International Zurich Seminar on Communications
February 29 – March 2, 2012
Sorell Hotel Zürichberg, Zurich, Switzerland
Proceedings
Acknowledgment of Support

The International Zurich Seminar on Communications (IZS) is organized jointly by the IEEE Switzerland Chapter on Digital Communication Systems and the Department of Information Technology and Electrical Engineering of ETH Zurich.

IEEE

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Conference Organization

General Co-Chairs

Helmut Bölcskei and Amos Lapidoth

Technical Program Committee

- Ezio Biglieri
- Pierre Chevillat
- Bernard Fleury
- Martin Hänggi
- Walter Hirt
- Gerhard Kramer
- Frank Kschischang
- Hans-Andrea Loeliger
- Stefan Mangold
- Thomas Mittelholzer
- Stefan Moser

- Nikolai Nefedov
- Bernhard Plattner
- Igal Sason
- Jossy Sayir
- Giorgio Taricco
- Emre Telatar
- Emanuele Viterbo
- Pascal Vontobel
- Michèle Wigger
- Armin Wittneben
- Raymond Yeung

Organizers of Invited Sessions

- Emmanuel Abbe
- Uri Erez
- Ioannis Kontoyiannis

- Emanuele Viterbo
- Michèle Wigger

Local Organization

Conference Secretaries

- Rita Hildebrand
- Silvia Tempel

Web and Publications

- Michael Lerjen
Table of Contents

Keynote Talks

Wed 09:00 – 10:00
Gerhard Kramer, TU München
Progress on Relaying and Noisy Network Coding

Thu 09:00 – 10:00
Peter Winzer, Alcatel-Lucent
Spatial Multiplexing and MIMO for Fiber Capacities beyond 100 Tb/s

Fri 09:00 – 10:00
Mitchell Trott, HP Labs
Why is Wyner-Ziv coding for video so difficult?

Session 1 Wed 10:20 – 12:00
Invited session organizer: Michèle Wigger

*The Multiple Access Channel with Correlated Sources and Cribbing Encoders . . . 8
E. Amir and Y. Steinberg

*The Source Coding Side of Secrecy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
P. Cuff

*A Comparative Study of Cooperation for the Two-User Interference Channel . . . 12
D. Tuninetti

*Refinement of the Random Coding Bound . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
Y. Altuğ and A. B. Wagner

*Source Coding with Side-Information at the Receivers and an Application . . . . . . 19
M. Wigger

Session 2: Shannon Theory Wed 13:20 – 15:00
Chaired by Michael Gastpar

On Gaussian Channels with MMSE Interference . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
R. Bustin and S. Shamai (Shitz)

On the Gaussian Listening-Helper Source-Coding Problem . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
S. I. Bross and H. Permuter

*Invited papers are marked by an asterisk.
Upper-Bounding Rate Distortion Functions based on the Minimum Mean-Square Error ................................................................. 28
M. Dörpinghaus, M. Rüngeler, and R. Mathar

On the Information Loss in Memoryless Systems: The Multivariate Case ........ 32
B. C. Geiger and G. Kubin

One-Bit Quantizers for Fading Channels ........................................... 36
T. Koch and A. Lapidoth

Session 3  Wed 15:30 – 16:50
Invited session organizer: Ioannis Kontoyiannis

*Model Switching and other Unconventional Fun with Experts ............... 40
S. de Rooij

*Some Old and New Relations between Information and Estimation .......... 43
K. Venkat and T. Weissman

*Polar Martingales of Maximal Spread ............................................ 44
E. Abbe

*Sumset Differential Entropy Bounds .............................................. 48
I. Kontoyiannis

Session 4  Thu 10:20 – 12:00
Invited session organizer: Emmanuel Abbe

*Latest and Greatest on Spatially Coupled Codes
R. Urbanke

*Coding Theory and Compressed Sensing through Convex Relaxations ........ 49
A. Saber Tehrani and A. G. Dimakis

*Hybrid Coding
Y.-H. Kim

*Polar Coding
E. Telatar

*Competitive Classification
A. Orlitsky

Session 5: Coding I  Thu 13:20 – 15:00
Chaired by Hans-Andrea Loeliger

Polar Codes for the Deterministic Broadcast Channel .......................... 51
N. Goela, E. Abbe, and M. Gastpar
Polar Codes for Coordination in Cascade Networks
R. Blasco-Serrano, R. Thobaben, and M. Skoglund

Polar Codes for Degraded Relay Channels
M. Karzand

Unambiguous Decoding of Generalized Reed–Solomon Codes beyond Half the Minimum Distance
A. Zeh, A. Wachter, and M. Bossert

Iteration-Constrained Design of IRA Codes
A. Grant, I. Land, and G. Wang

Session 6: Coding II
Thu 15:30 – 17:10

The Relation between Block Length and Reliability for a Cascade of AWGN Links
R. Subramanian

Secure Multiplex Coding over Interference Channel with Confidential Messages
X. Li and R. Matsumoto

Polymer Expansions for Cycle LDPC Codes
N. Macris and M. Vuffray

Factoring Linear Trellises
D. Conti and N. Boston

Information Theoretic Authentication and Secrecy Codes in the Splitting Model
M. Huber

Session 7
Fri 10:20 – 12:00

Invited session organizer: Uri Erez

*Anti-Structure Problems
R. Zamir

*Real-Time Coding of Markov Sources over Erasure Channels: When is Binning Optimal?
A. Khisti, F. Etezadi, and M. Trott

*Spinal Codes
D. Shah

*SuperNyquist Rateless Coding for Intersymbol Interference Channels
U. Erez and G. W. Wornell

*Successive Compute-and-Forward
B. Nazer
**Session 8**

Invited session organizer: Emanuele Viterbo

*What are Good Coding Schemes for Multicast in Heterogeneous Wireless Networks?*

_E. Soljanin_

*Tunable Sparse Network Coding* ................................................................. 107
_S. Feizi, D. E. Lucani, and M. Médard_

*Coding for Wireless Communications*  
_M. Bossert_

*The Flatness Factor in Lattice Network Coding: Design Criterion and Decoding Algorithm* .......................................................... 111
_J.-C. Belfiore and C. Ling_

*Wideband non-coherent Wireless Communications*  
_E. Viterbo_

**Session 9: Communication Theory**

Chaired by Tobias Koch

Analysis of the Stochastic Channel Model by Saleh & Valenzuela via the Theory of Point Processes ................................................................. 115
_M. L. Jakobsen, T. Pedersen, and B. H. Fleury_

_A. Schenk and R. F. H. Fischer_

_G. Levin and S. Loyka_

Complexity Analysis for ML-based Sphere Decoder Achieving a Vanishing Performance-Gap to Brute-Force ML Decoding .............................................. 127
_A. Singh, P. Elia, and J. Jaldén_

Symbol-by-Symbol APP Decoding based on Supercode Decoding ...................... 131
_J. Freudenberger and U. Kaiser_
The Multiple Access Channel with Correlated Sources and Cribbing Encoders

Eliron Amir and Yossif Steinberg
Dept. of EE, Technion, Israel

I. NOTATION

A discrete memoryless multiple access channel (MAC) is a quadruple \( \{X_1, X_2, P_{Y|X_1, X_2}, Y\} \) where \( X_1, X_2 \) are the input alphabets, \( Y \) the output alphabet, and \( P_{Y|X_1, X_2} \) a transition probability matrix from \( X_1 \times X_2 \) to \( Y \). For shorthand notation, we will refer to the MAC by \( P_{Y|X_1, X_2} \). A source pair is a triplet \( \{U, Y, P_{U|Y}\} \) where \( P_{U|Y} \) is a distribution on \( U \times Y \). We use the following notation: random variables are denoted by uppercase letters. Sequence of \( n \) letters. Sequence of \( n \) letters from alphabet \( \mathcal{X} \) is denoted by \( x^n \), and the substring \( x_i, x_{i+1}, \ldots, x_j \) as \( x^i_j \). When the dimension \( n \) is understood from the context, \( n \)-vectors are denoted by boldface letters: \( x^n = x \). Similar notation holds for random variables and random vectors, e.g., \( V, U, U^n \), etc. The channel and source are assumed memoryless, thus probabilities of \( n \)-sequences are given by

\[
P_{Y|X_1, X_2}(y|x_1, x_2) = \prod_{i=1}^{n} P_{Y|X_1, X_2}(y_i|x_{1,i}, x_{2,i})
\]

\[
P_{U,V}(u, v) = \prod_{i=1}^{n} P_{U,V}(u_i, v_{i+1})
\]

II. PROBLEM DEFINITION AND PREVIOUS RESULTS

We investigate a joint source-channel coding schemes for the discrete memoryless MAC with cribbing encoders. The sources, \( U \) and \( V \), deliver their output to two separate encoders 1 and 2. An \((n, m, \epsilon)\) joint source channel code for the source \( P_{U,V} \) and the MAC \( P_{Y|X_1, X_2} \) with strictly causal cribbing encoder consists of \( m+1 \) encoding functions

\[
f_1 : U^n \rightarrow X^n_1
\]

\[
f_{2,i} : V^n \times X^n_{i-1} \rightarrow X_{2,i} \quad i = 1, \ldots, m
\]

and a decoding function

\[
d^n : Y^n \rightarrow U^n \times V^n
\]

such that the probability of error is bounded from above by \( \epsilon \):

\[
P_e = Pr\{ (u^n, v^n) \neq d^n (y^n) \}
\]

\[
= \sum_{(u, v) \in U^n \times V^n} P_{U,V}(u, v) \cdot Pr\{ d^n (y^n) \neq (u, v) \} \leq \epsilon
\]

We say that encoder 2 is the cribbing encoder, as reflected in (1b). The rate of the code is defined as \( \rho = n/m \). A source pair \((U, V)\) is said to be transmissible via the MAC \( P_{Y|X_1, X_2} \) with strictly causal cribbing at rate \( \rho \) if for every \( \epsilon > 0, \delta > 0 \), and sufficiently large \( n \) there exists an \((n, n/(\rho - \delta), \epsilon)\) code for the source pair and the channel.

We will examine also the model of causal cribbing, where the cribbing encoder sees the output of the other encoder without the delay of (1b), that is, when the sequence of \( m \) encoders at channel input 2 is of the form

\[
f_{2,i} : V^n \times X^n_{i-1} \rightarrow X_{2,i} \quad i = 1, \ldots, m
\]

The definition of transmissibility of a source pair via the channel remains as in the case of strictly causal cribbing, with the only difference that (2) replaces (1b). The joint probability mass function for strictly causal cribbing is given by

\[
P(u, v, x_1, x_2, y) = \prod_{i=1}^{n} P_{U,V}(u_i, v_{i+1}) \prod_{i=1}^{n} P_{Y|X_1, X_2}(y_i|x_{1,i}, x_{2,i}, (v^n, x_{i-1}^{i-1}(u^n)))
\]

and for causal cribbing by

\[
P(u, v, x_1, x_2, y) = \prod_{i=1}^{n} P_{U,V}(u_i, v_{i+1}) \prod_{i=1}^{n} P_{Y|X_1, X_2}(y_i|x_{1,i}, x_{2,i}, (v^n, x_{i}^{i}(u^n)))
\]

We are interested in finding conditions for transmissibility of a source via the MAC with causal and strictly causal cribbing. We start by stating previous relevant results, on capacity regions of MAC with cribbing, and on distributed source coding.

Theorem 1. F. Willems & E. C. van der Meulen [4]: The capacity region for the strictly causal cribbing communication model is the collection of all pairs \((R_1, R_2)\) such that

\[
R_1 \leq H(X_1 | W) \quad (3)
\]

\[
R_2 \leq I(X_2; Y | X_1, W) \quad (4)
\]

\[
R_1 + R_2 \leq I(X_1, X_2; Y) \quad (5)
\]

for some distribution of the form

\[
P_{W,X_1, X_2, Y}(w, x_1, x_2, y) =
\]

\[
P_W(w) P_{X_1|W}(x_1 | w) \cdot P_{X_2|W}(x_2 | w) \cdot P_{Y|X_1, X_2}(y | x_1, x_2) \quad (6)
\]

and \( |W| \leq \min \{|X_1|, |X_2| + 1, |Y| + 2\} \)

The capacity region for the causal cribbing communication model is given by the collection of all rate pairs \((R_1, R_2)\) such that

\[
R_1 \leq H(X_1) \quad (7)
\]

\[
R_1 \leq I(X_2; Y | X_1) \quad (8)
\]

\[
R_1 + R_2 \leq I(X_1, X_2; Y) \quad (9)
\]
for
\[
P_{X_1,X_2,Y}(x_1,x_2,y) = \\
P_{X_1,X_2}(x_1,x_2) P_{Y|X_1,X_2}(y|x_1,x_2)
\]

\[\text{Theorem 2.} \text{ Slepian-Wolf [3]: For the distributed source coding problem with source } (U,V) \text{ drawn i.i.d } \sim \ P_{U,V}, \text{ the achievable rate region is given by:}
\]
\[
H(U|V) \leq R_1 \quad (11) \\
H(V|U) \leq R_2 \quad (12) \\
H(U,V) \leq R_1 + R_2 \quad (13)
\]

\[\text{III. STATEMENT OF RESULTS}
\]
\[\text{Theorem 3. A source } (U,V) \sim P_{U,V} \text{ is transmissible via the MAC } P_{Y|X_1,X_2} \text{ with strictly causal cribbing at rate } \rho = 1 \text{ if}
\]
\[
H(U|V) \leq H(X_1|W,V) \quad (14) \\
H(V|U) \leq I(Y;X_2|W,U,X_1) \quad (15) \\
H(U,V) \leq I(Y;X_1,X_2|U) \quad (16)
\]
\[\text{for some}
\]
\[
P_{W,U,V,X_1,X_2,Y}(w,u,v,x_1,x_2,y) = P_W(w) P_{U,V}(u,v) P_{X_1|W,U}(x_1|w,u) \\
P_{X_2|V,W}(x_2|w,v) P_{Y|X_1,X_2}(y|x_1,x_2)
\]

\[\text{where } P_W(\alpha) = P_U(\alpha) \text{ and } W = U.
\]

The proof of Theorem 3 is omitted due to space considerations.

\[\text{Theorem 4. A source } (U,V) \sim P_{U,V} \text{ is transmissible via the MAC } P_{Y|X_1,X_2} \text{ with causal cribbing at rate } \rho \text{ if and only if there exists a joint input distribution } P_{X_1,X_2} \text{ such that:}
\]
\[
\rho H(U|V) \leq H(X_1) \quad (17) \\
\rho H(V|U) \leq I(Y;X_2|X_1) \quad (18) \\
\rho H(U,V) \leq I(Y;X_1,X_2) \quad (19)
\]

The proof of the converse part of Theorem 4 is omitted. Two proofs of the achievability part of Theorem 4, based on Theorem 3 and based on a separation argument, are given in Section V.

Further results on transmission with distortion, and different models of cribbing, are reported elsewhere.

\[\text{IV. DISCUSSION}
\]
\[\text{As shown in Section V, Theorem 4 implies that a source } P_{U,V} \text{ is transmissible via the MAC } P_{Y|X_1,X_2} \text{ with causal cribbing at rate } \rho \text{ if and only if the Slepian-Wolf region scaled by } \rho \text{ intersects the capacity region of the MAC with causal cribbing. This implies that with causal cribbing, a separation strategy, where we use a Slepian-Wolf code to compress the source and a channel code to transmit the source codewords via the MAC, is optimal. In view of this, let us examine the achievable result for strictly causal cribbing. By Theorem 3, the following inequalities are sufficient conditions for the source } P_{U,V} \text{ to be transmissible via the channel } P_{Y|X_1,X_2} \text{ with strictly causal cribbing at rate } \rho = 1:
\]
\[
H(U|V) \leq H(X_1|W) \quad (20) \\
H(V|U) \leq I(Y;X_2|W,X_1) \quad (21) \\
H(U,V) \leq I(Y;X_1,X_2) \quad (22)
\]

\[\text{for some}
\]
\[
P_{W,U,V,X_1,X_2,Y}(w,u,v,x_1,x_2,y) = P_W(w) P_{U,V}(u,v) P_{X_1|W,X_1}(x_1|w) \\
P_{X_2|X_1,Y}(x_2|w,v) P_{Y|X_1,X_2}(y|x_1,x_2)
\]

\[\text{where } P_W(\alpha) = P_U(\alpha) \text{ and } U = W. \text{ Note that the right hand sides of (20)-(22) do not depend on the source, and coincide with the right hand sides of (3)-(5). The left hand sides of (20)-(22) coincide with the left hand sides of (11)-(13). This observation may suggest that if } P_{U,V} \text{ and } P_{Y|X_1,X_2} \text{ satisfy (20)-(22), then the source can be transmitted via the channel with a separation strategy, namely, by concatenating a Slepian-Wolf code that compresses the source, with a channel code for the MAC with strictly causal cribbing. Unfortunately, this is not the case - conditions (20)-(22) do not imply separation. This is demonstrated in the following example.}
\]

\[\text{Example 1. Consider the joint source } (U,V) \sim \text{ Bernoulli}(1/2), V = U, \text{ and the binary MAC given by:}
\]
\[
P_{Y|X_1,X_2}(y|x_1,x_2) = \begin{cases} 
1 & \text{if } x_1 = x_2 \\
0 & \text{otherwise}
\end{cases}
\]

\[\text{where } I[.] \text{ is the indicator function. Thus, if } x_1 = x_2 \text{ the inputs are connected directly to the output. If } x_1 \neq x_2 \text{ the output is uniformly distributed on } \{0,1\}, \text{ independently of the inputs.}
\]

The Slepian-Wolf region for this source is given by
\[
R_1 + R_2 \geq 1
\]

\[\text{We claim that the capacity region of this MAC with strictly causal cribbing does not contain the line } R_1 + R_2 = 1. \text{ Indeed, by (5)}
\]
\[
R_1 + R_2 \leq I(X_1,X_2;Y) \leq 1 - H(Y|X_1,X_2) \leq 1
\]

\[\text{where the last inequality holds with equality only if } Y \text{ is a deterministic function of } X_1 \text{ and } X_2. \text{ Since by (6) we have the Markov relations } X_1 \rightarrow W \rightarrow X_2, H(Y|X_1,X_2) = 0 \text{ only if } X_1 \text{ and } X_2 \text{ are determined by } W. \text{ But this, in turn, implies that } R_1 = R_2 = 0, \text{ by (3) and (4). Therefore the capacity region of this MAC with strictly causal cribbing does not contain the line } R_1 + R_2 = 1.
\]

\[\text{It is easy to verify that the source } (U,V) \text{ is transmissible via the channel at rate } \rho = 1. \text{ We can employ single letter code: connect } U \text{ and } V \text{ directly to } X_1 \text{ and } X_2, \text{ respectively. Moreover, transmissibility can be verified directly from conditions (20)-(22):}
\]
\[\text{the conditional entropies on the left hand side there are 0, so we can choose } P_W = P_U, X_1 = X_2 = W, \text{ and the transmissibility conditions are satisfied. Hence, (20)-(22) (and therefore also Theorem 3) predict that the source can be transmitted via the MAC with strictly causal cribbing, yet it cannot be transmitted by a separation strategy.}
\]

\[\text{V. PROOF OF THE ACHIEVABILITY PART OF THEOREM 4}
\]

\[\text{Proof: The achievability part of Theorem 4 can be proved based on Theorem 3, or based on a separation argument. We first give a proof based on separation.}
\]

\[\text{Given three nonempty intervals } T_1 = [a,b], T_2 = [c,d], \text{ and } T_3 = [e,f], \text{ a necessary and sufficient condition for the existence of a pair } (R_1,R_2) \text{ such that } R_1 \in T_1, R_2 \in T_2, \text{ and } R_1 + R_2 \in T_3, \text{ is that the endpoints satisfy}
\]
\[
a + c \leq f \quad \text{and} \quad b + d \geq e
\]
If the source $P_{UV}$ and the channel and input distribution $P_{Y|X_1,X_2}$, $P_{X_1,X_2}$ satisfy (17), (18), and (19), then necessarily the intervals

$$I_1 = \left[ \rho \left( U \mid V \right), H(X_1) \right]$$

$$I_2 = \left[ \rho \left( V \mid U \right), I(Y;X_2 \mid X_1) \right]$$

$$I_3 = \left[ \rho \left( U,V \right), I(Y;X_1,X_2) \right]$$

are nonempty. By basic properties of information functions we have

$$\rho(H(U \mid V) + \rho(H(V \mid U) \leq \rho(H(U, V) \leq I(X_1, X_2; Y)$$

$$= I(X_1; Y) + I(X_2; Y \mid X_1) \leq H(X_1) + I(X_2; Y \mid X_1) \leq H(X_1)$$

$$I(1; T; Y \mid X_1, W, U) = I(T; Y \mid X_1)$$

$$= I(X_2; Y \mid X_1) + I(T; Y \mid X_1, X_2) = I(X_2; Y \mid X_1)$$

$$I(T; Y \mid X_1, W, U) = I(X_1, X_2; Y)$$

and

$$P_{X_1,X_2,Y}(x_1, x_2, y) = P_{X_1}(x_1) \cdot \sum_{t : \|x_1\| = t} P_T(t)$$

where for (34) we used the same method presented in [4, (44), (45)], expressing $P_{X_2|X_1}(x_2 \mid x_1)$ as a function of $P_T(t)$:

$$P_{X_2|X_1}(x_2 \mid x_1) = \prod_{x_1} P_{X_2|X_1}(x_2, x_2 = t(x_1))$$

Define a set of conditions as follows:

$$H(U \mid V) \leq H(X_1 \mid W, V)$$

$$H(V \mid U) \leq I(T; Y \mid X_1, W, U)$$

$$H(U, V) \leq I(X_1, T; Y \mid U)$$

for

$$P_{W,U,V,X_1,T,Y}(w, u, v, x_1, t, y) = P_W(w) P_{Y|X_1,W}(x_1 \mid w, u) \cdot P_{T|W,U}(t \mid w, v) P_{Y|X_1,T}(y \mid x_1, t)$$

By Theorem 3, a source $P_{UV}(u, v)$ for which the conditions above hold, is transmissible by the derived MAC with strictly causal cribbing. If we now restrict the distribution in (33) to satisfy

$$P_{W,U,V,X_1,T,Y}(w, u, v, x_1, t, y) = P_W(w) P_{Y|UV}(u, v) P_{X_1}(x_1) \cdot P_T(t) P_{Y|X_1,T}(y \mid x_1, t)$$

then

$$H(X_1 \mid W, V) = H(X_1)$$

$$I(T; Y \mid X_1, W, U) = I(T; Y \mid X_1)$$

$$= I(X_2; Y \mid X_1) + I(T; Y \mid X_1, X_2)$$

$$= I(X_2; Y \mid X_1)$$

$$I(T; Y \mid X_1, W, U) = I(X_1, X_2; Y)$$

and

REFERENCES


The Source Coding Side of Secrecy

Paul Cuff – Princeton University

I. INTRODUCTION
Here we discuss achievability schemes and analysis techniques useful for source coding in adversarial settings. First we mention the equivalence between defining a “valid” joint distribution over all signals in a communication system and defining the behavior of the encoders and decoders. Then we show how to construct a joint distribution in reverse of the operation direction of the encoder, resulting in a communication system that is easy to analyze from the point of view of an optimal adversary. This relates to the work in [1] and [2].

II. CODING DEFINED BY A JOINT DISTRIBUTION
Questions of interest in information theory are usually defined in an operational context, involving encoders, decoders, messages, error probabilities, and distortions. All of these questions could be translated into statement and inquiries involving only probably distributions and their properties. For example, consider the definition of the rate-distortion function.

The standard definition of the rate-distortion function asks for the lowest communication rate such that there exists encoding functions and decoding functions that satisfy certain conditions concerning the range of the encoding function and the resulting average distortion. We can alternatively state an equivalent definition for the rate-distortion function as a question about existence of probability distributions that satisfy certain conditions, as follows.

Rate-distortion Function: Given a source distribution \( Q(x^n) \) and a distortion function \( d(\cdot, \cdot) \), what is the minimum rate \( R \) such that for all \( \epsilon > 0 \) and \( n \) sufficiently large there exists a joint distribution \( p(x^n, m, \hat{x}^n) \) such that the cardinality of \( M \) is less than \( 2^{nR} \), \( X^n - M - \hat{X}^n \) form a Markov chain, \( p(x^n) = Q(x^n) \), and \( E \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) < D + \epsilon \)? Let this be the value of the rate-distortion function \( R(D) \).

The above restatement of the rate-distortion function is trivial and not very helpful. However, in some problems, particularly those involving an adversary, we find this form of the problem statement to be revealing. For one thing, the operation of the encoders and decoders are implicitly specified in the joint distribution of the signals. For example, the encoder in the rate-distortion function statement above operates according to the induced conditional distribution \( p(m|x^n) \). But by changing the emphasis from defining how the encoder and decoder behave to instead defining a joint distribution to satisfy the constraints of the problem, a reverse-encoding construction is easily conceived. That is, we might find it useful to construct the inverse of the encoder, \( p(x^n|m) \), to have certain properties rather than design the actual working of the encoder explicitly.

III. REVERSE-CHANNEL ENCODER
In [1] we characterize the best lossless source coder that makes use of limited secret key to cause maximal distortion to an eavesdropper. An asymptotically optimal construction for the encoder is derived using two key concepts, discussed in the following two paragraphs. The communication is split into two parts, one message which has the secret key applied to it as a one-time pad, and the other non-secure message which is the crux of the analysis. The eavesdropper obtains the non-secure message, and we must be able to analyze the lowest distortion the eavesdropper can achieve given the received message.

The first key concept for constructing an optimal encoder is to design the encoder in reverse. Let \( M \) be the non-secure message obtained by the eavesdropper, and let \( X^n \) be the source sequence. The eavesdropper’s minimum average distortion when a particular message \( m \) is observed depends on the distribution \( p(x^n|m) \), which is the inverse of the encoder. Consider the following design. Using the standard technique of random codebook construction, let \( U \) be an auxiliary random variable correlated with \( X \) and construct a codebook of i.i.d. \( U^n \) sequences indexed by \( M \). Let \( X^n \) be conditionally distributed as the output of a memoryless channel \( \sim p(x|u) \) when the input to the channel is \( U^n(M) \). The memoryless nature of this constructed conditional distribution \( p(x^n|m) \) makes it trivial to characterize the minimum distortion incurred by the eavesdropper.

The last concept needed for the analysis is that of solvability. In the construction above, if \( M \) is uniformly distributed over a set of size \( 2^{nR} \), with \( R > I(X;U) \), then the resulting distribution \( p(x^n) \) will be close to the i.i.d. source distribution \( Q(x^n) \) in total variation. Consequently, the constructed joint distribution is close in total variation to a valid joint distribution. We close this gap with some lemmas about total variation distance.

