf{W}^{\mathsf{T}}] = \mathsf{K}_p.$$

Construct now the stochastic process $\{X_i\}$ using (9) initialized with $(X_{1-p}, \ldots, X_0)^{\mathsf{T}}$ being set to $\mathbf{W}$.

The resulting stochastic process thus satisfies the constraints (7). We next study its Rényi entropy rate. To that end, we study the Rényi entropy of the vector X_1^n. Let $f_{\mathbf{X}}$ denote its density, and let $f_{\mathbf{X}|\mathbf{w}_\ell}$ denote its conditional density given $\mathbf{W} = \mathbf{w}_\ell$, so

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{\ell=1}^{p} q_\ell f_{\mathbf{X}|\mathbf{w}_\ell}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n.$$

[3] The Row-ℓ Column-m element element of the matrix K_p is $\alpha_{|\ell-m|}$.

Consequently, by Lemma 2,

$$h_\alpha(f_{\mathbf{X}}) \geq \min_{1\leq \ell \leq p} h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell}). \tag{13}$$

We next study $h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell})$ for any given $\ell \in \{1,\ldots,p\}$. Recalling that $\mathbf{W}$ and $\{Z_i\}$ are independent, we conclude that, conditional on $\mathbf{W} = \mathbf{w}_\ell$, the random variables $X_1,\ldots,X_n$ are generated inductively via (9) with the initialization

$$(X_{1-p},\ldots,X_0)^\mathsf{T} = \mathbf{w}_\ell.$$

Conditionally on $\mathbf{W} = \mathbf{w}_\ell$, the random variables $X_1,\ldots,X_n$ are thus an affine transformation of $Z_1,\ldots,Z_n$. The transformation is of unit Jacobian, and thus

$$h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell}) = h_\alpha(Z_1,\ldots,Z_n), \quad \ell \in \{1,\ldots,p\}. \tag{14}$$

From this and (13) it follows that

$$h_\alpha(f_{\mathbf{X}}) \geq h_\alpha(Z_1,\ldots,Z_n).$$

Dividing by n and using (11) establishes the result.

We next turn to the case $0 < \alpha < 1$. For every $\mathsf{M} > 0$ arbitrarily large, we use Proposition 1 to construct $\{Z_i\}$ as above but with

$$\lim_{n\to\infty} \frac{1}{n} h_\alpha(Z_1,\ldots,Z_n) \geq \mathsf{M}.$$

The proof continues as for the case where α exceeds one. □

IV. Discussion

Theorem 3 has bearing on the spectral estimation problem, i.e., the problem of extrapolating the values of the autocovariance sequence from its first $p+1$ values. One approach is to choose the extrapolated sequence to be the autocovariance sequence of the stochastic process that—among all stochastic processes that have an autocovariance sequence that starts with these $p+1$ values—maximizes the Shannon rate, namely the p-th order Gauss-Markov process (Burg's theorem).

A different approach might be to choose some $\alpha > 1$ and to replace the maximization of the Shannon rate with that of the order-α Rényi rate. As we next argue, Theorem 3 shows that this would result in the same extrapolated sequence. Indeed, inspecting the proof of the theorem we see that the stochastic process $\{X_i\}$ that we constructed, while not a Gauss-Markov process, has the same autocovariance sequence as the p-th order Gauss-Markov process that satisfies the constraints. And, for $\alpha > 1$ the supremum can only be achieved by a stochastic process of this autocovariance sequence: for any other autocovariance function the Rényi rate is upper bounded by the Shannon rate (because $\alpha > 1$), and the latter is upper bounded by the Shannon rate of the Gaussian process, which, unless the autocovariance sequence is that of the p-th order Gauss-Markov process, is strictly smaller than the supremum (Burg's theorem).

Acknowledgment

Discussions with Stefan M. Moser are gratefully acknowledged.

References

[1] J. P. Burg, "Maximum entropy spectral analysis,," in *Proc. 37th Meet. Society of Exploration Geophysicists, 1967. Reprinted in* Modern Spectrum Analysis, *D. G. Childers, Ed. New York: IEEE Press, 1978 pp. 34–41.*, 1967.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2006.

[3] Z. Rached, F. Alajaji, and L. Campbell, "Rényi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1553–1561, May 2001.

[4] L. Golshani, E. Pasha, and G. Yari, "Some properties of Rényi entropy and Rényi entropy rate," *Information Sciences*, vol. 179, no. 14, pp. 2426–2433, 2009.

[5] L. Golshani and E. Pasha, "Rényi entropy rate for Gaussian processes," *Information Sciences*, vol. 180, no. 8, pp. 1486–1491, 2010.

[6] M. Khodabin, "ADK entropy and ADK entropy rate in irreducible-aperiodic Markov chain and Gaussian processes," *Journal of the Iranian Statistical Society*, vol. 9, no. 2, pp. 115–126, 2010.

[7] L. Wang and M. Madiman, "Beyond the entropy power inequality, via rearrangements," July 2013, arXiv:1307.6018.

[8] C. Bunte and A. Lapidoth, in *Proc. of the 2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, December 3–5 2014.

[9] M. Pourahmadi, *Foundations of Time Series Analysis and Prediction Theory*, ser. Wiley Series in Probability and Statistics. Wiley, 2001.

Inaedificatio

D. Stotz, E. Riegler, and H. Bölcskei
ETH Zürich

Abstract

We develop a unified framework for understanding the fundamental limits of a wide range of signal reconstruction problems such as image inpainting, super-resolution, signal separation, denoising, and recovery of signals that are impaired by, e.g., clipping, impulse noise, or narrowband interference. An information-theoretic formulation allowing for random signals leads us to an almost lossless analog signal separation problem and reveals Minkowski dimension as the foundational element of the theory. As a byproduct, we discover a new technique for showing that the intersection of generic subspaces with subsets of sufficiently small Minkowski dimension is empty. This result can be viewed as a measure-theoretic version of the null-space property widely used in compressed sensing theory.