REFERENCES
A Comparative Study of Cooperation for the Two-user Interference Channel

Daniela Tuninetti
University of Illinois at Chicago
Department of Electrical and Computer Engineering
851 S. Morgan St. (M/C 154)
Chicago, IL 60607, USA
Email: danielat@uic.edu

Abstract—The interference channel is a model for a wireless network where several source-destination pairs compete for the same resources. When the sources transmit simultaneously the destinations experience unwanted interference. Determining how to optimally manage the interference is a long standing open problem in network information theory. In this talk we review one emerging interference management technique: cooperation. Cooperation can occur among source nodes only (Source Cooperation or Generalized Feedback), or among destination nodes (Destination Cooperation), or among all nodes (General Cooperation). For the special case of Gaussian noise channels, we compare the different forms of cooperation in terms of symmetric generalized degrees of freedom. When known, we also discuss capacity results to within a constant number of bits.

I. CHANNEL MODEL

A general full-duplex cooperative interference channel is a 2K-node network with inputs $X_0, \ldots, X_{2K-1}$, outputs $Y_0, \ldots, Y_{2K-1}$ and a memoryless channel with transition probability $P := P[Y_0, \ldots, Y_{2K-1}|X_0, \ldots, X_{2K-1}]$. Sources have index 0 to $K-1$ and destinations $K$ to $2K-1$. Message $W_i$ is independent of everything else and uniformly distributed over $[1 : 2^{NR_i}]$, with $N \in \mathbb{N}$ denoting the block-length and $R_i \in \mathbb{R}^+$ the rate in bits per channel use, for $i \in [0 : K-1]$. Encoding at node $i \in [0 : 2K-1]$ at channel use $t \in [1 : N]$ is a deterministic function $X_{it} = X_{it}^{(t-1)}$ with the convention $W_K = \ldots = W_{2K-1} = \emptyset$. Decoding at node $i \in [K : 2K-1]$ after the $N$-th channel use is a deterministic function $W_{i-K}(Y_i^N)$. A rate vector $(R_0, \ldots, R_{K-1})$ is $\epsilon$-achievable if $\max_{i \in [0 : K-1]} P[W_i \neq W_i] \leq \epsilon$ for some $\epsilon \in (0,1]$. The capacity region is the convex closure of the set of rate vectors that are $\epsilon$-achievable for all $\epsilon > 0 [1]$. In this work we consider the case $K = 2$ only, which is depicted in Fig. 1.

II. MODELS OF COOPERATION

Several cooperation models have been analyzed in literature such as: cooperation only occurs among sources (Source Cooperation, also known as Generalized Feedback), or only among destinations (Destination Cooperation), or among all nodes (General Cooperation). In each case, cooperation can be in-band or out-of-band. For out-of-band cooperation, the network is effectively composed of two parallel networks: the underlying interference channel (from the source inputs to the destination outputs) and the cooperation channel (which is usually assumed to be deterministic/noiseless).

In the following we list some recent journal papers that deal with cooperation for the two-user interference channel and that we will compare in this talk. The interested reader can find the technical details in these works and the list of related preceding publications in the references therein. Some extensions to networks with more than two source-destinations pairs can be found for example in [14], [15].

A. Source Cooperation/Generalized Feedback

The destinations do not have an input to the channel, i.e.,

$$X_K = \ldots = X_{2K-1} = \emptyset.$$

A.1) Out-of-band. The channel input at source $i$, $i \in [0 : K-1]$, is a two-dimensional vector $X_i := [X_{i,d}, X_{i,s}]$, where $X_{i,d}$ affects the channel outputs at the destinations and $X_{i,s}$ the channel outputs at the sources. Effectively, there are two parallel channels: the underlying interference channel with transition probability $P[Y_{K}, \ldots, Y_{2K-1}|X_{0,d}, \ldots, X_{K-1,d}]$ and the cooperation channel with transition probability $P[Y_0, \ldots, Y_{K-1}|X_{0,s}, \ldots, X_{K-1,s}]$.

![Fig. 1. The memoryless cooperative interference channel (channel inputs in green, channel outputs in pink, transmitted messages in black, and decoded messages in blue).](image-url)
Conferencing Encoders is a special case of out-of-band source cooperation where the cooperation channel is deterministic [2].

Cognition is a special case of conferencing encoders where the channel output at some sources is constant (these sources are known as “primary users”) and the channel output at the other sources is such that all messages can be conveyed in a single channel use (these sources are known as the “cognitive users”) [4], [5].

A.2) In-band. The same inputs affect both the underlying interference channel and the cooperation channel. The cooperation channel is noisy [6], [7], [8], [9].

Noisy Output Feedback is a special case of in-band source cooperation where the channel output at a source is a degraded version of the channel output at its intended destination, i.e., \( Y_i = g_i(Y_{i+K}, Z_i) \) for some deterministic function \( g_i \) and some random process \( Z_i \) representing the “extra noise”, for all \( i \in [0 : K - 1] \) [11]. If the “extra noise” process is deterministic then the channel is known simply as interference channel with feedback [12].

Ultimate Limit of Source Cooperation. By sharing all messages among the sources we obtain an equivalent memoryless broadcast channel with input \( X_{	ext{eq}} := (X_0, \ldots, X_{K-1}) \) and outputs \( Y_K, \ldots, Y_{2K-1} \), whose capacity (which is not known in general [1]) gives the ultimate gain that can be obtained with source cooperation.

B. Destination Cooperation

The sources do not have an output from the channel, i.e.,

\[ Y_0 = \ldots = Y_{K-1} = 0. \]

B.1) Out-of-band. The channel output at destination \( i \), \( i \in [K : 2K - 1] \), is a two-dimensional vector \( Y_i := [Y_{i,d}, Y_{i,a}] \), where \( Y_{i,d} \) is affected by the channel inputs from the destinations and \( Y_{i,a} \) by the channel inputs from the sources. Effectively, there are two parallel channels: the underlying interference channel with transition probability \( P[Y_{0,s}, \ldots, Y_{K-1,s}|X_0, \ldots, X_{K-1}] \) and the cooperation channel with transition probability \( P[Y_{0,d}, \ldots, Y_{K-1,d}|X_K, \ldots, X_{2K-1}] \).

Conferencing Decoders is a special case of out-of-band destination cooperation where the cooperation channel is deterministic [3].

B.2) In-band. The channel outputs are affected by all channel inputs [6], [10].

Ultimate Limit of Destination Cooperation. By sharing all channel outputs at the destinations among the destinations we obtain an equivalent memoryless multiple access channel with inputs \( X_0, \ldots, X_{K-1} \) and output \( Y_{	ext{eq}} := (Y_K, \ldots, Y_{2K-1}) \), whose capacity region [1] gives the ultimate gain that can be obtained with destination cooperation.

C. General Cooperation

This is the most general case where all nodes cooperate.

C.1) Out-of-band. Channel outputs at the destinations are affected by the channel inputs from the sources only and channel outputs at the sources are affected by the channel inputs from the destinations only. Effectively, there are two parallel channels: the underlying interference channel with transition probability \( P[Y_K, \ldots, Y_{2K-1}|X_0, \ldots, X_{K-1}] \) and a cooperation channel with transition probability \( P[Y_0, \ldots, Y_{K-1}|X_K, \ldots, X_{2K-1}] \).

Rate-limited Feedback is a special case of out-of-band general cooperation where the cooperation channel is deterministic [13].

C.2) In-band. All channel inputs affect all channel outputs.

To the best of the author’s knowledge the case of in-band general cooperation has not been studied so far.

Ultimate Limit of General Cooperation. By sharing all messages among the sources and all channel outputs at the destinations among the destinations we obtain an equivalent memoryless point-to-point channel with input \( X_{	ext{eq}} := (X_0, \ldots, X_{K-1}) \) and output \( Y_{	ext{eq}} := (Y_K, \ldots, Y_{2K-1}) \), whose capacity gives the following sum-rate upper bound

\[ R_0 + \ldots + R_{K-1} \leq C_{\text{perfect coop}} := \max_{P[X_{	ext{eq}}]} I(X_{	ext{eq}}; Y_{	ext{eq}}). \]

which cannot be improved by feedback, i.e., \( C_{\text{perfect coop}} \) does not depend on \( Y_0, \ldots, Y_{K-1} \).

ACKNOWLEDGMENT

The author’s work on the Interference Channel with Generalized Feedback [7], [8] is part of Dr. S. Yang’s doctoral dissertation, while that on the Cognitive Interference Channel [4], [5] is part or Dr. S. Rini’s doctoral dissertation and was done in collaboration with Dr. N. Devroye.

REFERENCES


Refinement of the Random Coding Bound

Yücel Altuğ Student Member and Aaron B. Wagner Member

Abstract—We provide an improved random coding bound for a class of discrete memoryless channels that improves the pre-factor in front of the exponent. Specifically, while the tightest known bounds have a constant pre-factor, the presented bound has a pre-factor of order $O(1/\sqrt{N})$, where $N$ is the blocklength.

I. INTRODUCTION

Characterizing the interplay between the data rate, error probability, and blocklength of the best codes for a given discrete memoryless channel (D.M.C.) is a central problem in information theory. Although this interplay has been investigated starting from the early days of the field [1], [2], [3], [4], [5], [6], [7], [8], [9], it is still an active research topic [10], [11], [12], [13], [14]. Due to the difficulty of the problem, it is customary to consider formulations that are asymptotic in the blocklength to characterize this interplay.

One of the best-known formulations is “error exponents,” in which the rate is held fixed below the channel capacity and one characterizes the rate of decay of the error probability, which is exponential in $N$ for most channels [3], [4], [5], [6]. Another asymptotic characterization is the “normal approximation” in which the error probability is fixed at a positive constant in $(0,1/2)$ and one characterizes the speed with which the data rate converges to capacity [2], [13]. Between these two characterizations, one can consider the setup in which the error probability tends to zero and the rate simultaneously approaches capacity, with the goal of characterizing the speed of one convergence as a function of the other [14]. We call these three asymptotics the “small error probability,” “large error probability,” and “medium error probability” regimes, respectively. It should be noted that although there is an underlying relation between these three asymptotics via the minimum error probability of the $(N,R)$ codes, it is not possible to directly derive one regime’s conclusion from any of the other two.

In this paper, we consider upper bounds on the error probability of the best $(N,R)$ code in the small error probability regime. We prove a refinement of the random coding bound that has a pre-factor of $O(1/\sqrt{N})$ for a broad class of D.M.C.s, which includes positive channels with positive dispersion. This improves the best known pre-factor of $O(1)$ due to Fano [3] and Gallager [4]. It should be remarked that the improvement in the pre-factor is significant for rates close to the capacity, where the optimal error exponent itself is small and the pre-factors are therefore important. Although some improved finite-$N$ bounds could easily be extracted from the proofs in this paper, we defer the task of optimizing these bounds and numerically comparing them to the existing bounds for future work.

To better understand the asymptotic regimes mentioned above, noting the analogy between channel coding and the analysis of sums of i.i.d. random variables in probability theory is instructive. In particular, the small, medium, and large error probability regimes of channel coding correspond to large deviations, moderate deviations, and central limit theory, respectively, for i.i.d. sums. With this analogy in mind, the problem we consider in this paper is reminiscent of the “exact asymptotics” problem [18], [21, Theorem 3.7.4] in large deviations. The aim of that problem is to determine the pre-factor of the exponentially vanishing term in the large deviations theorem. Bahadur and Rao [18] characterized this pre-factor, including the constant, under some regularity conditions. Their result gives some hope that it may be possible to determine the optimal pre-factor, or at least its order, for channel coding in the small error probability regime. The current best pre-factors in the sphere packing bound [22] and random coding bound (this paper) do not coincide, however, and the order of the optimal pre-factor is only known for special cases such as the binary symmetric [15], binary erasure [15], and power-constrained A.W.G.N. [16] channels.

II. NOTATION, DEFINITIONS AND STATEMENT OF THE RESULT

A. Notation

Boldface letters denote vectors, regular letters with subscripts denote individual elements of vectors. Furthermore, capital letters represent random variables and lowercase letters denote individual realizations of the corresponding random variable. Throughout the paper, all logarithms are base-$e$. For a finite set $X$, $P(X)$ denotes the set of all probability measures on $X$. Similarly, for two finite sets $X$ and $Y$, $P(Y|X)$ denotes the set of all stochastic matrices from $X$ to $Y$. $\mathbb{R}$, $\mathbb{R}^+$ and $\mathbb{Z}^+$ denote the set of real, positive real and non-negative real numbers, respectively. $\mathbb{Z}$ denotes the set of positive integers. We follow the notation of Csiszár-Körner [9] for fundamental information theoretic quantities.

B. Definitions

Throughout the paper, let $W$ be a D.M.C. from $X$ to $Y$, $E_r(R,W)$ and $E_o(\rho,Q,W)$ denotes the random coding exponent (cf. [7, eq. (5.6.16)]) and the Gallager’s function (cf. [7, eq. (5.6.14)]), respectively. For any $W \in P(Y|X)$ and $Q \in P(X)$ the ensemble average error probability (resp. ensemble average error probability conditioned on the
message \( m \) of a random code with codewords generated by using \( Q \) along with a maximum likelihood decoder\(^1\) is denoted by \( \hat{P}_c(Q) \). \( \sigma^2(W) \) denotes the dispersion (cf. [2], [13]) of \( W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) \).

Further,
\[
\mathcal{S}_{Q,W} := \{(x, y) \in \mathcal{X} \times \mathcal{Y} : Q(x)W(y|x) > 0\},
\]
\[
\tilde{\mathcal{S}}_{Q,W} := \{(x, y, z) : Q(x)W(y|x)Q(z)W(y|z) > 0\},
\]
\[
\mathcal{X}_c(W) := \{x \in \mathcal{X} : W(y|x) > 0\}.
\]

Given a \((Q,W) \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y}|\mathcal{X})\) pair, the following property amounts to saying that “feasibility decoding is optimal” (F.D.I.O.) when a random code with distribution \( Q \) is used for transmission over the D.M.C. \( W \).

**Property 1**: (F.D.I.O.) For all \((x, y, z) \in \tilde{\mathcal{S}}_{Q,W}, W(y|x) = W(y|z)\).

**C. Statement of the result**

**Theorem 1**: Let \( W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) \) be arbitrary with \( \sigma^2(W) > 0 \) and \( R \in [R_{\text{cl}}(W), C(W)] \). If there exists \( Q \in \mathcal{P}(\mathcal{X}) \) achieving \( E_s(R, W) \) such that the F.D.I.O. property does not hold, then for all \( N \in \mathbb{Z}^+ \)
\[
\hat{P}_c(Q) \leq \frac{K_1}{\sqrt{N}} e^{-NE_s(R,W)},
\]
where \( K_1 \) is a positive constant that only depends on \( W \) and \( R \).

**Remark 1**: Using the standard expurgation arguments, Theorem 1 ensures the existence of an \((N,R)\) code with maximal error probability upper bounded by the right side of (4), with a different constant.

**Remark 2**: The proof of Theorem 1 is based on Fano’s proof of the random coding bound [3, pp. 324–331] with a refinement to prove the smaller pre-factor. In the literature, Fano’s argument has been superceded by Gallager’s, which is simpler and gives the same exponent. Fano’s argument, however, lends itself more readily to obtaining improved pre-factors.

**Remark 3**: Note that the assumption related to the F.D.I.O. property in Theorem 1 holds for a reasonably large class of channels, such as the class of positive channels with positive dispersion. Yet this assumption does not exclude only degenerate cases; note that the binary erasure channel (B.E.C.) is excluded. Moreover, the assumption is not necessary in order to have an upper bound on the error probability with a pre-factor of \( O(1/\sqrt{N}) \). Indeed, Elias [15] proves such an upper bound for B.E.C.

**Remark 4**: The above theorem applies to the ensemble average error probability at rates above the critical rate. For rates below the critical rate, the order of the pre-factor turns out to hinge on the F.D.I.O property, a fact that was heretofore overlooked in the literature. The next result, Theorem 2, corrects a small oversight by Gallager [17], who claims that for any D.M.C. \( W \) and rate \( R < R_{\text{cl}}(W) \)
\[
\min_{Q \in \mathcal{P}(\mathcal{X})} \hat{P}_c(Q) \sim \frac{g}{\sqrt{N}} e^{-NE_s(R,W)},
\]
for some constant \( g \). In fact, his argument is valid only for those \( W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) \) such that there exists \( Q \in \mathcal{P}(\mathcal{X}) \) with \( E_s(1,Q,W) = \max_{P \in \mathcal{P}(\mathcal{X})} E_s(1,P,W) \) and the F.D.I.O. property is not true. For the remaining channels, we have the following result.

**Theorem 2**: Consider an arbitrary \( W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) \) with \( C(W) > 0 \) and \( R < R_{\text{cl}}(W) \). If the F.D.I.O. property holds for all \( Q \in \mathcal{P}(\mathcal{X}) \) with \( E_s(1,Q,W) = \max_{P \in \mathcal{P}(\mathcal{X})} E_s(1,P,W) \), then for all \( N \in \mathbb{Z}^+ \)
\[
K_2 e^{-NE_s(R,W)} \leq \hat{P}_c(Q) \leq e^{-NE_s(R,W)},
\]
for some \( 0 < K_2 < 1 \) that depends on \( W \) and \( R \).

**III. PROOF OF THEOREM 1**

Let \( W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) \) with \( \sigma^2(W) > 0 \) be arbitrary. Fix an arbitrary \( R \in [R_{\text{cl}}(W), C(W)] \). Consider an arbitrary \( Q \in \mathcal{P}(\mathcal{X}) \) with \( E_s(R,Q,W) = E_s(R,W) \), such that the F.D.I.O. property does not hold, whose existence is guaranteed by the assumption of the theorem. For the sake of notational convenience, let \( \tilde{\mathcal{S}}_Q, \tilde{\mathcal{S}}_Q^r \) and \( \tilde{\mathcal{X}}_c \) denote \( \tilde{\mathcal{S}}_{Q,W}, \tilde{\mathcal{S}}_{Q,W} \) and \( \tilde{\mathcal{X}}_c(W) \) as given in (1), (2) and (3), respectively. Define
\[
P_{X,Y,Z}(x,y,z) := Q(x)W(y|x)Q(z),
\]
for all \((x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \). Also, let
\[
\tilde{P}_{X,Y,Z}(x,y,z) := \begin{cases} \frac{P_{X,Y,Z}(x,y,z)}{P_{X,Y,Z}(\tilde{\mathcal{S}}_Q)} & \text{if } (x,y,z) \in \tilde{\mathcal{S}}_Q \\ 0 & \text{else}. \end{cases}
\]

Define \( P_{X,Y,Z}^N(x,y,z) := \sum_{n=1}^N P_{X,Y,Z}(x_n,y_n,z_n) \) and \( \tilde{S}_Q^N \) (resp. \( \tilde{S}_Q^r \)) denote the \( N \)-fold cartesian product of \( \tilde{S}_Q \) (resp. \( \tilde{S}_Q^r \)). Hence,
\[
P_{X,Y,Z}^N(x,y,z|\tilde{S}_Q^N) = \prod_{n=1}^N \tilde{P}_{X,Y,Z}(x_n,y_n,z_n).
\]

One can show that
\[
\exists! \rho^* \in (0,1), \text{ s.t. } \frac{\partial E_s(\rho,Q,W)}{\partial \rho} \bigg|_{\rho=\rho^*_0(Q,W)} = R.
\]

Define
\[
D_0 := \mathcal{N} \left( \frac{\rho^*}{1 + \rho^*} \right),
\]
\[
\Lambda_1(s,t) := \log E_{\tilde{P}_{X,Y,Z}} \left[ e^{s \log \frac{Q(x)W(y|x)}{W(y|x)} + t \log \frac{Q(z)W(y|z)}{W(y|z)}} \right].
\]
for any \((s, t) \in \mathbb{R}^2 \). Observe that owing to the fact that \( \log \frac{Q(x)W(y|x)}{W(y|x)} \in \mathbb{R} \), for all \((x, y, z) \in \tilde{\mathcal{S}}_Q, \Lambda_1(.) \) is infinitely differentiable on \( \mathbb{R}^2 \).
Finally, for any $N \in \mathbb{Z}^+$, let
\[ D_N := \left\{ \frac{1}{N} \sum_{n=1}^{N} \log \frac{f(Y_n)}{W(Y_n|X_n)} > D_o \right\}. \]

From the standard random coding arguments, one can show that
\[ \bar{P}_{e,1}(Q) \leq P_{X,Y,Z}^{N} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log \frac{f(Y_n)}{W(Y_n|X_n)} \leq D_o, \right. \]
\[ \left. \frac{1}{N} \sum_{n=1}^{N} \log \frac{W(Y_n|X_n)}{W(Y_n|Z_n)} \leq 0 \right\}. \]  
(12)

In order to conclude the proof, we need to upper bound the two terms on the right side of (12).

We begin with the second term. To this end, define
\[ \Lambda^*(a) := \sup_{\lambda \in \mathbb{R}} \{a \lambda - \Lambda(\lambda)\}, \]  
for any $a \in \mathbb{R}$. By noting the convexity of $\Lambda(\cdot)$; (10), (13) and some algebra imply that
\[ \Lambda^*(D_o) = E_1(R, W). \]  
(14)

Owing3 to a result of Ney [19, eq. (2)], we deduce4 that
\[ P_{X,Y}^{N} \{D_N\} \leq \frac{k_1}{\sqrt{N}} e^{-N\Lambda^*(D_o)} = \frac{k_1}{\sqrt{N}} e^{-NE_1(R,W)}, \]  
(15)
for some $k_1 \in \mathbb{R}$ and for all sufficiently large $N$, where the equality follows from (14).

Next, we upper bound the first term on the right side of (12). First, since for any $(x, y, z) \notin \mathcal{S}_Q$, we have
\[ \log \frac{W(y|x)}{W(y|z)} = \infty, \quad Q \times W \times Q - (a.s.) \]
we have
\[ \alpha_N := P_{X,Y,Z}^{N} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log \frac{f(Y_n)}{W(Y_n|X_n)} \leq D_o, \right. \]
\[ \left. \frac{1}{N} \sum_{n=1}^{N} \log \frac{W(Y_n|X_n)}{W(Y_n|Z_n)} \leq 0 \right\} \]
\[ = P_{X,Y,Z}^{N} \left\{ \mathcal{S}_Q \right\}. \]  
(16)

where, in (16) we define
\[ \bar{\alpha}_N := \bar{P}_{X,Y,Z}^{N} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log \frac{f(Y_n)}{W(Y_n|X_n)} \leq D_o, \right. \]
\[ \left. \frac{1}{N} \sum_{n=1}^{N} \log \frac{W(Y_n|X_n)}{W(Y_n|Z_n)} \leq 0 \right\}. \]  
(17)

Define
\[ t^* := -1/(1+\rho^*), \quad s^* := 1 + 2t^*. \]  
(18)

One can show that
\[ \frac{\partial \Lambda_1(s,t)}{\partial s} \bigg|_{s=s^*} = 0, \quad \frac{\partial \Lambda_1(s,t)}{\partial t} \bigg|_{t=t^*} = 0. \]  
(19)

Before proceeding further, we define
\[ B := (-\infty, D_o] \times (-\infty, 0], \quad a_B := [D_o, 0]^T, \]  
\[ \Lambda_t^*(b) := \sup_{a \in \mathbb{R}^2} \{\langle b, a \rangle - \Lambda_1(a)\}, \]  
(20)
for any $b \in \mathbb{R}^2$. Note that owing to the convexity of $\Lambda_1(\cdot)$, one can show that
\[ \Lambda_t^*(a_B) = s^* D_o - \Lambda_1(s^*, t^*), \]  
(22)
where $s^*$ and $t^*$ are as defined in (18).

Moreover,
\[ \Lambda_0(0) := \log E_{P_{X,Y,Z}} \left[ e^{\lambda \log \frac{W(Y|X)}{W(Y|Z)}} \right] > 0. \]  
(24)

A result of Ney [19, Theorem] implies that5 there exists a unique dominating point (cf. [19, Definition]) $d_B \in B$, which also satisfies (cf. [20, eq. (3.3)])
\[ \Lambda_1^*(d_B) = \inf_{b \in B} \sup_{a \in \mathbb{R}^2} \{\langle b, a \rangle - \Lambda_1(a)\}. \]  
(25)

Moreover, [19, eq. (2)] implies that
\[ \frac{c_1}{N} e^{-N\Lambda_1^*(d_n)} \leq \bar{\alpha}_N \leq \frac{c_2}{\sqrt{N}} e^{-N\Lambda_1^*(a_a)}, \]  
(26)
for some $c_1, c_2 \in \mathbb{R}^+$ and all sufficiently large $N$. Next, one can prove that
\[ \Lambda_t^*(a_B) = \Lambda_t^*(d_B). \]  
(27)

Plugging (27) into (26) yields
\[ \bar{\alpha}_N \leq \frac{c_2}{\sqrt{N}} e^{-N\Lambda_t^*(a_a)}, \]  
(28)
for some $c_2 \in \mathbb{R}^+$ and all sufficiently large $N$. Moreover, it is possible to check that
\[ \Lambda_1(s^*, t^*) = -\log P_{X,Y,Z} \left\{ \mathcal{S}_Q \right\} + 2\Lambda \left( \frac{\rho^*}{1+\rho^*} \right). \]  
(29)

5Similar to the previous footnote, (24) ensures the fulfillment of the “mean does not belong to the set” condition of the theorem. The remaining conditions are easy to check.
Finally, plugging (18), (22) and (29) into (28) along with (10), (14) and some algebra yields
\[
\tilde{\alpha}_N \leq P_{X,Y,Z} \left\{ \tilde{S}_Q \right\} - N \frac{c_2}{\sqrt{N}} e^{-N[\bar{E}_e(R,W) + R]}.
\]  
(30)

By plugging (30) into (16) and noting P_{N}^{X,Y,Z} \{ \tilde{S}_Q \} = P_{X,Y,Z} \{ \tilde{S}_Q \}^N, we deduce that
\[
\alpha_N \leq \frac{c_2}{\sqrt{N}} e^{-N[\bar{E}_e(R,W) + R]}.
\]  
(31)

Lastly, (12), (15) and (31) imply that
\[
P_{e,1}(Q) \leq K_1 e^{-N\bar{E}_e(R,W)},
\]  
(32)

for some K_1 \in \mathbb{R}^+ and all sufficiently large N.

\[\blacksquare\]

IV. PROOF OF THEOREM 2

Let W ∈ P(\mathcal{Y} | \mathcal{X}) with C(W) > 0 and R < R_{c1}(W) be arbitrary. Assume that for all Q ∈ P(\mathcal{X}) with E_0(1, Q, W) = \max_{P \in P(\mathcal{X})} E_0(1, P, W), the F.D.I.O. property holds. Consider any such Q ∈ P(\mathcal{X}). For this (Q, W) pair, let P_{X,Y,Z} and \tilde{S}_Q and \tilde{X}_y denote the F.D.I.O. property is satisfied, (35) follows from (33) and (36) is true owing to the choice of Q ∈ P(\mathcal{X}) and the fact that R < R_{c1}(W) (e.g. [17, pg. 245]). Hence, the upper bound of (6) follows.

In order to establish the lower bound of (6), one can use Gallager’s arguments [17, pg. 245-246] by noting
\[
P_{N}^{X,Y,Z} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log \frac{W(Y_n | X_n)}{W(Y_n | Z_n)} \leq 0 \right\} = e^{-N\bar{E}_e(1, Q, W)}.
\]  
V. COMPARISON WITH THE SPHERE-PACKING BOUND

Our main result is an improvement of the random coding bound for rates greater than the critical rate, obtained by improving the pre-factor from O(1) to O(1/\sqrt{N}) for a broad class of channels. Even for the symmetric channels, however, this pre-factor does not match its lower bound counterpart [22], namely O(N^{-\frac{1}{2}(1+\bar{E}_e(R,W))}), where \{\bar{E}_e(R,W)\} is the slope of \{\text{Esp}(R)\}. It is worth noting, however, that for rates close to capacity, \{\text{Esp}(R)\} is nearly zero, so the order of the pre-factor is nearly determined in this regime.

REFERENCES

Source Coding with Side-Information at the Receivers and an Application

Michèle Wigger
Telecom ParisTech, Paris, France
Email: michele.wigger@telecom-paristech.fr

Abstract—We review results on the source-coding problem in Figure 1, i.e., on the simple network proposed by Gray and Wyner where additionally the receivers have side-information about the source. We also describe how this problem is relevant for coding over broadcast channels (BC) with feedback.

I. THE SOURCE CODING PROBLEM

We consider the source-coding problem in Figure 1, where a transmitter describes the source sequence $X^n = (X_1, \ldots, X_n)$ to Receivers 1 and 2, who observe the side-information $Y^n_1 = (Y_{1,1}, \ldots, Y_{1,n})$ and $Y^n_2 = (Y_{2,1}, \ldots, Y_{2,n})$, respectively. Here, $\{(X_t, Y_{1,t}, Y_{2,t})\}_{t=1}^n$ is a sequence of independent and identically distributed (IID) triples of joint law $P_{X,Y}$.

The transmitter can send a common message $M_0$ of rate $R_0$ to both receivers and private messages $M_1$ and $M_2$ of rates $R_1$ and $R_2$ to Receivers 1 and 2, respectively. The goal is that Receiver $i$, for $i \in \{1, 2\}$, produces a reconstruction sequence $\hat{X}_i^n = (\hat{X}_{i,1}, \ldots, \hat{X}_{i,n})$ that satisfies

$$D_i = \frac{1}{n} \sum_{t=1}^n d(\hat{X}_{i,t}, X_{i,t}) \leq D_1$$

for $i \in \{1, 2\}$, where $d$ is the symbol-wise distortion function and $D_1 \geq 0$ the distortion constraint. The described setup includes the case where $X^n = (X^n_1, X^n_2)$ and Receiver $i$ only wishes to reconstruct a lossy version of $X^n_i$.

Theorem I (\cite{6}): A triple $(R_0, R_1, R_2)$ is achievable if

$$R_0 + R_i \geq I(X; X_i, V_0|Y_i), \quad i \in \{1, 2\}$$

$$R_0 + R_1 + R_2 \geq I(X; X_1|Y_1, V_0) + I(X; X_2|Y_2, V_0) + \max_{i \in \{1, 2\}} I(X; V_0|Y_i)$$

for some $\hat{X}_1, \hat{X}_2, V_0$ such that $\hat{X}_1, \hat{X}_2, V_0 \rightarrow X \rightarrow Y_1, Y_2$ form a Markov chain and $E[d(\hat{X}_{i}, X_{i})] \leq D_1$, for $i \in \{1, 2\}$.

Though the region is not known to be optimal in general, it is optimal for all special cases where the rates-distortions region is known. In the following we review these special cases.

II. APPLICATION TO THE BC WITH FEEDBACK

Our lossy source-coding problem has application to coding over BCs with feedback, see especially \cite{5} but also \cite{6, 7}. The core idea of these schemes is the following. In a first stage the transmitter sends the desired messages at rates that do not allow for decoding at the receivers. Then, using the feedback, it identifies resolution information which it sends during the second stage. To this end, it first compresses the resolution information using a good lossy source-code for the setup in Figure 1, where the side-information at the receivers is given by their channel outputs in the first stage. Then it sends the compression indices over the BC so that the receivers can decode them. With the lossy reconstructions of the resolution information learned in the second stage and the outputs from the first stage the receivers then decode the desired messages.

REFERENCES


\[2\] C. Heegard and T. Berger, “Rate distortion when side information may be absent”, IT-31, no. 6, pp. 727–734, March 1985.


\[6\] L. Georgiadis and L. Tassiulas, Broadcast erasure channel with feedback capacity and algorithms, in Proc. 5th Workshop on Network Coding, Theory, & Applications, Lausanne, Switzerland, June 2009.

On Gaussian Channels with MMSE Interference

Ronit Bustin and Shlomo Shamai (Shitz)
Department of Electrical Engineering, Technion-Israel Institute of Technology.
Technion City, Haifa 32000, Israel, Email: bustin@tx.technion.ac.il, sshlomo@ee.technion.ac.il

Abstract—We examine codes, over the additive Gaussian noise channel, designed for for reliable communication at some specific signal-to-noise ratio (snr) and constrained by the permitted MMSE at some lower snr. We show that the maximum possible rate is the one attained by superposition codebooks. Moreover, the MMSE function of codes attaining this maximum rate under the MMSE constraint is completely defined for all snr. The problem is also extended to the maximization of the rate under two MMSE constraints. The optimal rate is again achieved by three-layers superposition codes.

I. INTRODUCTION

Capacity and capacity achieving codes have been the main concern of Information Theory from the very beginning. Trying to design capacity achieving codes is a central goal of many researchers in this field. Recently some emphasis has been given to the research of non-capacity achieving point-to-point codes [1], [2]. These codes, referred to as “bad” point-to-point codes [2], are heavily used in many multi-terminal wireless networks. In [3] it was shown that the mutual information and thus also the minimum mean square error (MMSE) of “good” (capacity achieving) point-to-point codes is known exactly, for no matter the specific structure of the code. Furthermore, it is known that “bad” codes can obtain lower MMSE at low signal-to-noise ratios (snrs) [2]. This advantage is meaningless in point-to-point communication, where all that matters is the performance at the receiver. However, in multi-terminal wireless networks, such as a cellular network, the case is different. In such networks there are two fundamental phenomena: interference from one node to another (an interference channel), and the potential cooperation between nodes (a relay channel). In the interference channel, where a message sent to an intended receiver acts as interference to other receivers in the network, a lower MMSE implies better interference cancelation, and thus improved rates for the interfered user. The performance of optimal point-to-point codes in the interference setting was the investigation in [4].

The best known achievable region for the two-user interference channel is given by the Han and Kobayashi (HK) scheme [5]. This scheme uses partial decoding of the interfering message at the receiver. Rate splitting (is, superposition coding) is a special case of the HK scheme, and is also point-to-point “bad” (see [1, Appendix VIII-C]). It was shown in [6] that these codes are close to optimal for the Gaussian interference channel, and in fact are within one bit from capacity. In this work we show that these codes are in fact optimal MMSE-wise.

Although it is known that we can obtain an advantage, MMSE-wise, when using “bad” codes, the question “how much better can we do, given a looser requirement on the rate?” is still open. In this work we answer this question for the additive Gaussian noise channel showing that superposition codebooks, optimal for a specific Gaussian broadcast channel (BC) are optimal in this sense. We further extend the question to the case of two MMSE constraints in lower snrs, and show that three-layers superposition codebooks attain the maximum possible rate. This result can also be extended to the general K receiver case.

A similar question has been raised in the work of Bandemer and El Gamal [7], where they provide the rate-disturbance region: for any given rate that can be transmitted reliably to the intended receiver, what is the minimum possible disturbance that can be attained at some interfered user. In [7] the authors measure the disturbance using the mutual information between the codeword and the output at the interfered user, rather than the minimum possible MMSE, as done here. We further discuss and compare the two measures in our concluding remarks (section IV).

II. PROBLEM FORMULATION AND PRELIMINARY RESULTS

In this work we are looking at the transmission of codewords, of length n, through a discrete memoryless standard Gaussian channel:

\[ Y = \sqrt{n}X + N \] (1)

where \( N \) is standard additive Gaussian noise. The codewords are constrained by the standard average power constraint: \( \forall x \in C_n \quad \frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq 1 \), where \( C_n \) stands for a code of n-dimensional codewords. We examine codebooks designed for reliable transmission at \( \gamma = \text{snr}_2 \) (reliable decoding of the codeword from \( Y(\gamma = \text{snr}_2) \)). Our main interest will be in examining non-optimal codes, alternatively known as “bad” codes [2], defined using code-sequences, as follows:

**Definition 1:** A non-optimal code-sequence \( C = \{ C_n \}_{n=1}^{\infty} \), for a channel with capacity \( C \), is a code-sequence with vanishing error probability and rate satisfying \( \lim_{n \to \infty} \frac{1}{n} \log M_n < C \) where \( M_n \) is the size of code \( C_n \).

Their associated MMSE defined as:

\[ \text{MMSE}^n(\gamma) = \lim_{n \to \infty} \text{MMSE}^n(\gamma) = \lim_{n \to \infty} \frac{1}{n} \text{Tr}(E_X(\gamma)) \] (2)

where \( E_X(\gamma) \) is the MMSE matrix when estimating the codeword \( X \) from the output of the channel \( Y = \sqrt{n}X + N \). We further define the following for abbreviation:

\[ I(\gamma) = \lim_{n \to \infty} I_n(\gamma) = \lim_{n \to \infty} \frac{1}{n} I(X; Y(\gamma)) \] (3)

20
Surely, for any such codes, the error probability for any $\gamma > \text{snr}_2$ is zero, when $n \to \infty$, since reliable transmission is guaranteed at $\text{snr}_2$. As a result $\text{MMSE}^c(\gamma)$ for these snrs is also zero. On the other hand, for $\gamma < \text{snr}_2$ the value of the error probability is not guaranteed to be any specific value. For an optimal code, it was shown in [3], that $I(\gamma)$ for $\gamma < \text{snr}_2$ follows that of the Gaussian i.i.d. input and thus $\text{MMSE}^c(\gamma)$ is also known exactly and descends gradually according to $1/\gamma$.

Our goal is to find a code-sequence which both attains a minimum required rate at $\text{snr}_2$, denoted as $I(\text{snr}_2) \geq \frac{1}{2}\log(1 + \alpha \text{snr}_2)$, for some predetermined $\alpha \in (0, 1]$, and secondly, obtains the minimum possible MMSE at some $\text{snr}_1 < \text{snr}_2$. Clearly, if $\alpha = 1$ we have the maximum rate at $\text{snr}_2$, attainable only by an optimal code-sequence. As mentioned above, this is a well understood case [3]. For $\alpha \leq \frac{\text{snr}_1}{\text{snr}_2}$ we also have an obvious solution, since for these low rates we can obtain zero $\text{MMSE}^c(\text{snr}_1)$ by simply choosing an optimal code-sequence of rate $\frac{1}{2}\log(1 + \text{snr}_1)$. Thus, the interesting range of parameters is $\frac{\text{snr}_1}{\text{snr}_2} < \alpha < 1$, for which the minimum attainable $\text{MMSE}^c(\text{snr}_1)$ is unknown. This problem is equivalent to maximizing the rate at $\text{snr}_2$, given some constraint on the maximum allowed $\text{MMSE}$ at $\text{snr}_1$. In this formulation the obvious extension, motivated by multi-user interference channels, is to maximize the rate at $\text{snr}_2$, given two MMSE constraints. In this work we proceed with this view of the problem.

A. The I-MMSE approach

The approach used in order to provide insight into the above mentioned problem is the I-MMSE approach, this to say that we make use of the fundamental relationship between the mutual information and the MMSE in the Gaussian channel and its generalizations [8], [9]. Even though we are examining a scalar Gaussian channel, the $n$-dimensional version of this relationship is required since we are looking at the transmission of $n$-dimensional codewords through the channel. In our setting the relationship is as follows:

$$I_n(\text{snr}) = \frac{1}{2} \int_0^{\text{snr}} \text{MMSE}^c(\gamma) \, d\gamma. \quad (4)$$

Taking the limit of $n \to \infty$ on both sides results with:

$$I(\text{snr}) = \frac{1}{2} \int_0^{\text{snr}} \text{MMSE}^c(\gamma) \, d\gamma. \quad (5)$$

The main property of the I-MMSE used for these proofs is an $n$-dimensional “single crossing point” property derived in [10] given here for completeness. This property is an extension of the scalar “single crossing point” property shown in [11]. In [10] the following function is defined as an arbitrary random vector X:

$$q_\text{A}(X, \sigma^2, \gamma) = \frac{\sigma^2}{1 + \sigma^2\gamma} \text{Tr}(A) - \text{Tr}(AE_X(\gamma)) \quad (6)$$

where $A$ is some $n \times n$ general weighting matrix. The following theorem is proved in [10].

Theorem 1 ([10]): Let $A \in S^n_+$ be a positive semidefinite matrix. Then, the function $\gamma \mapsto q_\text{A}(X, \sigma^2, \gamma)$, defined in (6), has no nonnegative-to-negative zero crossings and, at most, a single negative-to-nonnegative zero crossing in the range $\gamma \in [0, \infty)$. Moreover, let $\text{snr}_0 \in [0, \infty)$ be that negative-to-nonnegative crossing point. Then,

1. $q_\text{A}(X, \sigma^2, 0) \leq 0$.
2. $q_\text{A}(X, \sigma^2, \gamma)$ is a strictly increasing function in the range $\gamma \in [0, \text{snr}_0)$.
3. $q_\text{A}(X, \sigma^2, \gamma) \geq 0$ for all $\gamma \in [\text{snr}_0, \infty)$.
4. $\lim_{\gamma \to \infty} q_\text{A}(X, \sigma^2, \gamma) = 0$.

In this work, the matrix $A$ can be set to the identity matrix. The above property is valid for all natural $n$, thus we may also take $n \to \infty$.

B. Superposition Coding

An important family of non-optimal codes, that is, a family of codes that do not attain the point-to-point capacity at $\text{snr}_2$, is that of Gaussian superposition codes which are optimal for a degraded Gaussian BC [12]. As will be shown in the sequel these codes are optimal MMSE-wise. The analysis of this family was done by Merhav et. al. in [13, section 5.3] from a statistical physics perspective. As noted in [13], the MMSE of this family of codebooks undergoes phase transitions, that is, it is a discontinuous function of $\gamma$. The mutual information, $I(\gamma)$, and $\text{MMSE}^c(\gamma)$ of this family of codebooks is known exactly and given in the next theorem. An example is depicted in Figure 1.

Theorem 2 ([13] section 5.3): A superposition codebook designed for $(\text{snr}_0, \text{snr}_1, \ldots, \text{snr}_K)$ with the rate-splitting coefficients $(\beta_0, \ldots, \beta_{K-1})$ has the following $I(\gamma)$:

$$\begin{cases}
\frac{1}{2}\log(1 + \gamma), & \text{if } 0 \leq \gamma < \text{snr}_0 \\
\frac{1}{2}\log \left( \frac{1+\gamma}{1+\text{snr}_0} \prod_{j=1}^{K-1} \frac{1+\beta_j\gamma}{1+\beta_j\text{snr}_j} \right) + \frac{1}{2}\log(1 + \beta_{K-1}\gamma), & \text{if } \text{snr}_0 \leq \gamma \leq \text{snr}_{K-1} \\
\frac{1}{2}\log \left( \frac{1+\gamma}{1+\text{snr}_0} \prod_{j=1}^{K-1} \frac{1+\beta_j\gamma}{1+\beta_j\text{snr}_j} \right) + \frac{1}{2}\log(1 + \beta_{K-1}\text{snr}_K), & \text{if } \text{snr}_{K-1} < \gamma
\end{cases}$$

and the following $\text{MMSE}^c(\gamma)$:

$$\text{MMSE}^c(\gamma) = \begin{cases}
\frac{1}{\beta_j}, & 0 \leq \gamma < \text{snr}_0 \\
\frac{1}{\beta_j}, & \text{snr}_j \leq \gamma < \text{snr}_{j+1} \\
0, & \text{snr}_{j+1} \leq \gamma
\end{cases} \quad (7)$$

III. Main Results

The main result is given in the next theorem.

Theorem 3: Assuming $\text{snr}_1 < \text{snr}_2$ the solution of the following optimization problem:

$$\max_{\text{snr}_2} I(\text{snr}_2)$$

s.t. $\text{MMSE}^c(\text{snr}_1) \leq \frac{\beta}{1 + \beta\text{snr}_1}$

for some $\beta \in [0, 1]$, is the following

$$I(\text{snr}_2) = \frac{1}{2}\log(1 + \beta\text{snr}_2) + \frac{1}{2}\log \left( 1 + \frac{\text{snr}_1}{1 + \beta\text{snr}_1} \right)$$

(9)
and is attainable when using the optimal Gaussian superposition codebook designed for \((\text{snr}_1, \text{snr}_2)\) with a rate-splitting coefficient \(\beta\).

Proof: It is simple to verify that the optimal Gaussian superposition codebook complies with the above MMSE constraint and attains the maximum rate. Thus, the focus of the remainder of the proof is on deriving a tight upper bound on the rate. We prove an equivalent claim, by assuming a code of rate \(R_c = \frac{1}{2}\log (1 + \alpha \text{snr}_2)\), designed for reliable transmission at \(\text{snr}_2\), and deriving a lower bound on \(\text{MMSE}(\gamma)\).

If \(\text{osnr}_2 \leq \gamma \leq 1\) the lower bound is trivially zero using the optimal Gaussian codebook designed for \(\text{osnr}_2\). This is equivalent to setting \(\beta = 0\) in which case we obtain the rate \(I(\text{snr}_2) = \frac{1}{2}\log (1 + \text{snr}_1)\) using an optimal Gaussian codebook (no superposition). Thus, we assume \(\gamma \leq \text{osnr}_2\).

Using the trivial upper bound on \(I(\gamma) \leq \frac{1}{2}\log (1 + \gamma)\) (due to maximum entropy), we can lower bound the following difference, for any \(\gamma < \alpha \text{snr}_2\):

\[
I(\text{snr}_2) - I(\gamma) \geq I(\text{snr}_2) - \frac{1}{2}\log (1 + \gamma).
\]  

Using the I-MMSE relationship (5), the above translates to the following inequality:

\[
\frac{1}{2} \int_{\gamma}^{\text{snr}_2} \text{MMSE}(\tau) \, d\tau \geq R_c - \frac{1}{2}\log (1 + \gamma) \geq \frac{1}{2}\log (1 + \alpha \text{snr}_2) - \frac{1}{2}\log (1 + \gamma).
\]  

\[
\frac{1}{2} \int_{\gamma}^{\text{snr}_2} \text{MMSE}(\tau) \, d\tau \geq R_c - \frac{1}{2}\log (1 + \gamma) = \frac{1}{2}\log (1 + \alpha \text{snr}_2) - \frac{1}{2}\log (1 + \gamma).
\]  

Defining \(d\) through the following equality:

\[
\frac{1}{2}\log (1 + \alpha \text{snr}_2) - \frac{1}{2}\log (1 + \gamma) = \frac{1}{2}\log (1 + d\text{snr}_2) - \frac{1}{2}\log (1 + d\gamma).
\]  

\[
\frac{1}{2} \int_{\gamma}^{\text{snr}_2} \text{MMSE}(\tau) \, d\tau \geq \frac{1}{2} \log (1 + d\text{snr}_2) - \frac{1}{2} \log (1 + d\gamma).
\]  

where \(\text{mmse}_G(\tau)\) is the MMSE assuming a Gaussian random variable with variance \(d\) through the additive Gaussian channel at \(\text{snr} = \tau\). The single crossing point property (Theorem 1, with \(A = I\)) tells us that \(\text{MMSE}(\tau)\) and \(\text{mmse}_G(\tau)\) cross each other at most once, and after that crossing point \(\text{mmse}_G(\tau)\) remains an upper bound. From the inequality in (13) we can thus conclude that the single crossing point, if exists, will occur in the region \((\gamma, \infty)\). Thus, for \(\gamma\) we have the following lower bound:

\[
\text{MMSE}(\gamma) \geq \frac{d(\gamma)}{1 + d(\gamma)\gamma} = \frac{\alpha \text{snr}_2 - \gamma}{\text{snr}_2 - \gamma} \frac{1}{1 + \gamma}
\]  

A similar derivation can be done for any \(\gamma < \text{snr}_2\), and will result with a different \(d(\gamma)\). Specifically for \(\gamma = \text{snr}_1\) we obtain

\[
\text{MMSE}(\text{snr}_1) \geq \frac{\alpha \text{snr}_2 - \text{snr}_1}{\text{snr}_2 - \text{snr}_1} \frac{1}{1 + \text{snr}_1}.
\]  

Deriving \(\alpha\) as a function of the constraining \(\beta\), and substituting it in \(R_c = \frac{1}{2}\log (1 + \alpha \text{snr}_2)\) results with the superposition rate given in (9).

An interesting question to ask is whether there could be a different code that can attain maximum rate under the MMSE constraint at \(\text{snr}_1\) and also provide better MMSE for other values of \(\text{snr}\). The answer is to the negative, and is given in the next theorem, proof of which is omitted and can be found in [14].

Theorem 4: From the set of reliable codes of rate \(R_c = \frac{1}{2}\log (1 + \beta \text{snr}_2)\) the MMSE constraint at \(\text{snr}_1\), the superposition codebook provides the minimum MMSE for all \(\text{snr}\).

Two MMSE Constraints: We now wish to extend the result of Theorem 3 to the case of two constraints:

Theorem 5: Assuming \(\text{snr}_0 < \text{snr}_1 < \text{snr}_2\) the solution of the following optimization problem:

\[
\max \quad I(\text{snr}_2)
\]

s.t. \(\text{MMSE}(\text{snr}_1) \leq \frac{\beta_1}{1 + \beta_1 \text{snr}_1}\)

\[
\text{MMSE}(\text{snr}_0) \leq \frac{\beta_0}{1 + \beta_0 \text{snr}_0}
\]

for some positive \(\beta_1, \beta_0\) such that \(\beta_1 + \beta_0 \leq 1\) and \(\beta_1 < \beta_0\), is the following:

\[
I(\text{snr}_2) = \frac{1}{2} \log \left(1 + \beta_1 \text{snr}_2 \frac{1 + \beta_0 \text{snr}_1 + 1 + \text{snr}_0}{1 + \beta_1 \text{snr}_1 + 1 + \beta_0 \text{snr}_0}\right)
\]

and is attainable when using the optimal three-layers Gaussian superposition codebook designed for \((\text{snr}_0, \text{snr}_1, \text{snr}_2)\) with rate-splitting coefficients \((\beta_0, \beta_1)\).
When $\beta_0 < \beta_1$ the first constraint can be removed and we return to the case of a single constraint given in Theorem 3.

Proof: It is simple to verify that the optimal Gaussian three-layers superposition codebook complies with the above MMSE constraints and attains the maximum rate. Thus, we need to derive a tight upper bound on the rate. Deriving the upper bound begins with the usage of Theorem 3 on $I(\text{snr}_1)$, when we take into consideration only the constraint on $\text{MMSE}^c(\text{snr}_0)$. This provides the following upper bound:

$$I(\text{snr}_1) \leq \frac{1}{2} \log \left( 1 + \beta_0 \text{snr}_1 \right) + \frac{1}{2} \log \left( 1 + \frac{\text{snr}_0}{1 + \beta_1 \text{snr}_1} \right)$$  \hspace{1cm} (15)

On the other hand, using the I-MMSE approach we have:

$$I(\text{snr}_2) - I(\text{snr}_1) = \frac{1}{2} \int_{\text{snr}_1}^{\text{snr}_2} \text{MMSE}^c(\tau) \, d\tau \leq \frac{1}{2} \int_{\text{snr}_1}^{\text{snr}_2} \text{mmse}^c(\tau) \, d\tau$$

where $\text{mmse}^c(\tau)$ is the MMSE assuming a Gaussian random variable with variance $\beta_1$. This inequality is valid since according to the first constraint we have

$$\text{MMSE}^c(\text{snr}_1) \leq \frac{\beta_1}{1 + \beta_1 \text{snr}_1} = \text{mmse}^c(\text{snr}_1)$$  \hspace{1cm} (16)

thus, according to the single crossing point property

$$\text{MMSE}^c(\tau) \leq \text{mmse}^c(\tau), \quad \forall \tau \geq \text{snr}_1$$  \hspace{1cm} (17)

leading to the inequality in equation (16). Putting together (15) and (16) we obtain the desired upper bound. \qed

IV. DISCUSSION AND CONCLUSIONS

These results provide the engineering insight to the good performance of the HK superposition scheme on the two-user interference channel, as shown in [6]. From our results, that show that the HK superposition scheme, is optimal MMSE-wise, we can conclude that one cannot construct better codes of the type defined in [1, 2] that will beat HK, through the use of estimation. Note that, as mentioned in [1, section V], the codes constructed there have an important complexity advantage over HK codes. Furthermore, keep in mind that the HK scheme is efficient in the two-user interference channel and a simple approach in the general $K$-user interference channel. Finally, we expect that the I-MMSE approach to shed further light on the capacity region of the two-user interference channel itself.

As mentioned in the introduction, Corollary 2 in [7] can also be derived directly from the I-MMSE formulation. Using our notation, we begin with the MMSE rate. Since, $0 \leq I(\text{snr}_1) \leq \frac{1}{2} \log \left( 1 + \alpha^* \text{snr}_1 \right)$ there exists an $\alpha^* \in [0, 1]$ such that $R_d = \frac{1}{n} \log \left( 1 + \alpha^* \text{snr}_1 \right)$. Thus, the averaged MMSE of the code crosses the MMSE of a scalar Gaussian input with power $\alpha^*$ in the range $[0, \text{snr}_1]$. Now, using the I-MMSE,

$$I(\text{snr}_2) = \frac{1}{2} \log \left( 1 + \alpha^* \text{snr}_1 \right) + \int_{\text{snr}_1}^{\text{snr}_2} \text{MMSE}^c(\gamma) \, d\gamma$$

$$\leq \frac{1}{2} \log \left( 1 + \alpha^* \text{snr}_2 \right)$$  \hspace{1cm} (18)

where the last transition is due to the “single crossing point” property which ensures that the MMSE of the scalar Gaussian input with power $\alpha^*$ will remain an upper bound on $\text{MMSE}^c(\gamma)$ in the range $[\text{snr}_1, \infty)$. Moreover, the above derivations do not indicate a superposition coding scheme, but rather a Gaussian code with reduced power of $\alpha^*$. Such a scheme, which attains the required minimum rate at $\text{snr}_2$, does not attain the minimum MMSE at $\text{snr}_1$. Extending to two mutual information disturbance constraint is trivial, as only the most constraining one determines the result. Thus, the different measurement of “disturbance” suggested here is conceptually different than the one suggested in [7].

Finally, the two MMSE constraints solution, presented here, can be directly extended to the $K$ MMSE constraint, as shown in [15].

ACKNOWLEDGMENT

This work has been supported by the Israel Science Foundation (ISF) and the CORNET Consortium. Ronit Bustin is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

REFERENCES

On the Gaussian Listening-Helper Source-Coding Problem

Shraga I. Bross∗, and Haim Permuter†
∗ Engineering Department, Bar-Ilan University, Ramat Gan, 52900, Israel
Email: brosss@biu.ac.il
† Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva
Email: haimp@bgu.ac.il

Abstract—In the Gaussian listening-helper source-coding problem Encoder 1 and Decoder 1 observe respectively a pair of i.i.d. correlated Gaussian sources and both wish to communicate the first source to Decoder 2 subject to a distortion constraint. Encoder 1 sends a message to both decoders and then Decoder 1—the listening-helper—sends a message just to Decoder 2.

We derive an inner bound on the rate region of this problem by proposing a Gaussian achievable scheme. This inner bound coincides with the solution to the Gaussian optimization problem associated with a single-letter outer bound corresponding to this problem.

Index Terms—multi-terminal source coding, listening-helper problem.

I. INTRODUCTION

The Gaussian listening-helper source-coding problem that we consider is depicted in Fig. 1. Encoder 1 observes a scalar i.i.d. Gaussian source X that is correlated with a scalar i.i.d. source Y observed by Decoder 1. Encoder 1 wishes to communicate the source X to within certain distortion constraint to Decoder 2 by sending a message at rate R to both decoders. Decoder 1 (the listening-helper) is allowed to communicate to Decoder 2 by sending a message at rate R1. Decoder 2 uses both messages to form its estimate of the source X. The goal is to determine the set of all rate pairs (R, R1) that satisfy the distortion constraint.

We derive an inner bound on the rate region for this source-coding problem by proposing a Gaussian scheme in which Encoder 1 quantizes its observation using a Gaussian test channel and then compresses the quantized values using proper binning. Decoder 1 quantizes its observation using another test channel and finally, both Encoder 1 and Decoder 1 cooperate to provide Decoder 2 with the enumeration information it needs in order to resolve its ambiguity regarding the compressed information Encoder 1 had generated. Furthermore, we show that the Gaussian solution associated with an upper bound for our problem coincides with the proposed Gaussian achievable scheme.

The paper is organized as follows. In Section II we provide a formal definition of the Gaussian listening-helper source-coding model. In Section III we present our main result, while Section IV is devoted to the description of the main steps of the proof.

The work of S. Bross was supported by the Israel Science Foundation under Grant 497/09.

II. PROBLEM FORMULATION

The core problem we consider, the listening-helper source-coding problem, is described as follows. Let \(\{(X_i, Y_i)\}_{i=1}^{n}\) be a pair of zero-mean Gaussian random vectors generated by i.i.d drawings of a pair of jointly Gaussian random variables with variance \(\sigma_X^2\) and \(\sigma_Y^2 = \sigma_X^2 + \sigma_W^2\), respectively. With no loss of generality, we can write

\[
Y_i = a_i X_i + N_i^{(y)},
\]

where \(N_i^{(y)} \sim \mathcal{N}(0, \sigma_W^2)\) is independent of \(X_i\).

Encoder 1 which observes \(X^n\) sends a message to both decoders using an encoding function

\[
f^{(n)}: X^n \to \{1, 2, \ldots, M_f^{(n)}\},
\]

where \(X^n\) denotes the set of \(n\)-vectors over \(X\). Based on \(f^{(n)}(X^n)\) and its observation \(Y^n\) Decoder 1 sends a message to Decoder 2 using an encoding function

\[
f_1^{(n)}: \{1, 2, \ldots, M_f^{(n)}\} \times Y^n \to \{1, 2, \ldots, M_1^{(n)}\}.
\]

Decoder 2 then uses both received messages to estimate \(X^n\) using a reproduction function

\[
g^{(n)}: \{1, 2, \ldots, M_f^{(n)}\} \times \{1, 2, \ldots, M_1^{(n)}\} \to \hat{X}_2^n.
\]

Definition 1: A rate-distortion triple \((R, R_1, D_2)\) is achievable for the Gaussian listening-helper source-coding problem if there exists a block length \(n\), encoders \(f^{(n)}\) and \(f_1^{(n)}\), and a decoder \(g^{(n)}\) such that,

\[
R \geq \frac{1}{n} \log M_f^{(n)}
\]

\[
R_1 \geq \frac{1}{n} \log M_1^{(n)}
\]

\[
D_2 \geq \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \hat{X}_2,i)^2].
\]

Let \(\hat{R}(D_2)\) be the closure of the set of achievable rate-distortion triples. The rate region \(\hat{R}(D_2)\) for the Gaussian listening-helper source-coding problem is defined by

\[
\mathcal{R}(D_2) = \left\{ (R, R_1): (R, R_1, D_2) \in \hat{R}(D_2) \right\}.
\]
Since we are interested in the Mean-Square-Error (MSE) distortion constraint we can restrict the reproduction function to be the MMSE estimate of $X^n$ given $f^{(n)}(X^n)$ and $f_1^{(n)}(f^{(n)}(X^n),Y^n)$, that is
\[
\hat{X}^n = \mathbb{E} \left[ X^n | f^{(n)}(X^n), f_1^{(n)}(f^{(n)}(X^n), Y^n) \right].
\]

### III. Main Results

Our main result is a characterization of an inner bound on the rate-region for the Gaussian listening-helper source-coding problem.

**Proposition 1:** For a positive number $D_2 \leq \sigma_X^2$ the rate-region $R(D_2)$ contains the rate region $R$ defined by
\[
R = \bigcup (R, R_1) : R \geq \frac{1}{2} \log \left[ 1 + \frac{\sigma_X^2}{\sigma_{W_0}^2} \frac{\sigma_{W_0}^2}{\sigma_{W_0}^2} \right], R_1 \geq \frac{1}{2} \log \left[ 1 + \frac{\sigma_X^2}{\sigma_{W_0}^2} \frac{\sigma_{W_0}^2}{\sigma_{W_0}^2} \right],
\]
\[
R + R_1 \geq \frac{1}{2} \log \left[ \frac{\sigma_X^2}{\sigma_{W_0}^2} \frac{\sigma_{W_0}^2}{\sigma_{W_0}^2} \right], D_2 = \frac{1}{1 + \frac{\sigma_X^2}{\sigma_{W_0}^2} + \frac{\sigma_{W_0}^2}{\sigma_{W_0}^2}} .
\]

### A. A Gaussian Achievable Scheme

A natural Gaussian achievable scheme for the listening-helper problem is described as follows. Let $S(U,V)$ be the set of zero-mean jointly Gaussian random variables $U$ and $V$ such that
1) $U \circlearrowleft X \circlearrowright Y$ and $X \circlearrowleft U Y \circlearrowright V$ are Markov chains.
2) $\sigma_X^2(U,V) \leq D_2$.

For any $(U,V) \in S(U,V)$ and a large block-length $n$:
- Generate $e^{n I(X;U)+\epsilon}$, $\epsilon > 0$ independent codewords $u(1), \ldots, u(e^{n I(X;U)+\epsilon})$ randomly according to $p_U$. Assign these codewords uniformly into $e^{n R}$ bins. Label them $u(j,l), j \in \{1, \ldots, e^{n I(X;U)-R+\epsilon}\}, l \in \{1, \ldots, e^{n R}\}$. Share the codebook and its partition between Encoder 1 and both decoders.
- For each $u(j,l)$ generate $e^{n R_1+\epsilon_1}$, $\epsilon_1 > 0$ independent codewords randomly according to $p_{V|U}$. Label the codewords by $v(j,l,k)$, $k \in \{1, \ldots, e^{n R_1+\epsilon_1}\}$ and share this codebook between Decoder 1 and Decoder 2.

Given a source and a side-information sequence pair $(x,y)$:
1) Encoder 1 looks for the codeword $u(j,l)$ that is jointly typical with $x$, and sends its index $l$ to both decoders.
2) Encoder 1 and Decoder 1 send together the enumeration index $j$ to Decoder 2.
3) Decoder 1 looks for the codeword $v(j,l,k)$ that is jointly typical with $(u(j,l), y)$, and sends the index $k$ of that codeword to Decoder 2.

Decoder 1 can recover $u$ with high probability provided that
\[
R \geq I(X;U|Y).
\]

Decoder 1 can compute $v$ with high probability provided that
\[
R_1 \geq I(Y;V|U),
\]
and taking into account that Encoder 1 and Decoder 1 joint efforts in sending the index $j$ to Decoder 2 we obtain
\[
R + R_1 \geq I(X;U) + I(Y;V|U).
\]

Decoder 2 computes the MMSE estimate of the source sequence $X^n$ given $(U^n, V^n)$ and condition 2 above guarantees that the estimate meets the desired distortion.

The following lemma characterizes an achievable rate-region based on this scheme.

**Lemma 1:** A Gaussian achievable scheme as above which satisfies the long Markov chain $U \circlearrowleft X \circlearrowright Y \circlearrowright V$ achieves $R(D_2)$. Moreover, the Gaussian solution associated with an upper bound for our problem coincides with $R(D_2)$.

In Section IV.B we show that for Gaussian $(U,V) \in S(U,V)$ such that $U \circlearrowleft X \circlearrowright Y \circlearrowright V$ the rate region associated with (7)-(9) equals $R(D_2)$.

### IV. Proofs

The source-coding problem considered next will be used to show that the Gaussian solution associated with an upper bound for our main problem equals $R(D_2)$.

#### A. A source-coding subproblem

Let $(\{X_i, Y_i, Z_i\})_{i=1}^n$ be a sequence of independent drawings of a triplet of zero-mean jointly Gaussian random variables with covariance $K_{(X,Y,Z)}$. An encoder observes $Y^n$ while both the encoder and the decoder have access to the side-information $Z^n$. The encoder sends a message to the decoder using an encoding function
\[
f_E^{(n)} : Y^n \times Z^n \rightarrow \{1, 2, \ldots, M^{(n)}\}.
\]

Based on $f_E^{(n)}(Y^n, Z^n)$ and its side-information $Z^n$ the decoder forms an estimate for $X^n$ using a reproduction function
\[
g_D^{(n)} : \{1, 2, \ldots, M^{(n)}\} \times Z^n \rightarrow \hat{X}^n.
\]
Definition 2: A rate-distortion pair \((R_E, D)\) is achievable for the Gaussian source-coding sub-problem if there exists a block length \(n\), an encoder \(f_E^{(n)}\) and a decoder \(g_D^{(n)}\) such that,

\[
R_E \geq \frac{1}{n} \log M^{(n)}
\]
\[
D \geq \frac{1}{n} \sum_{i=1}^{n} E \left[ (X_i - \hat{X}_i)^2 \right].
\] (12)

Here also since we are interested in the MSE distortion constraint we can restrict the reproduction function to be the MMSE estimate of \(X^n\) given \(f_E^{(n)}(Y^n, Z^n)\) and \(Z^n\), that is

\[
\hat{X}^n = E \left[ X^n | f_E^{(n)}(Y^n, Z^n) \right].
\]

For a given distortion \(D > 0\), define \(P_0(D)\) as the set of laws \(p_{X^nZ^n}(x, y, z, v)\), \(x, y, z, v \in \mathbb{R}\), such that the marginal law \(f^{(n)}(x, y, z, v)\) identifies with \(p_{X^nZ^n}\), and

\[
E \left[ (X - E(X|V, Z))^2 \right] \leq D.
\]

Then, by a modification of [4, sec. 6.1.1], we have that

\[
R_E \geq R_{Y|Z}(D) = \min_{p_{X^nYZV} \in P(D)} I(Y; V|Z).
\] (13)

Consequently, define an optimization problem \(P\) as follows

\[
\max h(Y|V) \quad \text{subject to} \quad D \geq \sigma^2_{Y|V, Z},
\] (14)

then let \(P_G[D, (X, Y, Z), \sigma^2_{Y|V, Z}]\) be the corresponding optimization problem assuming \(V\) is Gaussian and denote by \(\phi \left( P_G[D, (X, Y, Z), \sigma^2_{Y|V, Z}] \right) \) the conditional variance \(\sigma^2_{Y|V, Z}\) associated with the optimal value of \(P_G\).

Consequently, \(P_G\) may be expressed as

\[
R_{Y|Z}(D) = \min_{p_{X^nYZV} \in P(D)} \frac{1}{2} \log \frac{\sigma^2_{Y|Z}}{\sigma^2_{Y|V, Z}}.
\] (15)

On the other hand, when \(Z^n\) is available just at the decoder, - i.e. \(f_E^{(n)} = f_E^{(n)}(Y^n)\), by a modification of [2], we have that

\[
\hat{R}_E \geq R^{WZ}_{Y|Z}(D) = \min I(Y; V|Z),
\] (16)

where the minimization is over the set of laws \(P(D)\) defined by

\[
p_{X^nYZV}(x, y, z, v) = p_{X^nYZ}(x, y, z) p_V(y|v).
\] (17)

Consequently,

\[
R^{WZ}_{Y|Z}(D) = \min_{P_{X^nYZV} \in P(D)} \frac{1}{2} \log \frac{\sigma^2_{Y|Z}}{\sigma^2_{Y|V, Z}}
\]
\[
\text{subject to} \quad D \geq \sigma^2_{Y|V, Z},
\]
\[
V^g \ni Y \ni (X, Z).
\] (18)

In [3, Section 3] the author shows that, when the encoder observes the source sequence \(X^n\) - i.e. \(Y = X\), both optimization problems (15) and (18) have the same solution and, in fact, the additional Markov relation \(Z \ni X \ni V^g\) in (18) characterizes the optimal solution of (15). The following lemma asserts that the same is true when the encoder observes a correlated random variable \(Y\) which satisfies a Markov relation.

Lemma 2: For jointly Gaussian random variables \((X, Y, Z)\) satisfying \(Z \ni X \ni Y\) both optimization problems (15) and (18) have the same solution and, in fact, the additional Markov relation in (18) characterizes the optimal solution of (15).

Proof: Omitted.

B. The Gaussian solution associated with an outer bound

Lemma 3: (Single-letter outer bound) If the rate-distortion triple \((R, R_1, D_2)\) is achievable then there exist random variables \((U, V)\) satisfying the Markov relation \(U \ni X \ni Y\), such that

\[
R + R_1 \geq I(X; U) + I(Y; V|U)
\]
\[
R \geq I(X; U|Y), \quad R_1 \geq I(Y; V|U)
\]
\[
D_2 \geq \sigma^2_{X|V, U}.
\] (19)

Proof: Assume that \((R, R_1, D_2)\) is achievable. Let \(T = f^{(n)}(X^n)\) and \(T_1 = f^{(n)}_1(f^{(n)}(X^n), Y^n)\), then the rate \(R\) can be lower bounded as follows

\[
nR \geq \log M^{(n)} \geq H(T) \geq H(T|Y^n)
\]
\[
= I(X^n; T|Y^n) = \sum_{k=1}^{n} I(X_k; T|Y^n, X^{k-1})
\]
\[
\geq \sum_{k=1}^{n} I(X_k; U_k|Y_k),
\] (20)

where \(X^{k-1} \triangleq (X_1, \ldots, X_{k-1}), U_k \triangleq (T, Y^{k-1}), Y^n \backslash k \triangleq (Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n)\), and the equality (a) follows by the Markov relation \(X_k \ni Y_k \ni Y^n \backslash X^{k-1}\).

Let \(V_k \triangleq T_1\), then the rate \(R_1\) can be lower bounded as follows

\[
nR_1 \geq \log M^{(n)}_1 \geq H(T_1) \geq H(T_1|T)
\]
\[
= I(Y^n; T_1|T) = \sum_{k=1}^{n} I(Y_k; T_1|T, Y^{k-1})
\]
\[
= \sum_{k=1}^{n} I(Y_k; V_k|U_k).
\] (21)
The sum-rate can be lower bounded as follows
\[
n(R + R_1) \geq \log M(n) + \log M(n)
\]
\[
= H(T) + H(T_1|T)
\]
\[
= H(T) + H(T_1|T)
\]
\[
= H(T|Y^n) + H(T; Y^n) + H(T_1|T)
\]
\[
= H(T|Y^n) + H(T_1|T) + \sum_{k=1}^{n} \log M(k|Y^k-1)
\]
\[
(b) \sum_{k=1}^{n} \log M(k|Y^k-1) \geq H(T) + H(T_1|T)
\]
\[
= H(T) + H(T_1|T)
\]
\[
= H(T, T_1)
\]
\[
(\sigma^2_{X|Y} + \sigma^2_{W_y})^{-1}
\]

Lemma 4: In the optimization problem \( P_G \) the conditional variance \( \sigma^2_{X|Y} \) is monotonically increasing with \( \sigma^2_{X|Y} \).

Proof: By the Markov relation \( U^G \circ X \circ Y \) and (1) we may assume that
\[
U^G = X + N^{(u)}
\]
where \( N^{(u)} \sim N(0, \sigma^2_{W_y}) \) is independent of \( X \). Then a straightforward computation yields the relation
\[
\sigma^2_{X|Y} = \sigma^2_{W_y} \left( \frac{\sigma^2_{W_y}}{\sigma^2_{X|U}} + a^2_y \right)^{-1}
\]
We may express \( P_G \) as a double optimization problem
\[
\min_{\sigma^2_{X|Y}} \frac{1}{2} \log \left[ 2\pi e \sigma^2_{X|Y} \right] + \frac{1}{2} \log \left[ 2\pi e \cdot \varphi \left( P_G[D_2, (X, Y, U^G), \sigma^2_{X|Y}] \right) \right]
\]
\[
\text{subject to } \begin{align*}
R_1 &\geq \frac{1}{2} \log \frac{\sigma^2_{X|Y} + \sigma^2_{W_y}}{\sigma^2_{X|Y}} \\
R &\geq \frac{1}{2} \log \frac{\sigma^2_{X|Y} + \sigma^2_{W_y}}{\sigma^2_{X|Y}} \text{ and } U^G \circ X \circ Y
\end{align*}
\]
\[
\min_{\sigma^2_{X|Y}} \frac{1}{2} \log \left[ 2\pi e \sigma^2_{X|Y} \right] + \frac{1}{2} \log \left[ 2\pi e \cdot \varphi \left( P_G[D_2, (X, Y, U^G), \sigma^2_{X|Y}] \right) \right]
\]
\[
\text{subject to } \begin{align*}
R &\geq \frac{1}{2} \log \frac{\sigma^2_{X|Y} + \sigma^2_{W_y}}{\sigma^2_{X|Y}} \\
R_1 &\geq \frac{1}{2} \log \frac{\sigma^2_{X|Y} + \sigma^2_{W_y}}{\sigma^2_{X|Y}} \text{ and } U^G \circ X \circ Y
\end{align*}
\]
\[
\text{such that for a fixed } \sigma^2_{X|Y}, \text{ the inner optimization solution is the source-coding sub-problem } P_G[D_2, (X, Y, U^G), \sigma^2_{X|Y}]
\]
defined in Section IV.A.

From Section IV.A we know that the the optimal solution to \( P_G[D_2, (X, Y, U^G), \sigma^2_{X|Y}] \) is attained by a law \( p_{XYUV} \) such that \( (U^G, X) \circ Y \circ V \). Thus, we can assume with no loss in generality that \( p_{XYUV} \) satisfies the long Markov chain \( U^G \circ X \circ Y \circ V \). Let \( (X, Y, U^G) \) be jointly Gaussian as per (1) and (23) then the choice of \( V^G = Y + N^{(v)} \) where \( N^{(v)} \sim N(0, \sigma^2_{W_v}) \) is independent of \( (X, N^{(u)}), (N^{(u)}) \) establishes that the optimal solution to \( P_G \) achieves \( R_G(D_2) \).

REFERENCES

Upper-Bounding Rate Distortion Functions based on the Minimum Mean-Square Error

Meik Dörpinghaus*, Matthias Rüngeler**, and Rudolf Mathar*
*Institute for Theoretical Information Technology, RWTH Aachen University, 52056 Aachen, Germany, {doerpinghaus, mathar}@ti.rwth-aachen.de
**Institute for Communication Systems and Data Processing, RWTH Aachen University, 52056 Aachen, Germany, ruengeler@ind.rwth-aachen.de

Abstract—We derive a new upper bound on the rate distortion function for arbitrary memoryless sources, which is based on the relation between mutual information and minimum mean-square error discovered by Guo et al. This upper bound is in general tighter than the well known upper bound given by the rate distortion function of a Gaussian source with an equal variance found by Shannon and becomes tight for Gaussian sources. We evaluate the new upper bound for various source distributions and compare it to the Shannon lower and upper bound and to the rate distortion function calculated with the Blahut-Arimoto algorithm. This shows that the new upper bound is quite tight.

I. INTRODUCTION AND SETUP

It is well known that the rate distortion function states which rate $R$ is at least required to describe a source $X$, such that it can be reconstructed up to a given distortion $D$ for a specific distortion measure.

Let $X$ be a random source that generates independent identically distributed (i.i.d.) symbols with respect to the probability distribution function (PDF) $p(x)$, i.e., $X \sim p(x)$. When encoding, source sequences $X^n$ consisting of $n$ source symbols are mapped onto indices

$$f_n : X^n \rightarrow \{1, 2, \ldots, 2^{nR}\}.$$  

(1)

Here, $R$ is the rate of the encoded sequence. By this mapping the source sequences are represented by $nR$ bits. Based on this index representation the decoder is able to generate an estimate $\hat{X}^n$ of the source sequence $X^n$, i.e., the decoding function is

$$g_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{X}^n.$$  

(2)

The encoding and decoding given by the functions $f_n$ and $g_n$ is often referred to as a $(2^{nR}, n)$-rate distortion code.

The question is: How close can the source sequences $X^n$ be reconstructed when they are encoded with rate $R$? I.e., how large is the expected distortion $D$ between the source sequences $X^n$ and their reconstruction $\hat{X}^n$ with

$$D = \mathbb{E}[d(X^n, g_n(f_n(X^n)))].$$

where

$$d(x^n, \hat{x}^n) = d(x^n, g_n(f_n(x^n))) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i).$$

(3)

is the distortion between the source sequence $x^n$ and its reconstruction $\hat{x}^n$. In addition, $d(x_i, \hat{x}_i)$ is the distortion between the individual source symbols $x_i$ and their estimates $\hat{x}_i$. The distortion $D$ and the rate $R$ form a rate distortion pair $(R, D)$. A rate distortion pair is achievable if a sequence of $(2^{nR}, n)$-rate distortion codes $(f_n, g_n)$ exists such that

$$\lim_{n \to \infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D.$$  

(4)

The rate distortion function $R(D)$ is defined as the minimum rate $R$ that is required to encode the source $X$ for a given distortion $D$. In [1], [2] Shannon has shown that for an i.i.d. source $X$ with distribution $p(x)$ and bounded distortion function $d(x^n, \hat{x}^n)$ the rate distortion function is given by

$$R(D) = \min_{p(x|\hat{x}) \in S} I(X; \hat{X})$$

(5)

where $I(X; \hat{X})$ is the mutual information of $X$ and $\hat{X}$ and

$$S = \left\{ p(\hat{x}|x) \left| \int \int p(\hat{x}|x)p(x)d(x, \hat{x})dxd\hat{x} \leq D \right\}.$$  

(6)

see also [3, Th. 10.2.1]. Minimization in (5) is hence over all $p(\hat{x}|x)$ for which $p(x, \hat{x})$ fulfills the distortion constraint $D$.

For certain distortion measures $d(x, \hat{x})$ and source distributions $p(x)$, $R(D)$ is known. E.g., it is a well known result that for a Gaussian source $X \sim N(0, \sigma_X^2)$ and the mean-square error as distortion measure, i.e., $E[(X - \hat{X})^2] \leq D$, the rate distortion function is given by

$$R(D) = \begin{cases} \frac{1}{2} \log \left( \frac{\sigma_X^2}{\sigma_y^2} \right) & \text{for } 0 \leq D \leq \sigma_X^2, \\ 0 & \text{for } D > \sigma_X^2. \end{cases}$$

(7)

While for a few combinations of distortion measures and source distributions the rate distortion functions are known, they are in general unknown. Existing bounds like the Shannon lower bound [2] and the upper bound given by the fact that the rate distortion function for a Gaussian source in (7) is an upper bound to the rate distortion function of arbitrarily distributed sources $X$ with the same variance $\sigma_X^2$, see [1], are in general not tight. This is the motivation for the present work. We derive a new upper bound on the rate distortion function for the mean-square error as distortion measure and arbitrary source distributions. This upper bound is based on the relation between the minimum mean-square error (MMSE) and the mutual information given in [4]. Recently, in [5] this relation has already been applied in the context of rate distortion theory. However, in contrast to the present work [5] does not consider minimization over the reproduction distribution for the study of $R(D)$.

1 All logarithms are to the base $e$ and, thus, all rates are in nats.
II. New Upper Bound on $R(D)$

**Theorem 1.** Let $X$ be a random source that generates i.i.d. symbols according to $X \sim p(x)$ with $\text{var}(X) = \sigma_X^2$. The minimal required rate $R$ such that $\mathbb{E}[(X - \hat{X})^2] \leq D$ is upper-bounded by

$$R(D) \leq \frac{1}{2} \log \left( 1 + \frac{\sigma_X^2}{\sigma_N^2} \right) - \frac{1}{2} \int_0^{\gamma} \left( \frac{\sigma_X^2}{1 + \sigma_X^2} - \text{mmse}(\gamma) \right) d\gamma$$

where $\text{mmse}(\gamma)$ is the minimum mean squared error when estimating $X$ disturbed by additive Gaussian noise $N$, i.e.

$$\text{mmse}(\gamma) = \mathbb{E} \left[ (X - \mathbb{E}[X|X+N])^2 \right]$$

and where $\gamma$ is the normalized SNR, i.e., $\gamma = \frac{\sigma_X^2}{\sigma_N^2}$. Furthermore, $\sigma_{X}^2$ is related to the distortion $D$ by

$$D = \text{mmse}(1/\sigma_N^2) = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2} - 2 \frac{d}{d\alpha} D(p_{X+N} \parallel p_{X'+N}).$$

Here $D(p_{X+N} \parallel p_{X'+N})$ is the Kullback-Leibler divergence between the PDF of $X+N$ and of $X'+N$ where $X'$ is a Gaussian random variable with the same mean and variance as $X$.

Theorem 1 has also the following interpretation. The rate distortion function $R(D)$ for an arbitrarily distributed source $X$ with variance $\sigma_X^2$ is upper-bounded by the mutual information of an AWGN channel with the input $X - \mathbb{E}[X]$ and a noise variance $\sigma_N^2$, which is chosen such that the distortion is equal to $D = \text{mmse}(1/\sigma_N^2)$. As $D$, i.e., the upper bound on $R(D)$, corresponds to the mutual information of an AWGN channel, $\text{mmse}(1/\sigma_N^2)$ and hence $D$ is given by two times the derivative of (9) with respect to $1/\sigma_N^2$ yielding (12), see [4] for the relation between mutual information and MMSE.

The representations in (9) and (12) using the Kullback-Leibler divergence allow an easy numerical evaluation of the upper bound on the rate distortion function. To calculate $D(p_{X+N} \parallel p_{X'+N})$ the PDF $p_{X+N}$ is required. As $X$ and the additive noise $N$ are independent, $p_{X+N}$ is given by the convolution of $p_X$ and $p_N = N(0, \sigma_N^2)$. In case $p_{X+N}$ cannot be obtained in closed-form, a numerical convolution is required to calculate the upper bound on $R(D)$. Thus, evaluation of (9) gives an upper bound on the rate $R$ for a given $\sigma_N^2$. The corresponding distortion $D$ can be calculated from (12). Hence, to evaluate the upper bound on the rate distortion function pairs of the upper bound on $R$ given by (9) and $D$ given by (12) have to be calculated by varying the parameter $\sigma_N^2$.

III. Proof of Theorem 1

As the rate distortion function is independent of the mean $\mathbb{E}[X]$ of the source, i.e., the rate distortion functions $R(D)$ for the sources $X$ and $X - \mathbb{E}[X]$ are equal, in the following proof we assume that $\mathbb{E}[X] = 0$ without loss of generality.

With (5) the rate distortion function is given by

$$R(D) = \min_{p(\hat{x}|x)\in S} \mathcal{I}(X; \hat{X})$$

where the minimization is over all $p(\hat{x}|x)$ such that

$$\mathbb{E}[(X - \hat{X})^2] \leq D$$

i.e., all $p(\hat{x}|x)$ in the set $S$ in (6) with $d(x, \hat{x}) = (x - \hat{x})^2$. As $R(D)$ is monotonically decreasing in $D$, all $p(\hat{x}|x)$ being solutions of (13) fulfill (14) with equality. Obviously, any choice $p(\hat{x}|x) \in S$ different from the optimal $p(\hat{x}|x)$ which minimizes $\mathcal{I}(X; \hat{X})$ yields an upper bound on $R(D)$:

$$R(D) \leq \mathcal{I}(X; \hat{X}).$$

Thus, we construct an upper bound on the rate distortion function $R(D)$ by choosing some $p(\hat{x}|x)$ satisfying

$$\mathbb{E}[(X - \hat{X})^2] = D$$

for which $\mathcal{I}(X; \hat{X})$ can be easily evaluated and at the same time is a sufficiently tight upper bound. Therefore, we define

$$Y = X + N$$

where $N$ is zero-mean additive white Gaussian noise with variance $\sigma_N^2$. Furthermore, let

$$\hat{X} = \mathbb{E}[X|Y]$$

be the MMSE estimate of $X$ from $Y$. Now, the variance $\sigma_N^2$ has to be chosen such that (16) is fulfilled. As the MMSE estimate $\hat{X}$ is a sufficient statistic of $Y$, see [6], it does not change the mutual information [3, Sect. 2.9] and it holds that

$$\mathcal{I}(X; \hat{X}) = \mathcal{I}(X; Y).$$

Thus, an upper bound on $R(D)$ is given by the mutual information of an AWGN channel with the source $X$ at its input and the variance of the additive Gaussian noise $\sigma_N^2$ depending on the distortion $D$ and implicitly given by (10) and (11).

Hence, to prove Theorem 1 it remains to show that (9) is equal to $\mathcal{I}(X; Y)$, i.e., the mutual information of the AWGN channel in (17) with the source $X$ at its input and noise variance $\sigma_N^2$. Therefore, we rewrite $\mathcal{I}(X; Y)$ as follows, see [7]²:

$$\mathcal{I}(Y; X) = h(Y) - h(Y|X) = h(Y) - h(N)$$

$$= - \int p_Y(y) \log(p_Y(y)) \, dy - h(N)$$

$$= - \int p_Y(y) \log(p_Y(y)) \, dy - \int p_Y(y) \log \left( \frac{p_Y(y)}{p_{Y'}(y)} \right) \, dy \quad \text{and}$$

$$= - \int p_{Y'}(y) \log(p_{Y'}(y)) \, dy - \int p_Y(y) \log \left( \frac{p_Y(y)}{p_{Y'}(y)} \right) \, dy \quad \text{and}$$

$$= h(N) - h(Y) - D(p_Y \parallel p_{Y'}) - h(N)$$

²Note that [7] states that inequality (8) in Theorem 1 holds with equality. However, this statement is wrong in general. The erroneous assumption in [7] is that the information rate of a system with optimal source and channel coding is equal to the information rate of an AWGN channel where the source symbols are transmitted without coding. It can be shown that this statement does not hold in general and, therefore, (8) does not hold with equality in general.
where $h(\cdot)$ denotes the differential entropy, $p_Y(y)$ is the PDF of $Y$, and $p_Y(y)$ is the PDF of a corresponding zero-mean Gaussian variable with the same variance
\[ \sigma_Y^2 = \sigma_X^2 + \sigma_N^2. \] (22)
Moreover, $D(p_Y \parallel p_Y)$ is the Kullback-Leibler divergence between $p_Y(y)$ and $p_Y$. Furthermore, (20) holds because
\[- \int p_Y(y) \log(p_Y(y)) \, dy = - \int p_Y(y) \log \left( \frac{\exp \left( \frac{-y^2}{2\sigma_Y^2} \right)}{\sqrt{2\pi\sigma_Y^2}} \right) \, dy \]
\[ = \frac{1}{2} \log \left( 2\pi e \sigma_Y^2 \right) = - \int p_Y(y) \log(p_Y(y)) \, dy. \] (23)
As $Y$ and $N$ are Gaussian, (21) is given by
\[ I(Y;X) = \frac{1}{2} \log \left( 2\pi e \sigma_X^2 \right) + D(p_{X|Y} \parallel p_{X|Y}) \]
\[ = \frac{1}{2} \log \left( 1 + \frac{\sigma_Y^2}{\sigma_X^2} \right) - D(p_{X+N} \parallel p_{X+N}) \] (24)
where for (24) we have used (22). This shows that (9) is the mutual information of an AWGN channel with noise variance $\sigma_Y^2$ and an arbitrarily distributed source $X$ with zero-mean and variance $\sigma_X^2$ at its input.

The alternative representation of the upper bound on $R(D)$ in (8) holds, as $D(p_Y \parallel p_Y)$ in (24) is given by
\[ D(p_Y \parallel p_Y) = h(Y) - h(Y) \] (25)
\[ = I(Y';Y) + h(Y'|X') - I(Y;X) = h(X' + N|X') - h(X + N|X) \] (26)
\[ = \frac{1}{2} \int_0^\infty \frac{\gamma}{1 + \sigma_X^2 \gamma} \, d\gamma \] (27)
where $Y'$ is a zero-mean Gaussian random variable with variance $\sigma_X^2$. If $X'$ is the input to the AWGN channel in (17), this results in the zero-mean Gaussian output $Y'$, yielding (26). Moreover, (25) and (27) follow from the results given in [8].

IV. PROPERTIES OF THE UPPER BOUND

A. Gaussian Sources

For a Gaussian source $X$, $\text{mmse}(\gamma)$ in the upper bound (8) becomes $\sigma_X^2/(1 + \sigma_X^2 \gamma)$ and, thus, the integral in (8) is zero. Hence, the upper bound of the AWGN channel. However, for a Gaussian source the capacity of the AWGN channel and the rate distortion function $R(D)$ in (7) are equal, as (11) becomes equal to
\[ D = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}. \] (28)
which follows from (12) as the Kullback-Leibler divergence in (12) is zero for Gaussian inputs. Thus, for a Gaussian source the upper bound on $R(D)$ in (8) and (9) is tight.

This discussion also shows another well known fact. In case we estimate the Gaussian source $X$ after transmission over an AWGN channel based on $Y$ using an MMSE estimator, the MMSE is given by $D$ in (28), see (10) to (12). To achieve this distortion no code at all is required. On the other hand, we can source encode the source $X$ using a rate-distortion code, then use a channel code to transmit it reliably over the AWGN channel. Following Shannon’s source channel coding separation theorem [1], this separation of source and channel coding is optimal. At the receiver the signal can be channel and source decoded. To achieve a maximum distortion $D$, the source encoder has to encode the source with rate $R(D)$ given in (7). For reliable transmission over the AWGN channel its capacity has to correspond to $R(D)$, i.e., for $0 \leq D \leq \sigma_X^2$
\[ \frac{1}{2} \log \left( 1 + \frac{P}{\sigma_N^2} \right) \leq R(D) \leq \frac{1}{2} \log \left( 1 + \frac{\sigma_X^2}{\sigma_N^2} \right) \] (29)
where $P$ is the required transmit power. Obviously $P = \sigma_X^2$.

However, this means that for a given transmit power, there is no advantage of the use of source and channel coding when a Gaussian source has to be transmitted over an AWGN channel. Directly transmitting uncoded source symbols over the AWGN channel yields the same distortion $D$ as perfect source and channel coding. In addition, the latter one implies an infinite delay. This is a well known result, see e.g. [9], [10].

B. Gap to Shannon’s Lower Bound

Shannon gave a lower bound on $R(D)$, which for the mean-square error distortion measure becomes [2]
\[ R(D) \geq h(X) - \frac{1}{2} \log(2\pi e D) = \text{LB}_{\text{Sha}}. \] (30)

The gap $\Delta$ between the new upper bound on $R(D)$ given in Theorem 1 and Shannon’s lower bound in (30) is given by
\[ \Delta = \text{UB}_{\text{new}} - \text{LB}_{\text{Sha}} = -h(X|Y) + \frac{1}{2} \log \left( 2\pi e \text{mmse} \left( \frac{1}{\sigma_N^2} \right) \right) \] (31)
where we have used (9), (30), (11), and (24). The variance of the MMSE estimate $E[X|Y]$ is equal to $\text{mmse}(1/\sigma_X^2)$. Furthermore, for a given variance Gaussian random variables are entropy maximizers and the second term on the RHS of (31) corresponds to the differential entropy of a Gaussian random variable with variance $\text{mmse}(1/\sigma_X^2)$. In general, the estimation error of the MMSE estimate $E[X|Y]$ is non-Gaussian. Thus, $h(X|Y)$ in (31) is upper-bounded by the second term on the RHS of (31). $\Delta$ corresponds to the negative difference of the entropy of $X$ conditioned on $Y$ and the corresponding entropy in case $X$ is Gaussian with the same variance. Hence, $\Delta$ becomes zero for $X$ being Gaussian.

C. Comparison of the New Upper Bound on $R(D)$ with Shannon’s Upper Bound on the Rate Distortion Function

Shannon has shown that for an arbitrary distributed source $X$ with variance $\sigma_X^2$, the rate distortion function is upper-bounded by the rate distortion function of a Gaussian source with the same variance, which is given in (7). I.e., Gaussian sources are hardest to encode, cf. [11].

The difference between this upper bound in (7) given by Shannon and the new upper bound on $R(D)$ stated in
Theorem 1 is for $D \leq \sigma_X^2$ given by

$$\text{UB}_{\text{Sh}} - \text{UB}_{\text{new}} = \frac{1}{2} \log \left( \frac{\sigma_X^2}{\text{mmse} \left( \frac{1}{\sigma_N^2} \right)} \right) - \frac{1}{2} \int_0^{1} \text{mmse}(\gamma) d\gamma$$

(32)

where we have used (7), (11), and (8). The difference on the RHS of (32) depends on mmse($\gamma$), which is bounded by $0 \leq \text{mmse} \left( \frac{1}{\sigma_N^2} \right) \leq \frac{\sigma_X^2}{1+\sigma_X^2}$, where the RHS is the MMSE in case $X$ is Gaussian. The upper bound on mmse($\gamma$) holds as Gaussian random variables are hardest to estimate [12, Prop. 15].

$\text{UB}_{\text{new}}$ is in general tighter than $\text{UB}_{\text{Sh}}$, as, by using mmse($\gamma$) $\leq \frac{\sigma_X^2}{1+\sigma_X^2}$ and the relation between the MMSE and the mutual information given in [4], (32) is lower-bounded by

$$\frac{1}{2} \log \left( 1 + \frac{\sigma_X^2}{\sigma_N^2} \right) - I(X + N; X) \geq 0$$

(33)

where the RHS of (33) follows from the fact that $I(X + N; X)$ is maximized if $X$ is Gaussian. For mmse($\gamma$) $\geq \frac{\sigma_X^2}{(1+\sigma_X^2)\gamma}$ (32) is equal to zero.

Finally, for an arbitrary distribution of the source $X$ and $D \rightarrow \sigma_X^2$, which by (11) corresponds to $\sigma_X^2 \rightarrow \infty$, the MMSE converges to the case where the source is Gaussian:

$$\lim_{1/\sigma_X^2 \rightarrow 0} \left\{ \text{mmse}(\gamma) - \frac{\sigma_X^2}{1+\sigma_X^2} \right\} = 0.$$  

(34)

Thus, $\text{UB}_{\text{Sh}} \rightarrow \text{UB}_{\text{new}}$ converges to zero for $D \rightarrow \sigma_X^2$.

D. Evaluation of the Upper Bound in Theorem 1

In Fig. 1 we have evaluated the new upper bound on the rate distortion function given by Theorem 1 for a uniform source distribution (Fig. 1(a)), a source with Laplace distribution and a bipolar input distribution (both Fig. 1(b)) all with zero-mean and variance $\sigma_X^2 = 1$, i.e.,

$$P_X(\text{uniform})(x) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } |x| \leq \sqrt{3}, \\ 0 & \text{otherwise}, \end{cases}$$

(35)

$$P_X(\text{Laplace})(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} |x| \right),$$

(36)

$$P_X(\text{bipolar})(x) = \begin{cases} \frac{1}{2} & \text{for } x = \pm 1, \\ 0 & \text{otherwise}. \end{cases}$$

(37)

For comparison the Shannon upper bound in (7) corresponding to the rate distortion function for a Gaussian source, the Shannon lower bound in (30), and the actual rate distortion function calculated numerically based on the Blahut-Arimoto algorithm [13], [14] are shown. The Shannon lower bound (30) is not shown for the bipolar source as $h(X)$ does not exist in this case.

Fig. 1 shows that the new upper bound and the Shannon lower bound on the rate distortion function become tight for $D \rightarrow 0$. On the other hand, for $D \rightarrow \sigma_X^2$ the new upper bound converges to the Shannon upper bound. The comparison of the new upper bound with the numerical evaluation of $R(D)$ based on the Blahut-Arimoto shows that the new upper bound is quite tight for the displayed source distributions.

The MATLAB source code is available on our websites.

REFERENCES


On the Information Loss in Memoryless Systems: 
The Multivariate Case

Bernhard C. Geiger∗, Gernot Kubin∗
∗Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{geiger,gernot.kubin}@tugraz.at

Abstract—In this work we give a concise definition of information loss from a system-theoretic point of view. Based on this definition, we analyze the information loss in memoryless input-output systems subject to a continuous-valued input. For a certain class of multiple-input, multiple-output systems the information loss is quantified. An interpretation of this loss is accompanied by upper bounds which are simple to evaluate.

Finally, a class of systems is identified for which the information loss is necessarily infinite. Quantizers and limiters are shown to belong to this class.

I. INTRODUCTION

In the XXXI. Shannon lecture Han argued that information theory links information-theoretic quantities, such as entropy and mutual information, to operational quantities such as source, channel, capacity, and error probability [1]. In this work we try to make a new link to an operational quantity not mentioned by Han: information loss. Information can be lost, on the one hand, in erasures or due to superposition of noise as it is known from communication theory. Dating back to Shannon [2] this loss is linked to the conditional entropy of the input given the output, at least in discrete-amplitude, memoryless settings. On the other hand, as stated by the data processing inequality (DPI, [3]), information can be lost in deterministic, noiseless systems. It is this kind of loss that we will treat in this work, and we will show that it makes sense to link it to the same information-theoretic quantity.

The information loss in input-output systems is very sparsely covered in the literature. Aside from the DPI for discrete random variables (RV) and static systems, some results are available for jointly stationary stochastic processes [4]. Yet, all these results just state that information is lost, without quantifying this loss. Only in [5] the information lost by collapsing states of a discrete-valued stochastic process is quantified as the difference between the entropy rates at the input and the output of the memoryless system.

Conversely, energy loss in input-output systems has been deeply analyzed, leading to meaningful definitions of transfer functions and notions of passivity, stability, and losslessness. Essentially, it is our aim to develop a system theory not from an energetic, but from an information-theoretic point of view. So far we analyzed the information loss of discrete-valued stationary stochastic processes in finite-dimensional dynamical input-output systems [6], where we proposed an upper bound on the information loss and identified a class of information-preserving systems (the information-theoretic counterpart to lossless systems). In [7] the information loss of continuous RVs in memoryless systems was quantified and bounded in a preliminary way. In this work, extending [7], we analyze the information loss for static multiple-input, multiple-output systems which are subject to a continuous input RV. Unlike in our previous work, we permit functions which lose an infinite amount of information and present the according conditions. Aside from that we provide a link between information loss and differential entropy, a quantity which is not invariant under changes of variables. The next steps towards an information-centered system theory are the analysis of discrete-time dynamical systems with continuous-valued stationary input processes and a treatment of information loss in multirate systems.

In the remainder of this paper we give a mathematically concise definition of information loss (Section II). After restricting the class of systems in Section III, in Section IV we provide exact results for information loss together with simple bounds, and establish a link to differential entropies. Finally, in Section V we show under which conditions the information loss becomes infinite. An extended version of this paper, accompanied by several examples illustrating the theoretical results, is available in [8].

II. A DEFINITION OF INFORMATION LOSS

When talking about the information loss induced by processing of signals, it is of prime importance to accompany this discussion by a well-based definition of information loss going beyond, but without lacking, intuition. We try to meet this objective with the following

Definition 1. Let X be an RV1 on the sample space $\mathcal{X}$, and let $\hat{Y}$ be obtained by transforming $X$. We define the information loss induced by this transform as

$$L(X \rightarrow \hat{Y}) = \sup_{\mathcal{P}} \left( I(\hat{X}; X) - I(\hat{X}; \hat{Y}) \right)$$

(1)

where the supremum is over all partitions $\mathcal{P}$ of $\mathcal{X}$, and where $\hat{X}$ is obtained by quantizing $X$ according to the partition $\mathcal{P}$ (see Fig. 1).

This Definition is motivated by the data processing inequality (cf. [3]), which states that the expression under the supremum is always non-negative: Information loss is the worst-case reduction of information about $\hat{X}$ induced by

1Note that $X$ and all other involved RVs need not be scalar-valued.
transferring $X$. We now try to shed a little more light on Definition 1 in the following

**Theorem 1.** The information loss of Definition 1 is given by the conditional entropy of the input given the output, i.e.,

$$L(X \rightarrow Y) = H(X|Y).$$

(*Proof:* We start by noticing that

$$I(\hat{X}; X) - I(\hat{X}; Y) = H(\hat{X}|Y) \leq H(X|Y)$$

by the definition of mutual information and since both $\hat{X}$ and $Y$ are functions of $X$. The inequality in (3) is due to data processing ($\hat{X}$ is a function of $X$). We now show that in the supremum over all partitions equality can be achieved.

To this end, observe that among all partitions of the sample space of $X$ there is a sequence $\{\mathcal{P}_n\}$ of increasingly fine partitions such that

$$\lim_{n \to \infty} \hat{X}_n = X$$

where $\hat{X}_n$ is the quantization of $X$ induced by partition $\mathcal{P}_n$. By the axioms of entropy (e.g., [9, Ch. 14]), $H(X_n|Y)$ is an increasing sequence in $n$ with limit $H(X|Y)$. Thus, equality in (3) is achieved, which completes the proof. \hfill $\blacksquare$

This Theorem shows that the supremum in Definition 1 is achieved for $\hat{X} \equiv X$, i.e., when we compute the difference between the self-information of the input and the information the output of the system contains about its input. In addition to that, the Theorem suggests a natural way to measure the information loss via measuring mutual information, as it is depicted in Fig. 1. As we will see later (cf. Theorem 3), the considered partition does not have to be even infinitely fine.

**III. PROBLEM STATEMENT**

Let $X = [X_1, X_2, \ldots, X_N]$ be an $N$-dimensional RV with a probability measure $P_X$ absolutely continuous w.r.t. the Lebesgue measure $\mu$ ($P_X \ll \mu$). We require $P_X$ to be concentrated on $\mathcal{X} \subseteq \mathbb{R}^N$. This RV, which possesses a unique probability density function (PDF) $f_X$, is the input to the following multivariate, vector-valued function:

**Definition 2.** Let $g: \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N$, be a surjective, Borel-measurable function defined in a piecewise manner:

$$g(x) = \begin{cases} g_1(x), & \text{if } x \in \mathcal{X}_1 \\ g_2(x), & \text{if } x \in \mathcal{X}_2 \\ \vdots \end{cases}$$

where $x = [x_1, x_2, \ldots, x_N]$ and $g_i: \mathcal{X}_i \rightarrow \mathcal{Y}_i$ bijectively. Furthermore, let the Jacobian matrix $J_{g_i}()$ exist on the closures of $\mathcal{X}_i$. In addition to that, we require the Jacobian determinant, $|\det J_{g_i}()|$, to be non-zero $P_X$-almost everywhere.

In accordance with previous work [7] the $\mathcal{X}_i$ are disjoint sets of positive $P_X$-measure which unite to $\mathcal{X}$, i.e., $\bigcup_i \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ if $i \neq j$. Clearly, also the $\mathcal{Y}_i$ unite to $\mathcal{Y}$, but need not be disjoint. This definition ensures that the preimage $g^{-1}[y]$ of each element $y \in \mathcal{Y}$ is a countable set.

Using the method of transformation [9, pp. 244] one obtains the PDF of the $N$-dimensional output RV $Y = [Y_1, Y_2, \ldots, Y_N]$ as

$$f_Y(y) = \sum_{x \in g^{-1}[y]} \frac{f_X(x)}{|\det J_{g_i}(x)|}$$

where the sum is over all elements of the preimage. Note that since $Y$ possesses a density, the corresponding probability measure $P_Y$ is also absolutely continuous w.r.t. the Lebesgue measure.

**IV. MAIN RESULTS**

We now state our main results:

**Theorem 2.** The information loss induced by a function $g$, satisfying Definition 2 is given as

$$H(X|Y) = \int_X f_X(x) \log \left( \frac{\sum_{y \in g^{-1}[g(x)]} f_X(x)}{|\det J_{g_i}(x)|} \right) dx.$$  

(*Proof:* The proof can be found in the Appendix of [8] and, in a modified version for univariate functions, in [7]. Note that for univariate functions the Jacobian determinant is replaced by the derivative of the function. \hfill $\blacksquare$

**Corollary 1.** The information loss induced by a function $g$, satisfying Definition 2 is given as

$$H(X|Y) = h(X) - h(Y) + E \{ \log |\det J_{g_i}(X)| \}$$

(*Proof:* The proof is obtained by recognizing the PDF of $Y$ inside the logarithm in (7) and by splitting the logarithm. \hfill $\blacksquare$

This result is particularly interesting because it provides a link between information loss and differential entropies already anticipated in [9, pp. 660]. There, it was claimed that

$$h(Y) \leq h(X) + E \{ \log |\det J_{g_i}(X)| \}$$

where equality holds iff $g$ is bijective. While (9) is actually another version of the DPI, Corollary 1 quantifies how much information is lost by processing. In addition to that, a very similar expression denoted as folding entropy has been presented in [10], although in a completely different setting analyzing the entropy production of autonomous dynamical systems.

We now introduce a discrete RV $W$ which depends on the set $\mathcal{X}_i$ from which $X$ was taken. In other words, we have $W =$

\[ \]
One can interpret this RV as being generated by a vector quantization of $X$ with a partition $P = \{X_i\}$. With this new RV we can state

**Theorem 3.** The information loss is identical to the uncertainty about the set $X_i$ from which the input was taken, i.e.,

$$H(X|Y) = H(W|Y). \quad (10)$$

The proof closely follows the proof provided in [7] and thus is omitted. However, this equivalence suggests a way of measuring information loss by means of proper quantization: Since $H(W|Y) = I(W;X) - I(W;Y)$ the loss can be determined by measuring mutual informations, which in this case are always finite (or, at least, bounded by $H(W)$).

The interpretation derived from Theorem 3 allows us now to provide upper bounds on the information loss:

**Theorem 4.** The information loss is upper bounded by

$$H(X|Y) \leq \int_Y f_Y(y) \log \text{card}(g^{-1}[y]) dy \quad (11)$$

$$\leq \log \left( \sum_i \int_{Y_i} f_Y(y) dy \right) \quad (12)$$

$$\leq \max_y \log \text{card}(g^{-1}[y]) \quad (13)$$

where $\text{card}(B)$ is the cardinality of the set $B$.

**Proof:** We give here only a sketch of the proof: The first inequality results from bounding $H(W|Y = y)$ by the entropy of a uniform distribution on the preimage of $y$. Jensen’s inequality yields the second line of the Theorem. The coarsest bound is obtained by replacing the cardinality of the preimage by its maximal value.

In this Theorem, we bounded the information loss given a certain output by the cardinality of the preimage. While the first bound considers the fact that the cardinality may actually depend on the output itself, the last bound incorporates the maximum cardinality only. In cases where the function from Definition 2 is defined not on a countable but on a finite number of subdomains, this finite number can act as an upper bound (cf. [7]). Another straightforward upper bound, which is independent from the bounds in Theorem 4 is obtained from Theorem 3 by removing conditioning:

$$H(X|Y) \leq H(W) = - \sum_i p_i \log p_i \quad (14)$$

where $p_i = P_X(X_i) = \int_{X_i} f_X(x) dx$.

A further implication of introducing this discrete RV $W$ is that it allows us to perform investigations about reconstructing the input from the output. Currently, a Fano-type inequality bounding the reconstruction error by the information loss is under development. In addition to that, new upper bounds on the information loss related to the reconstruction error of optimal (in the *maximum a posteriori* sense) and of simpler, sub-optimal estimators are analyzed.

V. FUNCTIONS WITH INFINITE INFORMATION LOSS

We now drop the requirement in Definition 2 that the $g$ possesses partial inverses to analyze a wider class of Borel-measurable functions $g: X \to Y$. We keep the requirement that $P_X \ll \mu$ and thus $X$ possesses a density $f_X$ (positive on $X$ and zero elsewhere). We maintain

**Theorem 5.** Let $g: X \to Y$ be a Borel-measurable function and let the continuous RV $X$ be the input to this function. If there exists a set $B \subseteq Y$ of positive $P_Y$-measure such that the preimage $g^{-1}[y]$ is uncountable for every $y \in B$, then the information loss is infinite.

**Proof:** With $B \subseteq Y$ we notice that

$$H(X|Y) \geq \int_B H(X|Y = y) dP_Y(y) \quad (15)$$

where the integral is now written as Lebesgue integral, since $P_Y$ now not necessarily possesses a density.

Since on $B$ the preimage of every element is uncountable, we obtain with [4] and the references therein $H(X|Y = y) = \infty$ for all $y \in B$, and, thus, $H(X|Y) = \infty$.

Note that the requirement of $B$ being a set of positive $P_Y$-measure cannot be dropped, as one of the examples provided in [8] illustrates. We immediately obtain the following

**Corollary 2.** Let $g: X \to Y$ be a Borel-measurable function and let the continuous RV $X$ be the input to this function. If the probability measure of the output, $P_Y$, possesses a non-vanishing discrete component, the information loss is infinite.

**Proof:** According to the Lebesgue-Radon-Nikodym theorem [11, pp. 121] every measure can be decomposed in a component absolutely continuous w.r.t. $\mu$ and a component singular to $\mu$. The latter part can further be decomposed into a singular continuous and a discrete part, where the latter places positive $P_Y$-mass on points. Let $y^*$ be such a point, i.e., $P_Y(y^*) > 0$. As an immediate consequence, $P_X(g^{-1}[y^*]) > 0$, which is only possible if $g^{-1}[y^*]$ is uncountable ($P_X \ll \mu$).

This result is also in accordance with intuition, as the analysis of a simple quantizer shows: While the entropy of the input RV is infinite ($I(X:Y) = \infty$ for $X \to Y$; cf. [9, pp. 654]), the quantized output can contain only a finite amount of information ($I(\hat{X}:Y) = H(Y) < \infty$). In addition to that, the preimage of each possible output value $y$ is a set of positive $P_X$-measure. The loss, as a consequence, is infinite.

While for the quantizer the preimage of each possible output value is a set of positive measure, there certainly are functions for which some outputs have a countable preimage and some whose preimage is a non-null set. An example of such a system is the limiter [9, Ex. 5-4]. For such systems it can be shown that both the information loss $I(X \to Y) = H(X|Y)$ and the information transfer $I(X;Y)$ are infinite.

Finally, there exist functions $g$ for which the preimages of all output values $y$ are null sets, but which still fulfill the conditions of Theorem 5. Functions which project $X$ on a lower-dimensional subspace of $\mathbb{R}^N$ fall into that category.
VI. EXAMPLE: A TWO-DIMENSIONAL TRANSFORM WITH FINE INFORMATION LOSS

In this Section we illustrate our theoretical results with the help of an example. The transform is taken to be 2-dimensional. Let \( X \) be uniformly distributed on the square \( X = [-a, a] \times [-a, a] \), or equivalently, the two constituting RVs \( X_1 \) and \( X_2 \) are independent and uniformly distributed on \( [-a, a] \). In other words, \( f_X(x) = 1/4a^2 \) for all \( x \in X \), we have \( f_X(x_i) = 1/2a \) for \( x_i \in [-a, a] \) and \( i = 1, 2 \).

We consider a function \( g \) performing the mapping:
\[
Y_1 = X_1 \\
Y_2 = |X_1 - X_2|
\]

The corresponding Jacobian matrix is a triangular matrix
\[
J_g(x) = \begin{bmatrix}
\frac{1}{\text{sgn}(x_1 - x_2)} & 0 \\
\text{sgn}(x_2 - x_1) & \frac{1}{\text{sgn}(x_2 - x_1)}
\end{bmatrix}
\]

where \( \text{sgn}() \) is the sign-function. From this immediately follows that the magnitude of the determinant of the Jacobian matrix is unity for all possible values of \( X \), i.e., \( |\det J_g(x)| = 1 \) for all \( x \in X \). The subsets of \( X \) on which the partitioned functions \( g_i \) are bijective are no intervals in this case; they are the triangular halves of the square induced by \( x_1 = x_2 \) (see Fig. 2).

The preimage of \( g(x) \) is, in any case,
\[
\{[x_1, x_2], [x_1, 2x_1 - x_2] \cap X\}.
\]

The transform \( g \) is bijective whenever \( [x_1, 2x_1 - x_2] \notin X \), i.e., if \( |2x_1 - x_2| > a \).

With the PDF of \( X \) and of its components we obtain for the information loss
\[
H(X|Y) = \int_{-a}^{a} \int_{-a}^{a} \frac{1}{4a^2} \log \left( \frac{\frac{1}{2a} + \frac{\sqrt{4a^2 - (2x_1 - x_2)^2}}{2a}}{\sqrt{4a^2 - (2x_1 - x_2)^2}} \right) dx_1 dx_2
\]

which is non-zero only if \(-a \leq 2x_1 - x_2 \leq a\) (numerator and denominator cancel otherwise; no loss occurs in the bijective domain of the function). As a consequence,
\[
H(X|Y) = \int_{-a}^{a} \int_{-a}^{a} \frac{\log 2}{4a^2} dx_1 dx_2 = \int_{-a}^{a} \frac{1}{4a} dx_2 = \frac{1}{2}
\]

The information loss is identical to a half bit. This is intuitive by looking at Fig. 2, where it can be seen that any information loss occurs only on one half of the domain \( X \) (shaded in stronger colors). By destroying the sign information, in this area the information loss is equal to one bit.

VII. CONCLUSION

In this work, we proposed a mathematically concise definition of information loss for the purpose of establishing a system theory from an information-theoretic point of view. For a certain class of multivariate, vector-valued functions and continuous input variables this information loss was quantified, and the result is accompanied by convenient upper bounds. We further showed a connection between information loss and the differential entropies of the input and output variables.

Finally, a class of systems has been identified for which the information loss is infinite. Vector-quantizers and limiters belong to that class, but also functions which project the input space onto a space of lower dimensionality.

REFERENCES

One-Bit Quantizers for Fading Channels

Tobias Koch
University of Cambridge
Cambridge, CB2 1PZ, UK
Email: tobi.koch@eng.cam.ac.uk

Amos Lapidoth
ETH Zurich
CH-8092 Zurich, Switzerland
Email: lapidoth@ist.ee.ethz.ch

Abstract—We study channel capacity when a one-bit quantizer is employed at the output of the discrete-time average-power-limited Rayleigh-fading channel. We focus on the low signal-to-noise ratio regime, where communication at very low spectral efficiencies takes place, as in Spread Spectrum and Ultra-Wideband communications. We demonstrate that, in this regime, the best one-bit quantizer does not reduce the asymptotic capacity of the coherent channel, but it does reduce that of the noncoherent channel.

I. INTRODUCTION

We study the effect on channel capacity of quantizing the output of the discrete-time average-power-limited Rayleigh-fading channel using a one-bit quantizer. This problem arises in communication systems where the receiver uses digital signal processing techniques, which require the analog received signal to be quantized using an analog-to-digital converter (ADC). The effects of quantization are particularly pronounced when high-resolution ADCs are not practical and low-resolution ADCs must be used [1].

We focus on the low signal-to-noise ratio (SNR) regime, where communication at very low spectral efficiencies takes place (as in Spread-Spectrum and Ultra-Wideband communications). For the average-power-limited real-valued Gaussian channel, it is well-known that, in this regime, a symmetric one-bit quantizer (which produces 1 if the channel output is nonnegative and 0 otherwise) reduces the capacity by a factor of 2/π, corresponding to a 2dB power loss [2]. It was recently shown that, by allowing for asymmetric one-bit quantizers with corresponding asymmetric signal constellations, these two decibels can be recovered in full [3]. A similar result was shown for the average-power-limited complex-valued Gaussian channel [4]: using binary on-off keying and a radial quantizer (which produces 1 if the magnitude of the channel output is above some threshold and 0 otherwise), one can achieve the low-SNR asymptotic capacity of the unquantized channel by judiciously choosing the threshold and the on-level as functions of the SNR. Here we extend [3], [4] to Rayleigh-fading channels. Specifically, we study the capacity per unit-energy [5] of such channels when the channel output is quantized using a one-bit quantizer.

For coherent fading channels, where the receiver has perfect channel knowledge, we show that quantizing the channel output with a one-bit quantizer causes no loss in the capacity per unit-energy. As in [4], the capacity per unit-energy can be achieved using binary on-off keying and a radial quantizer by choosing the threshold as a function of the SNR and the fading, with the threshold and the on-level both tending to infinity as the SNR tends to zero. This result might mislead one to think that quantizing the channel output with a one-bit quantizer causes no loss in the capacity per unit-energy also for noncoherent fading channels, where the receiver does not have perfect channel knowledge. Indeed, in the absence of a quantizer the capacity per unit-energy does not depend on whether the receiver has perfect channel knowledge or not [6], [7]. Since this capacity per unit-energy can be achieved using binary on-off keying with diverging on-level, it might therefore seem plausible that also in the presence of a quantizer the capacity per unit-energy would not depend on whether the receiver has perfect channel knowledge or not. But this is not the case: in contrast to the coherent case, quantizing the output of the noncoherent Rayleigh-fading channel with a one-bit quantizer reduces the capacity per unit-energy.

The rest of the paper is organized as follows. Section II describes the channel model and introduces the capacity per unit-energy. Section III presents the main results. Section IV discusses the capacity per unit-energy when the real and the imaginary part of the channel output are quantized separately with one-bit quantizers. And Section V presents the proofs of the main results.

II. CHANNEL MODEL AND CAPACITY PER UNIT-ENERGY

We consider a discrete-time Rayleigh-fading channel whose complex-valued output \( \hat{Y}_k \) at time \( k \in \mathbb{Z} \) corresponding to the channel input \( x_k \in \mathbb{C} \) (where \( \mathbb{C} \) and \( \mathbb{Z} \) denote the set of complex numbers and the set of integers) is given by

\[
\hat{Y}_k = H_k x_k + Z_k, \quad k \in \mathbb{Z}.
\]  

(1)

Here \( \{Z_k, k \in \mathbb{Z}\} \) and \( \{H_k, k \in \mathbb{Z}\} \) are independent sequences of independent and identically distributed (i.i.d.), zero-mean, circularly-symmetric, complex Gaussian random variables, the former with unit variance and the latter with variance \( \sigma^2 \). We say that the channel is coherent if the receiver is cognizant of the realization of \( \{H_k, k \in \mathbb{Z}\} \) and that it is noncoherent if the receiver is cognizant only of the statistics of \( \{H_k, k \in \mathbb{Z}\} \).

The receiver does not have access to the channel outputs \( \{\hat{Y}_k, k \in \mathbb{Z}\} \) but only to a quantized version thereof. Specifically, the complex channel output \( \hat{Y}_k \) is fed to a one-bit
quantizer which produces $Y_k = 1$ if $\tilde{Y}_k$ is in the quantization region $\mathcal{D}$ and $Y_k = 0$ otherwise, for some Borel set $\mathcal{D} \subset \mathbb{C}$. In the coherent case, $\mathcal{D}$ may depend on the fading $\{H_k, k \in \mathbb{Z}\}$.

We assume that the average power of the channel inputs is limited by $P$. The capacity of the above channel is [8, 9]

$$C(P) = \sup_{\|X\| \leq P} I(X;Y|H), \quad \text{coherent case}$$  \hspace{1cm} (2)

$$C(P) = \sup_{\|X\| \leq P} I(X;Y), \quad \text{noncoherent case}$$  \hspace{1cm} (3)

where the suprema on the right-hand side (RHS) of (2) and (3) are over all distributions on $X$ satisfying $\mathbb{E}[\|X\|^2] \leq P$ and over all quantization regions $\mathcal{D}$. (Since the above channel is memoryless, we omit the time indices.)

The capacity per unit-energy is given by [5, Th. 2]

$$\hat{C}(0) = \frac{\sup_{P>0} C(P)}{P}$$  \hspace{1cm} (4)

It can be shown that

$$\hat{C}(0) = \lim_{P \to 0} \frac{C(P)}{P}$$  \hspace{1cm} (5)

Thus, the capacity per unit-energy is equal to the slope at zero of the capacity-vs-power curve. It can be further shown that [5, Th. 3] (see also [6])

$$\hat{C}(0) = \sup_{\xi \neq 0, \mathcal{D}} \frac{D(P_{Y|H,X=\xi} \| P_{Y|H,X=0} \!| \! P_H)}{\|\xi\|^2}$$  \hspace{1cm} (6)

in the coherent case and

$$\hat{C}(0) = \sup_{\xi \neq 0, \mathcal{D}} \frac{D(P_{Y|X=\xi} \| P_{Y|X=0} \!| \! P_H)}{\|\xi\|^2}$$  \hspace{1cm} (7)

in the noncoherent case. Here $D(\|\cdot\|)$ denotes relative entropy

$$D(P||Q) = \begin{cases} \int \log \left( \frac{dP}{dQ} \right) dP, & \text{if } P \ll Q \\ \infty, & \text{otherwise} \end{cases}$$

(where $P \ll Q$ indicates that $P$ is absolutely continuous with respect to $Q$); $D(\|\cdot\|)$ denotes conditional relative entropy

$$D(P_{Y|H,X=\xi} \!| \! P_H) = \int D(P_{Y|H=h,X=\xi} \!| \! P_{Y|H=h,X=0}) dP_H(h);$$

$P_H$ denotes the distribution of the fading $H$; $P_{Y|X=x}$ denotes the output distribution given that the input is $x$; and $P_{Y|H=h,X=x}$ denotes the output distribution conditioned on $(H,X) = (h,x)$.

By the Data Processing Inequality [10, Th. 2.8.1], the capacity per unit-energy of the quantized channel is upper-bounded by that of the unquantized channel [7, 6]

$$\hat{C}(0) \leq \frac{1}{\sigma^2}. \hspace{1cm} (8)$$

We show that in the coherent case this upper bound holds with equality, while in the noncoherent case it is strict.

III. MAIN RESULT

We restrict ourselves to radial quantizers, for which

$$\mathcal{D} = \{ \tilde{y} \in \mathbb{C} : |\tilde{y}| \geq T \}, \quad \text{for some } T > 0. \hspace{1cm} (9)$$

In the noncoherent case—as we show in Section V-B—such quantizers are optimal in the sense that they maximize the relative entropy on the RHS of (7) for every $\xi \neq 0$. In the coherent case such quantizers need not be optimal in the above sense. However, they suffice to achieve the capacity per unit-energy. And such quantizers have the practical advantage of not requiring knowledge of the phase of $\tilde{y}$.

**Theorem 1:** Consider the above channel model, and assume that the channel output is quantized using a one-bit quantizer.

1) In the coherent case,

$$\hat{C}(0) = \frac{1}{\sigma^2}$$  \hspace{1cm} (10)

which can be achieved by some radial quantizer (9) with $T$ depending on $H$ and $\xi^2$.

2) In the noncoherent case,

$$\hat{C}(0) < \frac{1}{\sigma^2}$$  \hspace{1cm} (11)

with the inequality being strict.

Proof: See Section V.

■

IV. QUANTIZING THE REAL AND IMAGINARY PART

Instead of quantizing $\tilde{Y}$ using a one-bit quantizer, often the real and imaginary parts of $\tilde{Y}$ are quantized separately using a one-bit quantizer for each; see, e.g., [11]–[14], [6]. Thus, the first quantizer produces $Y_{R,k} = 1$ if $\text{Re}(\tilde{Y}_k) \in D_R$ and $Y_{R,k} = 0$ otherwise, and the second quantizer produces $Y_{I,k} = 1$ if $\text{Im}(\tilde{Y}_k) \in D_I$ and $Y_{I,k} = 0$ otherwise, for some Borel sets $D_R, D_I \subset \mathbb{R}$. (Here $\mathbb{R}$ denotes the set of real numbers. $\text{Re}(\cdot)$ denotes the real part, and $\text{Im}(\cdot)$ denotes the imaginary part.) In the coherent case, $D_R$ and $D_I$ may depend on the fading $\{H_k, k \in \mathbb{Z}\}$.

The capacity per unit-energy of this channel is given by (6) or (7), but with $Y$ replaced by $(Y_{R}, Y_{I})$, and with $\mathcal{D} \subset \mathbb{C}$ replaced by $(D_R, D_I) \subset \mathbb{R} \times \mathbb{R}$.

For symmetric quantizers, i.e., for

$$D_k = D_1 = \{ u \in \mathbb{R} : u \geq 0 \} \hspace{1cm} (12)$$

it follows from [11] and [15, Th. 2] that, in the coherent case,

$$\hat{C}_{\text{sym}}(0) = \frac{2}{\pi \sigma^2}. \hspace{1cm} (13)$$

In the noncoherent case, symmetric quantizers result in zero capacity and hence, by (4), in zero capacity per unit-energy. Indeed, for (12)

$$\text{Pr}(Y_{R} = 1 \mid X = x) = \text{Pr}(Y_{I} = 1 \mid X = x) = \frac{1}{2}, \quad x \in \mathbb{C}.$$}

Since, conditioned on $X$, the random variables $Y_{R}$ and $Y_{I}$ are independent, this implies that the capacity is zero. Thus, quantizing the real and imaginary parts of the Rayleigh-fading

2Here and throughout this paper, $\xi$ refers to the parameter in (6) or (7).
channel using symmetric one-bit quantizers reduces the capacity per unit-energy by a factor of \(2/\pi\) in the coherent case, and it reduces it to zero in the noncoherent case. In the following, we show that if we allow for asymmetric quantizers, then we can fully recover the loss of \(2/\pi\) incurred on the coherent Rayleigh-fading channel. For the noncoherent channel, we show that asymmetric quantizers achieve a positive capacity per unit-energy, albeit strictly smaller than \(1/\sigma^2\).

**Theorem 2:** Consider the above channel model, and assume that the real and imaginary parts of \(Y\) are quantized separately using a one-bit quantizer for each.

1) In the coherent case,

\[
\hat{C}(0) = \frac{1}{\sigma^2} \tag{14}
\]

which can be achieved by some quantization regions

\[
D_k^0 = \{ u \in \mathbb{R} : u \geq T_k \} \tag{15}
\]

\[
D_1^0 = \{ u \in \mathbb{R} : u \geq T_1 \} \tag{16}
\]

where \(T_k\) and \(T_1\) depend on \(\text{Re}(H\xi)\) and \(\text{Im}(H\xi)\), respectively.

2) In the noncoherent case,

\[
\frac{2Q(1)}{\sigma^2} \leq \hat{C}(0) < \frac{1}{\sigma^2} \tag{17}
\]

with the upper bound being strict. Here \(Q(\cdot)\) denotes the Gaussian \(Q\)-function [16, Eq. (1.3)]. The lower bound can be achieved by the quantization regions \((15)\) and \((16)\) with \(T_k = T_1 = (|\xi|^2 + \sigma^2)/2\).

**Proof:** Omitted. \(\square\)

V. PROOF OF THEOREM 1

**A. Part 1)**

We show that a radial quantizer \((9)\) achieves the rate per unit-energy \(1/\sigma^2\). Together with \((8)\), this proves Theorem 1. To this end, we first note that, conditioned on \((H, X) = (h, x)\), the squared magnitude of \(\frac{1}{h}Y\) is a noncentral chi-square distribution with degree 2 and noncentrality parameter \(\frac{1}{2} |h|^2 |x|^2\) [16, p. 8]. Consequently, a radial quantizer yields [16, Sec. 2-E]

\[
\text{Pr}(Y = 1 | H = h, X = x) = Q_1\left(\sqrt{\frac{2}{\sigma^2}} |h||x|, \sqrt{\frac{2}{\sigma^2}} T\right)
\]

where \(Q_1(\cdot, \cdot)\) denotes the first-order Marcum \(Q\)-function [16, Eq. (2.20)]. Furthermore, for \(x = 0\) this becomes

\[
\text{Pr}(Y = 1 | H = h, X = 0) = e^{-\frac{x^2}{2\sigma^2}}.
\]

We thus obtain

\[
\begin{align*}
D(&P_{Y|\xi = x} \| P_{Y|\xi = 0} ) \big| P_H) \\
&= \mathbb{E}\left[ Q_1\left(\sqrt{\frac{2}{\sigma^2}} |H||\xi|, \sqrt{\frac{2}{\sigma^2}} T\right) \log \frac{1}{1 - e^{-\frac{t}{2\sigma^2}}} \right] \\
&+ \mathbb{E}\left[ \left(1 - Q_1\left(\sqrt{\frac{2}{\sigma^2}} |H||\xi|, \sqrt{\frac{2}{\sigma^2}} T\right) \log \frac{1}{1 - e^{-\frac{t}{2\sigma^2}}} \right) \\
&- \mathbb{E}\left[ H_b\left(Q_1\left(\sqrt{\frac{2}{\sigma^2}} |H||\xi|, \sqrt{\frac{2}{\sigma^2}} T\right) \right) \\
&\geq \mathbb{E}\left[ Q_1\left(\sqrt{\frac{2}{\sigma^2}} |H||\xi|, \sqrt{\frac{2}{\sigma^2}} T\right) \right] - \log 2 \tag{18}
\end{align*}
\]

where \(H_b(\cdot)\) denotes the binary entropy function, i.e.,

\[
H_b(p) \triangleq \begin{cases} p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}, & 0 < p < 1 \\ 0, & \text{for } p = 0 \text{ or } p = 1. \end{cases}
\]

Here the inequality follows because the second term in the first step is nonnegative, and because the binary entropy function is upper-bounded by \(\log 2\).

We choose \(T = \mu |h||\xi|\) for some fixed \(\mu \in (0, 1)\) and lower-bound the RHS of \((18)\) using the general lower bound on the first-order Marcum \(Q\)-function [16, Sec. C-2, Eq. (C.24)]

\[
Q_1(\alpha, \beta) \geq 1 - \frac{1}{2} \left( e^{-\frac{\alpha - \beta^2}{2}} - e^{-\frac{\alpha + \beta^2}{2}} \right), \quad \alpha > \beta > 0.
\]

We thus obtain for the first term on the RHS of \((18)\)

\[
\begin{align*}
\mathbb{E}\left[ Q_1\left(\sqrt{\frac{2}{\sigma^2}} |H||\xi|, \sqrt{\frac{2}{\sigma^2}} T\right) \right] &\geq \frac{\mu^2 \mathbb{E}[|H|^2]}{\sigma^2} \frac{\sigma^2}{2} \\
&- \frac{1}{2} \mathbb{E}\left[ \exp \left( -\frac{|H|^2}{\sigma^2} - (\mu - 1)^2 \frac{\mu^2 |H|^2}{\sigma^2} \right) \left| \frac{|H|^2}{\sigma^2} \right|^2 \right] \\
&+ \frac{1}{2} \mathbb{E}\left[ \exp \left( -\frac{|H|^2}{\sigma^2} - (\mu + 1)^2 \frac{\mu^2 |H|^2}{\sigma^2} \right) \left| \frac{|H|^2}{\sigma^2} \right|^2 \right]
\end{align*}
\]

\[
\geq \frac{\mu^2 \mathbb{E}[|H|^2]}{\sigma^2} \frac{\sigma^2}{2} - \frac{\mu^2}{2} e^{-\frac{|H|^2}{\sigma^2}} \tag{19}
\]

where the last step follows because \(0 \leq xe^{-\alpha x} \leq 1/(\alpha 0\) for every \(x > 0\) and \(\alpha > 0\).

Combining \((19)\) with \((18)\), and computing its ratio to \(|\xi|^2\) in the limit as \(|\xi|^2\) tends to infinity, yields

\[
\hat{C}(0) \geq \frac{\mu^2 \mathbb{E}[|H|^2]}{\sigma^2} = \frac{\mu^2}{\sigma^2} \tag{20}
\]

Theorem 1 follows then by letting \(\mu\) tend to one.

**B. Part 2)**

We first note that, by the Data Processing Inequality for relative entropy [10, Sec. 2.9], the relative entropy on the RHS of \((7)\) is upper-bounded by the relative entropy corresponding to the unquantized channel, i.e., [6, Eq. (64)]

\[
D(P_{Y|X = \xi} \| P_{Y|X = 0}) \leq \frac{1}{\sigma^2} \log \left( 1 + \frac{|\xi|^2}{\sigma^2} \right). \tag{21}
\]

38
Consequently, the capacity per unit-cost \( (7) \) is strictly smaller than \( 1/\sigma^2 \) unless \( |\xi| \) tends to infinity. It thus remains to show that

\[
\lim_{|\xi| \to \infty} \sup_{D} \frac{D(P_{Y|X=\xi} \parallel P_{Y|X=0})}{|\xi|^2} < \frac{1}{\sigma^2} \tag{22}
\]

To this end, we first note that, for every \( \xi \neq 0 \), the supremum in (22) over all quantization regions \( D \) can be replaced with the supremum over all radial quantizers (9). Indeed, for every quantization region satisfying

\[
\Pr(Y = 1 \mid X = \xi) = \beta, \quad 0 < \beta < 1
\]

the relative entropy

\[
D(P_{Y|X=\xi} \parallel P_{Y|X=0}) = \beta \log \frac{\Pr(Y = 1 \mid X = 0)}{\Pr(Y = 1 \mid X = 0)} + (1 - \beta) \log \left( 1 - \Pr(Y = 1 \mid X = 0) \right) - H_b(\beta) \tag{23}
\]

is a convex function of \( \Pr(Y = 1 \mid X = 0) \). Thus, for every \( 0 < \beta < 1 \), the RHS of (23) is maximized for the quantization region that minimizes (or maximizes) \( \Pr(Y = 1 \mid X = 0) \) while holding \( \Pr(Y = 1 \mid X = \xi) = \beta \) fixed. By the Neyman-Pearson lemma [17], such a quantization region has the form

\[
D_\star = \left\{ \hat{y} \in \mathbb{C} : \frac{f(\hat{y})(0)}{f(\hat{y})(\xi)} \leq \Lambda, \quad \Lambda > 0 \right\} \tag{24}
\]

(or the complement thereof), where \( f(\hat{y}|x) \) denotes the conditional density of \( Y \), conditioned on \( X = x \), and \( \Lambda \) is such that \( \Pr(Y \in D_\star \mid X = \xi) = \beta \). (Note that for every \( 0 < \beta < 1 \) there exists such a \( \Lambda \) since, for the above channel model (1), \( \Pr(Y \in D_\star \mid X = \xi) \) is a continuous, strictly increasing function of \( \Lambda > 0 \).) The likelihood ratio on the RHS of (24) is readily evaluated as

\[
f(\hat{y}(0))/(f(\hat{y})(\xi)) = \left( 1 + \frac{|\xi|^2}{\sigma^2} \right) e^{-|\xi|^2/(2\sigma^2)}, \quad \hat{y} \in \mathbb{C}.
\]

Consequently, \( D_\star \) is the same as (9) with

\[
T = \sigma \left( 1 + \frac{\sigma^2}{|\xi|^2} \right)^{\frac{1}{2}}.
\]

We thus obtain that, for every \( \xi \neq 0 \), the relative entropy \( D(P_{Y|X=\xi} \parallel P_{Y|X=0}) \) is maximized by a radial quantizer (9). For such a quantizer, we have

\[
\Pr(Y = 1 \mid X = x) = \exp\left( -\frac{T^2}{|x|^2 + \sigma^2} \right)
\]

which yields

\[
D(P_{Y|X=\xi} \parallel P_{Y|X=0}) = e^{-|\xi|^2/(2\sigma^2)} \log \frac{1}{e^{-\sigma^2/2}} + \left[ 1 - e^{-|\xi|^2/(2\sigma^2)} \right] \log \frac{1}{e^{-\sigma^2/2}} - H_b\left( e^{-|\xi|^2/(2\sigma^2)} \right) \leq \frac{T^2}{\sigma^2} e^{-|\xi|^2/(2\sigma^2)} - \frac{1 - e^{-|\xi|^2/(2\sigma^2)} \log \left( 1 - e^{-|\xi|^2/(2\sigma^2)} \right)}{\sigma^2} \leq \frac{T^2}{\sigma^2} e^{-|\xi|^2/(2\sigma^2)} + \frac{1}{\sigma^2} \tag{25}
\]

where the second step follows because \( H_b(\cdot) \geq \log \frac{1}{e^2} \geq 0 \) and \( \exp(-T^2/(|\xi|^2 + \sigma^2)) \geq \exp(-T^2/\sigma^2) \), and the third step follows because \(-x \log x \leq \frac{x}{2}, \quad 0 < x < 1\). The first term on the RHS of (25) is maximized for \( T^2 = |\xi|^2 + \sigma^2 \). The RHS of (25) is thus upper-bounded by

\[
\lim_{|\xi| \to \infty} \sup_D D(P_{Y|X=\xi} \parallel P_{Y|X=0}) \leq \frac{|\xi|^2}{\sigma^2} + \frac{2}{\sigma^2} \tag{26}
\]

Dividing the RHS of (26) by \( |\xi|^2 \), and computing the limit as \( |\xi| \) tends to infinity, yields

\[
\lim_{|\xi| \to \infty} \sup_D D(P_{Y|X=\xi} \parallel P_{Y|X=0}) \leq \frac{1}{\sigma^2} \leq \frac{1}{\sigma^2} \tag{27}
\]

This proves Theorem 2.

Acknowledgment

Enlightening discussions with A. Martinez and comments from S. Verdú are gratefully acknowledged.

References

Model Switching and other Unconventional Fun with Experts

Steven de Rooij
Centrum Wiskunde & Informatica (CWI)
Science Park 123, Amsterdam, Netherlands
Email: s.de.rooij@cwi.nl

I. UNIVERSAL CODING AND REGRET

Universal codes combine a number of codes into one that achieves code lengths close to those of the best of the bunch. We consider the strongest notion of universality, where we compete with the best code for any data sequence, not just in expectation. When this is the goal, a natural performance measure is regret: the difference between the universal code length and the length of the best original code. Given a fixed finite set of $K$ codes, a simple universal combination can be obtained with a two-part scheme, where the first $\log_2 K$ bits are used to identify the original codes to be used for the data, and the second part is the actual encoding of the data using that code. The regret of this code is bounded by the length of the first part of the code. (From here on, all logarithms are binary and we ignore the rounding of code lengths to integer numbers of bits.)

This idea can be taken a step further. Consider the case where a good code is available for any reasonably small segment of the input sequence, but either the data or the codes themselves are so changeable that none of the $K$ original codes perform very well on the entire data sequence. In such cases, a stronger regret guarantee is desirable. For simplicity, fix the number of source symbols at $n$. Then, starting from $K$ codes, first construct a new set of all “tracking” codes that split the input sequence into $m$ blocks, and use one of the $K$ original codes for each block. The goal is now to minimise the regret with respect to this extended set of codes. Counting how many such codes there are is a simple combinatorial exercise; we can subsequently use a two-part code as before to obtain a regret $\mathcal{R}(x^n, m)$ satisfying

$$\mathcal{R}(x^n, m) \leq \log \left( \frac{n - 1}{m - 1} \right) + m \log K. \quad (1)$$

This solution is not quite satisfactory. First, it requires finding the best tracking code; while this can be done using dynamic programming, we will see that it is possible to obtain similar regret guarantees with much faster algorithms. Second, the number of segments $m$ is a parameter of the algorithm, but $m$ is typically unknown so we would like to have a regret guarantee that is valid for all $m$ simultaneously. Third, the code is not sequential, making it unusable in an online setting.

A number of very simple and elegant algorithms have been developed that avoid these issues. Some derive from the information theory literature [Wi07], [VW98], [SM99], but there are also important, and strongly related results in online learning that deserve more exposure [HW98], [Vov99], [BW02], [CBL06], [GLL08]. This text is based on our paper [KdR08], which outlines the common structure to these algorithms.

II. EXPERT SEQUENCE PRIORS

Universal codes of the kind described above are most conveniently defined and analysed in the Bayesian framework. As a first step, we switch to probability theory and replace the length functions $L_1, \ldots, L_K$ of the $K$ original codes by probabilistic sources $P_1, \ldots, P_K$, such that $-\log P_k(x^n) = L_k(x^n)$ for all $k$ and all data sequences $x^n = x_1, \ldots, x_n \in \mathcal{X}^n$ (for simplicity assume the alphabet to be discrete). Such probabilistic sources can now be used in a sequential prediction setting; the accumulated logarithmic loss exactly corresponds to the code length:

$$\sum_{t=1}^{n} -\log P_k(x_t|x^{t-1}) = -\log P_k(x^n = x^n) = L(x^n).$$

In the online learning literature, the prediction strategies $P_k$ are called “experts” because, unless we have additional knowledge about the data generating process, which we do not wish to assume, we have to use their predictions.

We now extend the model class as follows. Let $k^n \in \{1, \ldots, K\}^n$ denote a sequence of expert identifiers. Define

$$P_k^n(x^n) = P_k(X_1) \cdot P_k(X_2|X_1) \cdot \ldots \cdot P_k(X_n|X_{n-1}).$$

We do not know which sequence of experts will yield the best predictions, but we can represent our uncertainty in the usual Bayesian way in the form of a prior $\pi$ on sequences of experts. We can use this prior to make predictions by conditioning the Bayesian marginal distribution:

$$P(X_{n+1}|x^n) = \frac{P(x^{n+1}|x^n)}{P(x^n)} = \frac{\sum_{k^{n+1}} P_{k^{n+1}}(x^{n+1}, x_{n+1})}{\sum_{k^n} P_{k^n}(x^n)} \quad (2)$$

It is now also very easy to obtain a reasonable regret bound, by dropping every term from the marginal likelihood except the one we are interested in. Choose any reference expert sequence $\tilde{k}^n$, and let $m$ be the number of blocks it has. Then

$$\mathcal{R}(x^n, \tilde{k}^m) = -\log P(x^n) - (-\log P_{\tilde{k}^m}^n)$$

$$= -\log \sum_{k^n} P_{k^n}(x^n) \pi(k^n) + \log P_{\tilde{k}^m}(x^n)$$

$$\leq -\log \left( P_{\tilde{k}^m}(x^n) \pi(\tilde{k}^m) \right) + \log P_{\tilde{k}^m}(x^n)$$

$$= -\log \pi(\tilde{k}^m).$$

40
A. Example

In general, the best sequence of experts may have a very low prior probability, so it will be hard to compete with. However, we can construct the prior such that expert sequences with relatively few blocks receive a high prior probability. For example, we may set

$$\pi(k^n) = K^{-m} \alpha^{n-1} \left( \frac{1 - \alpha}{\alpha} \right)^{m-1}, \quad (4)$$

where $m$ is the number of blocks in $k^n$, and $\alpha \in [0,1]$ is a hyperparameter that can be interpreted as the “switching rate”: it is the probability of a new block starting at any given time. It may be verified that this prior sums to one.

An algorithm to efficiently compute the predictions of this distribution, called Fixed Share, was first published in COLT 1995; see [HW98] for the journal version. For Fixed Share, the regret bound (3) reduces to

$$R(x^n, k^n) \leq m \log K - n \log \alpha - (m - 1) \log \left( \frac{1 - \alpha}{\alpha} \right).$$

The bound is optimised by setting the hyperparameter to $\alpha^* = (n - m + 1)/(n - 1)$, which yields

$$R(x^n, k^n) \leq m \log K + (n - 1)H(\alpha^*), \quad (5)$$

where $H$ denotes the binary entropy function. Note that this bound is very close to (1); the slack is necessary because this new regret bound is valid for all $m$, whereas (1) is tight for a single fixed $m$.

B. Second Example: Hyperpriors

In case the expert sequence prior is burdened by an unknown parameter, we can “integrate it out” using a hyperprior. Let $\pi^{(\alpha)}$ denote an expert sequence prior $\pi$ with hyperparameter $\alpha$, and let $P^{(\alpha)}$ denote the tracking distribution with expert sequence prior $\pi^{(\alpha)}$. Now impose a hyperprior $\rho$ on $\alpha$. We then obtain the following parameterless distribution:

$$P(X^n) = \int P^{(\alpha)}(X^n) \rho(\alpha) \, d\alpha$$

$$= \sum_{k^n} P_{k^n}(X^n) \int P^{(\alpha)}(k^n) \rho(\alpha) \, d\alpha.$$  

Note that the final integral itself defines an expert sequence prior. We will call it $\sigma$. Thus, to obtain a parameterless tracking distribution we do not have to look at the data; it suffices to mix the expert sequence prior using $\rho$.

In case of the example of the previous section, with $\pi^{(\alpha)}$ given as in (4) and hyperprior $\rho(\alpha) = \text{Beta} \left( \frac{1}{2}, \frac{1}{2} \right)$, we find

$$\sigma(k^n) = K^{-m} \frac{\Gamma(n - m + \frac{1}{2}) \Gamma(m + \frac{1}{2})}{\pi \Gamma(n + 1)}.$$  

Around the time that Fixed Share was published, but completely independently, Paul Volf and Frans Willems developed an algorithm dubbed the “Switching Method”, that efficiently predicts according to this distribution [Wii07], [VW98]. It can be shown that its regret exceeds (5) by at most $\frac{1}{2} \log n + 1$ bits, a modest price to pay for a completely parameterless strategy.

In the next section we proceed to show how Fixed Share and the Switching Method can be implemented efficiently.

III. Computation using HMMs

In general, the predictions of an expert distribution based on an expert sequence prior (2) may require summing an exponential number of terms, but the complexity can be reduced dramatically if the expert sequence prior admits a simple sufficient statistic. This is the case for the two example priors of the previous section: in Fixed Share, the marginal prior distribution on the next expert $\pi(K_{n+1}|k^n)$ can be expressed as a function of only the current expert $k_n$. Namely,

$$\pi(K_{n+1} = k|k^n) = \frac{\alpha}{R} + 1 \{ k = k_n \} (1 - \alpha).$$

Such expert distributions can be represented by a Hidden Markov Model (HMM), where the hidden state is such a sufficient statistic. The value of the hidden variable determines the distribution on experts; the experts’ predictions are weighed according to this distribution. It is convenient to represent the prior in the form of a state transition diagram for the hidden states. For the Fixed Share algorithm with four experts, the diagram looks as follows:

All straight transitions have probability $1 - \alpha$; they represent that no new block is started at that sample size. The transitions into the centre nodes have probability $\alpha$; they represent a switch to a new block. The centre nodes are “silent states” where no prediction has to be produced; such states redistribute some of the probability mass over the experts. The arcs out of the silent states all have probability $1/K$.

Using the Forward algorithm with these state transitions, the distribution on the hidden state can be updated in $O(K)$, so that the entire data set can be processed in $O(n \cdot K)$.

The switching method is only slightly more complicated. In this case, the currently used expert is not sufficient to determine the probability of starting a new block: we also need to know how many blocks there were previously. The following state transition diagram visualises the state transition diagram required to calculate $\sigma(K_{n+1}|k^n)$. For simplicity it is for two experts, and we also assume that the first expert always predicts first. Then the current expert is determined by the number of switches modulo 2.
Here the upward arcs represent the starting of a new block, and the straight arcs represent staying in the same block. For the $n$th node from the left, the $s$th row from the bottom, the probability of the upward arc is $(s + 1/2)/(n + 1)$ and the probability of the straight arc is one minus that; in other words the Krichevsky-Trofimov estimator is used to predict whether or not a switch occurs.

Note that the time complexity of the forward algorithm on all symbols has increased to $O(n^2 \cdot K)$.

IV. THERE IS MORE

The power of expert tracking algorithms is twofold. First, they are black box: any online prediction strategy can be used as an expert. As such they can be applied very generally. One example is in [vEGdR11], where the estimators associated with each a series of model classes are used as “experts”. The paper subsequently uses an algorithm similar to Fixed Share to improve the performance of Bayesian Model Averaging under many circumstances. Second, very attractive computational complexity can sometimes be achieved. For example, the Fixed Share algorithm runs exactly as fast as Bayesian Model Averaging itself.

Based on our paper [KdR08], we have seen two expert tracking algorithms as examples of how you can do more with a bunch of prediction strategies than what everyone always seems to be doing, which is to just take the Bayesian mixture. The preprint of the full paper discusses several other expert sequence priors for tracking, as well as more sophisticated methods to analyse the resulting performance.

Much of current online learning research is related to prediction with expert advice, but it may not be instantly recognisable as such, as over time the problem has been generalised substantially. Rather than that the learner uses a set of experts who make predictions in the form of probability distributions, is subsequently penalized with logarithmic loss, before finally observing the actual outcome, in the more general setting both experts and outcomes are abstracted away altogether: the learner chooses an action from some space of allowed actions, and receives an arbitrary loss. This more general setup allows the theory to encompass online linear optimisation and prediction problems with loss functions other than logarithmic loss. It also makes the problem harder, and perhaps less directly applicable to information theory. As a starting point, see [CBL06].

REFERENCES


42
Some Old and New Relations between Information and Estimation
Kartik Venkat∗, Tsachy Weissman∗

Elegant relationships between classical quantities in information and estimation arise in the presence of additive Gaussian noise. In this talk, we review and present these relationships as expectations of random quantities, and subsequently present some of our recent findings pertaining to their ‘pointwise’ analogues.

Duncan, in [1], showed that for the continuous-time additive white Gaussian noise channel, the minimum mean squared filtering (causal estimation) error is twice the input-output mutual information for any underlying signal distribution. Another discovery was made by Guo et al. in [2], where the derivative of the mutual information was found to equal half the minimum mean squared error in non-causal estimation. By combining these two intriguing results, the authors of [2] also establish the remarkable equality of the causal mean squared error (at some ‘signal to noise’ level snr) and the non-causal error averaged over ‘signal to noise’ ratio uniformly distributed between 0 and snr. There have been extensions of these results to the presence of mismatch. In this case, the relative entropy and the difference of the mismatched and matched mean squared errors are bridged together: Mismatched estimation in the scalar Gaussian channel was considered by Verdú in [3]. In [4], a generalization of Duncan’s result to incorporate mismatch for the full generality of continuous time processes is provided. In [5], Kadota et al. generalize Duncan’s theorem to the presence of feedback.

In [6], the pointwise analogues of these relationships are obtained, giving considerable insight into the above results. As an illustration, consider Duncan’s 1970 result, which can equivalently be expressed saying that the difference between the input-output information density and half the causal estimation error is a zero mean random variable (regardless of the distribution of the channel input). We characterize this random variable explicitly, rather than merely its expectation. Classical estimation and information theoretic quantities emerge with new and surprising roles. For example, the variance of this random variable turns out to be given by the causal MMSE (which, in turn, is equal to twice the mutual information by Duncan’s result).

REFERENCES

∗ Stanford University. Email: kvenkat@stanford.edu, tsachy@stanford.edu
Polar martingales of maximal spread

Emmanuel Abbe
Faculté Informatique et Communications
Ecole Polytechnique Federale de Lausanne
Email: emmanuel.abbe@epfl.ch

Abstract—This paper investigates polar codes for q-ary input channels using other kernels than \([1, 0]\) to construct the generator matrix. The kernel is optimized by maximizing the spread of the polar martingale and it is shown that in certain cases, the resulting construction has a significantly improved error probability (even at small block length like \(n = 1024\)).

I. INTRODUCTION

The generator matrix of a polar code [1] of block length \(n\) is obtained by deleting some rows of the matrix \(G_n = [1, 0]^\otimes \log_2 n\). Which rows to delete depends on the channel and on the targeted error probability (or rate). For a symmetric discrete memoryless channel \(W\), the rows to be deleted are indexed by

\[
B_{\epsilon,n} := \{i \in [n] : I(U_i; Y^n U^{i-1}) \leq 1 - \epsilon\},
\]

where \(\epsilon\) is a parameter governing the error probability. \(U^n\) has i.i.d. components which are uniform on the input alphabet, \(X^n = U^n G_n\) and \(Y^n\) is the output of \(n\) independent uses of \(W\) when \(X^n\) is the input. The key observation at the root of this construction can be explain by considering \(n = 2\). Let \(u_1, u_2 \in \mathbb{F}_2 = \{0, 1\}\)

\[
x_1 = u_1 + u_2
\]

\[
x_2 = u_2
\]

and let \(Y_1, Y_2\) be the output of two independent uses of a binary input channel \(W\) when \(x_1, x_2\) are the inputs. Note that the binary input channels

\[
W^- : u_1 \rightarrow Y_1, Y_2
\]

\[
W^+ : u_2 \rightarrow Y_1, Y_2, u_1
\]

satisfy the relations

\[
I(W^-) \leq I(W) \leq I(W^+)
\]

\[
I(W^-) + I(W^+) = 2I(W)
\]

where \(I(W)\) denotes the mutual information of the channel \(W\) when the input distribution is uniform over \(\mathbb{F}_2\). Therefore, we have transformed two independent uses of the channel \(W\) into two uses of the channels \(W^-\) and \(W^+\), with overall the same amount of mutual information. If one had to transmit only one bit uniformly drawn, using \(W^+\) rather than \(W^-\) would lead to a lower error probability. One can then iterate this argument, after \(\log_2(n)\) iterations, one obtains the channels \(u_i \mapsto Y^n u^{i-1}\). For a given number of information bits to be transmitted (rate), selecting the channels having largest mutual informations leads to the least error probability. As explained in the next section, polarization happens in the sense that as \(n\) tends to infinity, these synthesized channels have mutual information tending to either 0 or 1. Hence, sending information bits on the high mutual information channels (equivalently, deleting rows of \(G_n\) corresponding to low mutual information channels) allows to achieve communication rates as large as the mutual information of the original binary input channel. The construction extends to q-ary input alphabets when \(q\) is prime, using the same matrix \(G_n = [1, 0]^\otimes \log_2 n\), while carrying the operations over \(\mathbb{F}_q\).

It is tempting to investigate what happens if one keeps the Kronecker structure for the generator matrix but modifies the kernel \([1, 0]\). For binary input alphabets, there might not be another interesting choice. One can consider kernel matrices which are not \(2 \times 2\) but \(k \times k\). In this case, [3] shows that a significant improvement in the error probability rate can be achieved, however, to the cost of suffering from a complexity increase and a much larger block length requirement. Not as much has been studied concerning modifying the kernel matrix over \(\mathbb{F}_q\). In [4], the error probability of non-binary polar codes constructed on the basis of Reed-Solomon matrices is calculated using numerical simulations on q-ary erasure channels. It is confirmed that 4-ary polar codes can have significantly better performance than binary polar codes. In this paper, our goal is to investigate potential improvements at finite block length using modified kernels over \(\mathbb{F}_q\). We propose to pick kernels not by optimizing the polar code exponent as in [4] but by maximizing the polar martingale spread. The resulting improvement is illustrated with numerical simulations. Preliminary theoretical results are provided and it is shown how the proposed criteria connects to new entropy maximization problems.

A. Polar martingale

In order to show that polarization happens, namely that

\[
|\{i \in [n] : I(U_i; Y^n u^{i-1}) \in (\epsilon, 1 - \epsilon)\}| \rightarrow 0,
\]

it is helpful to rely on a random process having a uniform measure on the possible realizations of \(I(U_i; Y^n u^{i-1})\), so that counting the number of such mutual informations in \((\epsilon, 1 - \epsilon)\) is obtained by evaluating the probability that the process lies in this interval. The process is defined by taking \(\{B_n\}_{n \geq 1}\) to be

1If the channel is symmetric the generator matrix is indeed obtained by deleting rows, otherwise in addition to deleting rows one may also have to translate the codewords (affine code)
i.i.d. uniform random variables valued in \{−, +\} and defining the binary (or q-ary with q prime) random input channels \{W_n, n \geq 0\} by

\[
W_0 := W, \\
W_n := W_{B_n}^n, \quad \forall n \geq 1.
\]

The polarization result is then expressed as

\[
P\{I(W_n) \in (\varepsilon, 1 - \varepsilon)\} \to 0.
\]

Here, the key point is that the martingale \(I(W_n)\) is a random walk in \([0, 1]\) which is ‘unstable’ at any points \(I(W) \in (0, 1)\) as it must move from at least \(I(W^+) - I(W^-) > 0\) in this range. The plot of \(I(W^+) - I(W^-)\) for different values of \(I(W)\) is provided in Figure 1.

When restricting the channels to be ‘additive noise’, one can equivalently work with the ‘source’ setting, in which case the corresponding martingale spread is at least \(H(X + X') - H(X)\), where \(X, X'\) are i.i.d. under the noise distribution. In [2], the rate of convergence of the polar martingale is studied as a function of the block length. We will investigate in this paper the performance at finite block length.

II. Problem: Kernels with Maximal Spread

Being interested in the performance of polar codes at finite block length, we start in this paper with the optimization of the kernel matrix over \(F_q\) for a block length of 2. Namely, we investigate the following optimization:

\[
K^*(W) = \arg \max_{K \in M_2(F_q)} I(W^+(W, K))
\]

where \(W^+(W, K)\) is the channel \(u_2 \mapsto Y_1Y_2u_1, (Y_1, Y_2)\) are the output of two independent uses of \(W\) when \((x_1, x_2) = (u_1, u_2)K\) are the inputs. We call \(K^*\) the 2-optimal kernel for \(W\).

We consider in particular channels which are ‘additive noise’, in which case one can equivalently study the ‘source’ version of this problem as follows:

\[
\lambda^*(\mu) = \arg \max_{\lambda \in \mathbb{P}_q} H(X + \lambda X')
\]

where \(X, X'\) are i.i.d. under \(\mu\). As discussed next, the connection with previous problem is obtained by choosing

\[
K^*(W) = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}
\]

where \(\mu\) is the distribution of the noise of the channel \(W\).

III. Equivalent Problems and Definitions

A general kernel is a matrix over \(F_q\) which is invertible. For \(2 \times 2\) matrices, let \(K = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\) be such a matrix and consider \((X_1, X_2)\) to be i.i.d. under \(\mu\) over \(F_q\) and \((U_1, U_2) = (X_1, X_2)K\). Then, since \(K\) is invertible, we have

\[
2H(\mu) = H(X_1, X_2) = H(U_1, U_2) = H(U_1) + H(U_2 | U_1)
\]

and

\[
H(U_1) - H(\mu) = H(\mu) - H(U_2 | U_1)
\]

is the entropy spread gained by using the transformation \(K\). To maximize the spread, one may maximize \(H(U_1) = H(ax_1 + cx_2)\) over the choice of \(a\) and \(c\), or simply \(H(X_1 + cX_2)\) over the choice of \(c\) (this leads to the definition of \(\lambda^*\)). Hence, the maximization problem depends on the variable \(c\). \(a\) can be set to 1, and \(b\) and \(d\) need to ensure that \(K\) is invertible, which leads w.l.o.g. to a kernel of the form \(K = \begin{bmatrix} 1 & b \\ c & 1 \end{bmatrix}\).

Note that to maximize the spread, one may alternatively minimize \(H(U_2 | U_1) = H(X_2 | X_1 + cX_2)\).

IV. Results for \(F_3\)

**Lemma 1.** Let \(X, X'\) be i.i.d. random variables valued in \(F_3\), then

\[
H(X + X') \leq H(X - X').
\]

(The arithmetic is over \(F_3\).)

**Proof outline:** Let \(p\) denote the distribution of \(X + X'\) and \(\tilde{p}\) denote the distribution of \(X - X'\). Note that

\[
\|F(p)\|_2 = \|F(\tilde{p})\|_2
\]

where \(F\) is the Fourier transform and \(\|\cdot\|_2\) the \(l_2\)-norm. Hence,

\[
\|p\|_2 = \|\tilde{p}\|_2,
\]

This implies that \(p\) and \(\tilde{p}\) lie on an Euclidean ball (and on the simplex). Moreover, note that \(\tilde{p}\) is such that \(\tilde{p}(1) = \tilde{p}(2)\) and
Fig. 2. Plot of $H(X)$ (horizontal axis) vs. $H(X + X') - H(X)$ (in blue) and $H(X - X') - H(X)$ (in red – top curves) for various $X, X'$ i.i.d. on $F_3$. As confirmed by Lemma 1, the gap $H(X - X') - H(X)$ is larger than $H(X + X') - H(X)$.

Fig. 3. Block error probability (in log$_{10}$ scale) of a polar code using the 2-optimal kernel (red curve – lower curve) vs. original kernel (blue curve) for a block length of $n = 1024$ and an additive noise channel over $F_3$ with noise distribution $\{0.7, 0.3, 0\}$.

Proof of Lemma 2: The condition $|S_\mu + \gamma S_\nu| = |S_\mu|^2$ ensures that knowing $x_1 + \gamma x_2$ with $x_1, x_2 \in S_\mu$ allows to exactly recover both $x_1$ and $x_2$.

Examples of distributions satisfying (16): Let $\mu$ over $F_5$ be such that $S_\mu = \{0, 1\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2\}$ and $S_\mu + 2S_\mu = \{0, 1, 2, 3\}$, and (16) is verified. In this case, using $\gamma = 1$ can only provide a strictly smaller spread since it will not set $H(X_2|X_1 + \gamma X_2) = 0$. It is hence better to use the 2-optimal kernel $[[1, 0]]$ rather than the original kernel $[[1, 1]]$, and as illustrated in Figure 4, this leads to significant improvements in the error probability at finite block length. Also note that a channel with noise $\mu$ satisfying (16) has positive zero-error capacity, which is captured by the 2-optimal kernel as shown with the rapid drop of the error probability (it is 0 at low enough rates since half of the synthetized channels have noise entropy exactly zero). If $\mu$ is close to a distribution satisfying (16), the error probability is also significantly improved with respect to the original kernel $[[1, 0]]$.

Over $F_{11}$, let $\mu$ be such that $S_\mu = \{0, 2\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2, 4\}$ and (16) does not hold. However, picking $\gamma = 3$ leads to $3S_\mu = \{0, 3, 6\}$ and (16) holds. Therefore, $\gamma^*$ varies with $q$.

Over $F_q$, defining $k = \lceil \sqrt{q} - 1 \rceil$, if $S_\mu = \{0, 1, \ldots, k - 1\}$ and $\gamma = k$, (16) holds.

VI. DISCUSSION

We have shown that over $F_q$, the martingale spread can be significantly enlarged by using 2-optimal kernels rather than the original kernel $[[1, 1]]$. In some cases, this leads to significant improvements on the error probability of polar codes, even at low block length ($n = 1024$). We have considered additive noise channels, and shown that the improvement is particularly significant when the noise distribution
Fig. 4. Block error probability (in $\log_{10}$ scale) of a polar code using the 2-optimal kernel (red curve – lower curve) vs. original kernel (blue curve) for a block length of $n = 1024$ and an additive noise channel over $F_5$ with noise distribution $\{0.5, 0.5, 0, 0, 0\}$. This channel takes any symbol of $F_5$ to itself with probability $1/2$ and shifts any symbol circularly with probability $1/2$.

is concentrated on ‘small’ supports. For distributions which have full support, the improvement may not be significant. It would be interesting to investigate the improvement for the AWGN channel, and check whether this channel falls in the cases where the improvement is significant. It would also be interesting to consider polar codes that have a kernel which changes adaptively in the recursive construction. This will require a more elaborate decoding algorithm but may not hurt the complexity. Finally, it would be interesting to show analytical results on the error probability rate or exponent of 2-optimal kernels, to consider extensions to $k \times k$ rather than $2 \times 2$ kernels, or to consider other criteria for the optimization of the kernel.

Acknowledgement

I would like to thank Emre Telatar for helpful comments and stimulating discussions.

References

Sumset Differential Entropy Bounds

Ioannis Kontoyiannis
Department of Informatics
Athens University of Economics & Business
Patission 76, Athens 10434, Greece
Email: yiannis@aueb.gr
Web: http://pages.cs.aueb.gr/yiannisk/

Abstract—Terence Tao recently developed a series of new inequalities for the (discrete) Shannon entropy, which parallel sumset and inverse sumset bounds in additive combinatorics. We examine some of their natural analogs for differential entropy. Since the main ingredient in Tao’s proofs (the submodularity of the entropy function) fails in the continuous case, interesting new phenomena arise. We review some known results and present a series of new differential entropy inequalities. These results are based, in part, on joint work with Mokshay Madiman.

I. ADDITIVE COMBINATORICS AND ENTROPY

The field of additive combinatorics [see, e.g., [7] for an introduction] is the theory of additive structures in sets equipped with a group structure. The prototypical example is the study of arithmetic progressions in sets of integers, as opposed to the multiplicative structure that underlies prime factorization and much of classical combinatorics and number theory. There have been several major developments and a lot of high-profile mathematical activity in additive combinatorics in recent years, with perhaps the most famous example being the celebrated Green-Tao theorem on the existence of arbitrarily long arithmetic progressions within the primes.

An important collection of tools in the study of additive combinatorics is a variety of sumset inequalities. Here, the sumset $A + B$ of two discrete sets $A$ and $B$ is defined as, $A + B = \{a + b : a \in A, b \in B\}$, and a sumset inequality is an inequality connecting the cardinality $|A + B|$ of $A + B$ with the cardinalities of $A$ and $B$. Therefore, roughly and somewhat incorrectly speaking, it might be said that additive combinatorics gives bounds on the sizes of discrete additive sets.

In this setting, we recall the natural connection between entropy and set cardinality established by the AEP: The entropy $H(X)$ can be thought of as the logarithm of the effective cardinality of the alphabet of a discrete random variable $X$. This suggests a correspondence between bounds for the cardinalities of sumsets like, e.g., $|A + B|$, and corresponding bounds for the entropy of sums of discrete random variables, e.g., $H(X + Y)$. This connection appears to have first been identified by Imre Ruzsa, and in the last few years it has also been explored in different directions by, among others, Tao and Vu [6], Lapidoth and Pete [1], Ruzsa [4], Madiman, Marcus and Tetali [3], Madiman and Kontoyiannis [2], and Tao [5].

II. SUMSET BOUNDS AND DIFFERENTIAL ENTROPY

This analogy can be carried further: The continuous AEP states that the differential entropy $h(X)$ of a continuous random vector $X$ can be thought of as the logarithm of the (Euclidean or Lebesgue) “volume of the effective support” of $X$. It is then natural to consider whether the recent discrete sumset entropy bounds can be extended to the continuous case.

Taking as our starting point the results in Tao’s recent work [5], we provide natural differential entropy analogs for the following inequalities:

- Ruzsa triangle inequality
- Sum-difference inequality
- Plünnecke-Ruzsa inequality
- Iterated sum bound
- Balog-Szemeredi-Gowers lemma
- Ruzsa’s covering lemma
- Green-Ruzsa-Freiman inverse-sumset theorem

The difficulty as well as the excitement of this work stem from the fact that, in the continuous case, the natural generalizations of the proofs of the discrete versions of many of these results fail at a deep level. One reason is that the main ingredient in earlier proofs is the submodularity of the discrete entropy functional, which does not hold for differential entropy. And, further, it is the overall structure of the proofs that does not carry on to the continuous case: Not the method, but the actual intermediate steps fail to hold for differential entropy. Therefore, it is necessary to employ different tools and fundamentally new proof strategies.

REFERENCES

Coding Theory and Compressed Sensing through Convex Relaxations

Arash Saber Tehrani, Alexandros G. Dimakis
Dept. of Electrical Eng. University of Southern California
Los Angeles, CA, USA
emails: {saberteh, dimakis}@usc.edu

Abstract—We survey recent connections between convex relaxations used in coding theory and in sparse approximation theory. One important conclusion arising from our results is that graph girth can be used to certify that sparse compressed sensing matrices have good sparse approximation guarantees. This allows us to present the first deterministic measurement matrix constructions that have an optimal number of measurements for \( \ell_2/\ell_1 \) sparse approximation.

I. INTRODUCTION

Sparse approximation theory and compressed sensing have recently received significant attention in signal processing and statistics. We focus on two developments that connect this area with coding theory and allow the translation of results and techniques. The first involves mathematical links [1]–[3] between two linear programming relaxations: basis pursuit for CS reconstruction and the fundamental polytope relaxation for the decoding of binary linear codes [4], [5]. In coding theory, the linear-program (LP) relaxation was introduced by Feldman, Wainwright and Karger [4] and is closely related to the work by Koetter and Vontobel [5]. This work defined the so-called fundamental polytope and structures called pseudocodewords that cause problems to channel decoders (see also the work of Wiberg [6] and Forney et al. [7]). Our recent results [1], [8] connect this theory to the basis pursuit LP relaxation used in compressed sensing. The high-level message is that ‘good’ channel codes can be used over the reals as provably good CS measurement matrices under basis pursuit. This connection allowed us to solve a well-known open problem in compressed sensing, the deterministic construction of measurement matrices with an order-optimal number of rows [1], [8].

The second connection involves the nature of iterative message-passing algorithms used both for channel coding and compressed sensing. A number of iterative algorithms (e.g., [9]–[13]) have been proposed for compressed sensing recovery, with the benefit of faster decoding complexity and comparable performance to basis pursuit. Very similar iterative message-passing algorithms have been developed and used for the decoding of error-correcting codes [14]. More interestingly, the techniques for their analysis have recently found applicability for compressed sensing. Channel-coding-inspired message-passing decoders for compressed sensing problems have been also discussed in [15]–[17].

The fact that both iterative message-passing algorithms and LP relaxations are mathematically connected for these two problems is not an accident: For channel coding, the LP relaxation is currently understood as an approximation to iterative message-passing [18] and the theory of pseudocodewords allows a rigorous non-asymptotic theory for decoding and design of error-correcting codes under polynomial time decoding algorithms.

II. RECENT RESULTS

Assume we observe \( m \) linear measurements of an unknown vector \( \mathbf{e} \in \mathbb{R}^n \):

\[ \mathbf{H} \cdot \mathbf{e} = \mathbf{s}, \]

where \( \mathbf{H} \) is a real-valued matrix of size \( m \times n \), called the measurement matrix. When \( m < n \), this is an underdetermined system of linear equations and one fundamental compressed sensing problem involves recovering \( \mathbf{e} \) assuming that it is also \( k \)-sparse, i.e., it has \( k \) or less non-zero entries.

The sparse approximation problem goes beyond exactly sparse vectors and requires the recovery of a \( k \)-sparse vector \( \mathbf{\hat{e}} \) that is close to \( \mathbf{e} \), even if \( \mathbf{e} \) is not exactly \( k \)-sparse itself. Recent breakthrough results [19], [20] showed that it is possible to construct measurement matrices with \( m = O(k \log(n/k)) \) rows that recover \( k \)-sparse signals exactly in polynomial time. These results rely on randomized matrix constructions and establish that the optimal number of measurements will be sufficient with high probability over the choice of the matrix and/or the signal. Unfortunately, the required properties of Restricted Isometry Property (RIP) [20], Nullspace [21] and high expansion (expansion quality \( \epsilon < 1/6 \)) have no known ways to be deterministically constructed or efficiently checked.

There are several explicit constructions of measurement matrices (e.g., [22], [23]) which, however, require a slightly suboptimal number of measurements (\( m \) growing super-linearly as a function of \( n \), for \( k = p \cdot n \)).

We focus on the linear sparsity regime where \( k \) is a fraction of \( n \) and optimal number of measurements will also be a fraction of \( n \). The explicit construction of measurement matrices with an optimal number of rows is a well-known open problem in compressed sensing theory (see e.g., [24] and references therein). A closely related issue is that of checking or certifying in polynomial time that a given candidate matrix has good recovery guarantees or satisfies the RIP.

Consider a sparse matrix \( \mathbf{H} \) in \( \{0,1\}^{m \times n} \) that has \( d_v \) ones per row and \( d_c \) ones per column. If the bipartite graph
corresponding to $H$ has $\Omega(\log n)$ girth, then for $k = p \cdot n$ and an optimal number of measurements $m = c_2 \cdot n$, we show that $H$ offers $\ell_1/\ell_1$ sparse approximation under basis pursuit decoding. Our technical requirement of girth, unlike expansion or RIP, is easy to check and several deterministic constructions of matrices with $m = c \cdot n$ and $\Omega(\log n)$ exist, starting with the early construction in Gallager’s thesis [25], and progressive edge-growth Tanner graphs.

Our result is a weak bound, also known as a ‘for-every signal’ guarantee [24]. This means that we have a fixed deterministic matrix and show the $\ell_1/\ell_1$ sparse approximation guarantee with high probability over the support of the signal. To the best of our knowledge, this is the first deterministic construction of matrices with an optimal number of measurements.

Our proof relies on the discussed connections of channel decoding LP and compressed sensing. We rely on a primal-based density evolution technique initiated by Koetter and Vontobel [26] and analytically strengthened in the break-through paper of Arora et al. [27] that established the best known finite-length threshold results for LDPC codes under LP decoding. In our presentation we will survey this area, present recent results and discuss open problems.

REFERENCES


Polar Codes for the Deterministic Broadcast Channel

Naveen Goela*, Emmanuel Abbe*, and Michael Gastpar*†

* Ecole Polytechnique Fédérale de Lausanne, Faculté Informatique et Communications, CH-1015 Lausanne, Switzerland
Email: {naveen.goela, emmanuel.abbe, michael.gastpar}@epfl.ch

† University of California, Berkeley, Department of E.E.C.S., Berkeley, CA 94720-1770 USA
Email: {ngoela, gastpar}@eecs.berkeley.edu

Abstract—We construct and analyze polar codes for the deterministic broadcast channel (DBC) which is comprised of one source transmitting independent messages to m receivers. For any input distribution, polar codes with rate-splitting achieve all optimal rates within the corresponding multi-user rate region, a polyhedron in $\mathbb{R}^m_+$. For fixed m, the encoding and decoding complexities remain $O(n \log n)$ with $O(2^{-n^{\beta}})$ average block error probability where n is the block length and $0 < \beta < \frac{1}{2}$. In addition, a related joint polarization coding scheme without rate-splitting achieves points on the dominant face of the polyhedron. Our proof technique relies on bounding the total variation distance between probability measures defined on a set of length-nm binary sequences.

Index Terms—Polar codes, deterministic broadcast channel (DBC), total variation distance.

I. INTRODUCTION

Traditionally, Arvikan’s channel polarization [1] was shown to achieve the capacity of binary-input symmetric channels with low encoding and decoding complexity. Subsequently, source polarization was introduced for lossless compression of binary sources [2], lossless Slepian-Wolf communication based on a reduction to the single-source setting [2], and lossy source compression [3]. In this note, we analyze the performance of polar codes for the noiseless multi-receiver deterministic broadcast channel (DBC). Several researchers have characterized (functional and random coding) dualities between the DBC, Slepian-Wolf distributed source coding, and deterministic multiple-access channel (MAC), see e.g. [4]. Polar codes were analyzed for the latter two scenarios using both single-user and joint polarization techniques in [5] and [6]; however, polar codes for the DBC shed new light on explicit, low complexity, capacity-achieving codes in contrast to random binning and joint typicality encoding strategies.

A. The Multi-Receiver DBC

Definition 1: (Deterministic BC) The DBC with m receivers consists of an input alphabet $\mathcal{X}$, output alphabets $\{\mathcal{Y}_i\}_{i=1}^m$, and m deterministic functions $y_i = f_i(x)$ for $i \in [m]$, $x \in \mathcal{X}$, $y_i \in \mathcal{Y}_i$.

A $((2^{nR_1}, 2^{nR_2}, \ldots, 2^{nR_m}), n)$ code for the DBC with independent messages consists of an encoder

$$x^n : 2^{nR_1} \times 2^{nR_2} \times \cdots \times 2^{nR_m} \rightarrow \mathcal{X}^n,$$

and m decoders specified by $\hat{W}_i : \mathcal{Y}_i^n \rightarrow 2^{nR_i}$. The probability of error $P_e(n)$ is defined to be the probability that the decoded message is not equal to the transmitted message, i.e.,

$$P_e(n) = P \left( \bigvee_{i=1}^m \hat{W}_i(Y_i^n) \neq W_i \right),$$

where the message $(W_1, W_2, \ldots, W_m)$ is assumed to be uniformly distributed over $2^{nR_1}\times2^{nR_2}\times\cdots\times2^{nR_m}$.

Theorem 1: (Channel Capacity Region [7] [8]) Let $R = (R_1, R_2, \ldots, R_m)$. The capacity region $\mathcal{C}_{DBC}$ given deterministic functions $f_i(\cdot)$ is specified as

$$\mathcal{C}_{DBC} := \text{clo}\left( \bigcup_{X} \mathcal{R}(X) \right),$$

where

$$\mathcal{R}(X) = \left\{ R \in \mathbb{R}^m_+ \middle| \sum_{i \in S} R_i < H(Y(S)), \forall S \subseteq [m] \right\}.$$ (2)

The notation $Y(S) = \{Y_i : i \in S\}$, clo$(S)$ represents closure, and co$(S)$ the convex hull operation over set $S$. The union in Eqn. (1) is over random variable $X \sim P_X(x)$ which induces a joint distribution over $X$, $Y_1$, $Y_2$, ..., $Y_m$ via $Y_i = f_i(X)$.

B. Random Coding vs. Low Complexity Codes

Shannon-theoretic coding strategies for broadcast channels (see e.g. [9] or [10, Lec. 9] for a survey) rely on random coding. Specifically for the DBC, random binning and joint typicality encoding yield the optimal rates. At the encoder, assuming for a moment $m = 2$ receivers, a fixed $P_X(x)$ induces a joint distribution $P_Y(y_1, y_2)$. All $y_1^n$ and $y_2^n$ sequences are randomly assigned bin indices $i \in [2^{nR_1}]$ and $j \in [2^{nR_2}]$ respectively. The rates $R_1$ and $R_2$ are selected to ensure that at least one jointly typical pair $(y_1^n, y_2^n)$ exists in each $(i, j)$ product bin. To send information i to the first receiver, and information j to the second receiver, the encoder selects a jointly typical pair $(y_1^n, y_2^n)$ in the product bin $(i, j)$ and computes the $x^n$ sequence which results in $y_1^n$ and $y_2^n$. A valid $x^n$ codeword exists due to the deterministic property of $f_1(\cdot)$ and $f_2(\cdot)$. Each receiver decodes its message by looking up the bin index of its received sequence. Random binning and joint typicality encoding require a search over an exponential number of sequences.

Designing low-complexity codes for the DBC requires "binning" in a sense, and constraint satisfaction to ensure an $x^n$ codeword exists which generates the proper $y_1^n$ and $y_2^n$ output sequences. If $(y_1^n, y_2^n)$ are jointly typical with respect to
Fig. 1. The Blackwell channel with \( f_1(x) = \max(x-1, 0) \) and \( f_2(x) = \min(x, 1) \).

the distribution \( P_{Y_1,Y_2}(y_1,y_2) \) induced by \( P_X(x) \), a matching codeword \( x^* \) is guaranteed because \( P_{Y_1,Y_2}(y_1,y_2) \) has zero mass over inconsistent pairs \((y_1,y_2)\). In [11], the authors propose sparse linear coset codes to emulate binning (i.e. the syndrome is the message), combined with survey propagation to enforce the DBC channel constraints. In [12], the authors describe a reinforced belief propagation algorithm with non-linear coding for the DBC. However, going beyond ad-hoc strategies in the present paper, we design polar codes that are low-complexity and provably optimal for achieving points in the DBC capacity region.

C. Simplifying Assumptions

We make the following simplifications to Definition 1: (a) The number of receivers \( m = 2 \); (b) Polar code design is for achieving vertices of the polyhedron given in Eqn. (2) for fixed \( P_X(x) \); (c) The output alphabets \( Y_i \) are all \( F_2^m \). The first two simplifications are without loss of generality. In particular (a) is for compactness of notation and our coding scheme applies for any distribution \( P_X \), and for (b) it was shown that an arbitrary point in the achievable rate region of a DBC with \( m \) receivers maps to a corresponding vertex point in the achievable region of a DBC with \((2m-1)\) receivers if rate-splitting is used [14]. Item (c) is a mild restriction due to the algebraic construction of polar codes. It is possible to generalize the polar code construction to include output alphabets of prime, prime power, and different cardinalities.

Example 1: (Blackwell Channel) In Fig. 1, the Blackwell channel is depicted with \( X = \{0,1,2\} \) and \( Y_1 = F_2^2 \). For any fixed distribution \( P_X(x) \), it is seen that \( P_{Y_1,Y_2}(y_1,y_2) \) has zero mass for the pair \((1,0)\). Let \( \alpha \in \{1/2,3/2\} \), and \( h_\alpha(.) \) denote the binary entropy function. Due to the symmetry of this channel, the capacity region is the union of two regions,

\[
\{(R_1,R_2) : R_1 \leq h_\alpha(\alpha) \leq h_\alpha(\alpha), R_1 + R_2 \leq h_\alpha(\alpha + \alpha), \} \cup \{(R_1,R_2) : R_1 \leq h_\alpha(\alpha / 2), R_2 \leq h_\alpha(\alpha), R_1 + R_2 \leq h_\alpha(\alpha + \alpha), \}
\]

where the first region is achieved with input distribution \( P_X(0) = P_X(1) = \frac{1}{2} \), and the second region is achieved with \( P_X(1) = P_X(2) = \frac{1}{2} \) [10, Lec. 9]. The sum rate is maximized for a uniform input distribution which yields a pentagonal achievable rate region: \( R_1 \leq h_0(\frac{1}{4}), R_2 \leq h_0(\frac{3}{4}), R_1 + R_2 \leq \log_2 3 \). We construct polar codes to achieve the vertices of this pentagon, and more generally the vertices of a polyhedron in \( R^m_+ \).

II. POLAR CODES

Using the simplifications of Sec. I-C for compactness, we analyze coding over \( n \) channel uses of the DBC. Let \( m = 2 \) receivers, \( i \in [m] \) and \( j \in [n] \). The channel input is given by \( X = (X_1,X_2,...,X_n) \) where \( X_j \sim P_X \) are i.i.d. random variables. The channel output \( Y_{ij} \sim f_i(X_j) \) and \( Y_{ij},Y_{kj} \sim P_{Y_1,Y_2} \) where \( P_{Y_1,Y_2} \) is the joint distribution on \( F_2^m \) induced by \( P_X(x) \).

A. Polarization

Define the following row vectors of a matrix,

\[
\begin{bmatrix}
\bar{Y}_1 \\
\bar{Y}_2
\end{bmatrix}
\triangleq
\begin{bmatrix}
Y_{1,1} & Y_{1,2} & Y_{1,3} & \ldots & Y_{1,n} \\
Y_{2,1} & Y_{2,2} & Y_{2,3} & \ldots & Y_{2,n}
\end{bmatrix}.
\]

(3)

The joint distribution of all output variables is

\[
P_{Y_1,Y_2}(y_1,y_2) = \prod_{j=1}^{n} P_{Y_{1j},Y_{2j}}(y_{1j},y_{2j}).
\]

(4)

The polar algebraic transformation is defined in terms of an invertible matrix \( G_n \in F_2^{m \times n} \) where \( m = 2^\ell \) for \( \ell \geq 1 \) [2]. Denote the polar transformation of the output variables by

\[
\begin{bmatrix}
\bar{U}_1 \\
\bar{U}_2
\end{bmatrix} =
\begin{bmatrix}
\bar{Y}_1 \\
\bar{Y}_2
\end{bmatrix} G_n,
\]

(5)

where the transformed variables are defined

\[
\begin{bmatrix}
\bar{U}_1 \\
\bar{U}_2
\end{bmatrix}
\triangleq
\begin{bmatrix}
U_{1,1} & U_{1,2} & U_{1,3} & \ldots & U_{1,n} \\
U_{2,1} & U_{2,2} & U_{2,3} & \ldots & U_{2,n}
\end{bmatrix}.
\]

(6)

Theorem 2 (Polarization [2]): Let the random variables \( \bar{Y}_1, \bar{Y}_2, \bar{U}_1, \bar{U}_2 \) be related according to Eqsns. (3)-(6). As \( n \to \infty \), for any \( \epsilon \in (0,1) \),

\[
\begin{cases}
\{ j \in [n] : H(U_{1j}|U_{1j}^{(i-1)}) \geq 1 - \epsilon \} \to nH(Y_1), \\
\{ j \in [n] : H(U_{2j}|U_{2j}^{(i-1)},\bar{Y}_{1}) \geq 1 - \epsilon \} \to nH(Y_2|Y_{1}).
\end{cases}
\]

Theorem 2 shows how the randomness of \( \bar{Y}_1 \) and \( \bar{Y}_2 \) is extracted into \( nH(Y_1) + nH(Y_2|Y_1) \) nearly uniform components.

B. Message Indices

To design a polar code for the DBC, we identify the indices of variables in \( \bar{U}_1 \) and \( \bar{U}_2 \) that have (conditional) entropy close to one. The Bhattacharyya parameter is closely related to the conditional entropy.

Definition 2: (Bhattacharyya Parameter): Let \( (S,V) \sim P_{S,V} \) where \( S \in \{0,1\} \) and \( V \in V \) where \( V \) is an arbitrary countable alphabet. The Bhattacharyya parameter \( Z(S|V) \in [0,1] \) is defined

\[
Z(S|V) = 2 \sum_{v \in V} P_v(v) \sqrt{P_{S|V}(0|v)P_{S|V}(1|v)}.
\]

(7)

As shown by Arikan in [2], \( H(S|V) \) is close to 0 or 1 if and only if \( Z(S|V) \) is close to 0 or 1. We define the message indices for \( m = 2 \) receivers based on the Bhattacharyya parameter of variables comprising the random vectors \( \bar{U}_1, \bar{U}_2, \bar{Y}_1, \bar{Y}_2 \).

\[
\mathcal{I}_1^{(n)} = \{ j \in [n] : Z(U_{1j}|U_{1j}^{(i-1)}) \geq 1 - 2\delta_1 \},
\]

(8)

\[
\mathcal{I}_2^{(n)} = \{ j \in [n] : Z(U_{2j}|U_{2j}^{(i-1)},\bar{Y}_{1}) \geq 1 - 2\delta_2 \},
\]

(9)
where $\delta_n > 0$. According to the rate of polarization [15], we will select $\delta_n \downarrow 0$. As a result of Theorem 2 and the properties of the Bhattacharyya source parameter, for any $\tau > 0$ there exists an $n$ large enough so that both $|I_1^{(n)}| \geq n(H(Y_1) - \tau)$ and $|I_2^{(n)}| \geq n(H(Y_2|Y_1) - \tau)$.

### III. Bounding The Total Variation Distance Between Probability Measures

To code over $n$ channel uses of the DBC, we insert uniformly random message bits into the position indexed by message sets $I_1^{(n)}$ and $I_2^{(n)}$. We now begin to establish that for any $\tau > 0$, with $n \to \infty$, the message rates $R_1 \geq H(Y_1) - \tau$ and $R_2 \geq H(Y_2|Y_1) - \tau$ are achievable with $P_{\sigma}^{(n)} \to 0$. For a DBC with $m = 2$ receivers, the rate pair $(R_1, R_2)$ approaches one vertex point of the pentagonal rate region (defined more generally as a polyhedron in Eqn. 2). Other optimal vertex points of the polyhedral rate region are achieved similarly by changing the ordering of receivers.

#### A. Two Probability Measures on a Set of Binary Sequences

We define two probability measures on the set of all $mn$-length binary sequences (here we assume $m = 2$) which will help in bounding the average block error probability $P_{\sigma}^{(n)}$ of our encoding/decoding schemes. The i.i.d. joint distribution $P_{\sigma_1, \sigma_2}(\bar{u}_1, \bar{u}_2)$ defined in Eqn. (4) specifies a joint distribution $P_{\sigma_1, \sigma_2}(\bar{u}_1, \bar{u}_2)$ via the one-to-one linear transform $G_n$ of Eqn. (5). We write this distribution as

$$P_{\sigma_1, \sigma_2}(\bar{u}_1, \bar{u}_2) = P_{U_1}(\bar{u}_1)P_{U_2|U_1}(\bar{u}_2|\bar{u}_1).$$

We further factorize the above as follows,

$$P_{U_{1j}}(\bar{u}_1) = \prod_{j=1}^{n} P_{U_{1j}}|U_{11}^{j-1}(u_{1j}|u_{11}^{j-1}),$$

$$P_{U_{2j}|U_{1j}}(\bar{u}_2|\bar{u}_1) = \prod_{j=1}^{n} P_{U_{2j}|U_{1j}^{j-1}, U_{1j}}(u_{2j}|u_{1j}^{j-1}, \bar{u}_1).$$

We now define closely related conditional probability distributions $\forall j \in [n]$,

$$Q_{U_{1j}|U_{11}^{j-1}, U_{1j}}(u_{1j}|u_{11}^{j-1}) \begin{cases} \frac{1}{2}, & \text{if } j \in I_1^{(n)}, \\ \frac{1}{2}, & \text{otherwise}. \end{cases}$$

$$Q_{U_{2j}|U_{21}^{j-1}, U_{2j}}(u_{2j}|u_{21}^{j-1}, \bar{u}_1) \begin{cases} \frac{1}{2}, & \text{if } j \in I_2^{(n)}, \\ \frac{1}{2}, & \text{otherwise}. \end{cases}$$

Then analogous to Eqn. (11) and Eqn. (12),

$$Q_{U_{1j}}(\bar{u}_1) \triangleq \prod_{j=1}^{n} Q_{U_{1j}|U_{11}^{j-1}, U_{1j}}(u_{1j}|u_{11}^{j-1}),$$

$$Q_{U_{2j}|U_{21}^{j-1}, U_{2j}}(\bar{u}_2|\bar{u}_1) \triangleq \prod_{j=1}^{n} Q_{U_{2j}|U_{21}^{j-1}, U_{2j}}(u_{2j}|u_{21}^{j-1}, \bar{u}_1).$$

Essentially $Q_{U_{1j}}(\bar{u}_1) = (\frac{1}{2})^{2^{(n)}}g_1(\bar{u}_1)$ and $Q_{U_{2j}|U_{21}^{j-1}, U_{2j}}(\bar{u}_2|\bar{u}_1) = (\frac{1}{2})^{|I_2^{(n)}|}g_2(\bar{u}_2|\bar{u}_1)$ where $g_1(\bar{u}_1)$ and $g_2(\bar{u}_2|\bar{u}_1)$ are nearly indicator functions; i.e., the distributions are approximately composed of i.i.d. Bernoulli(1/2) random variables.

### B. Total Variation Bound

**Theorem 3:** (Total Variation Bound for Joint Distributions)

Let probability measure $P_{\sigma_1, \sigma_2}(\bar{u}_1, \bar{u}_2)$ be defined as in Eqn. (10). Let $Q_{\sigma_1, \sigma_2}(\bar{u}_1, \bar{u}_2) = Q_{U_{1j}}(\bar{u}_1)Q_{U_{2j}|U_{1j}}(\bar{u}_2|\bar{u}_1)$ be defined via Eqn. (13) and Eqn. (14). Then the total variation distance between $P$ and $Q$ is bounded as

$$\sum_{\bar{u}_1, \bar{u}_2} |Q(\bar{u}_1, \bar{u}_2) - P(\bar{u}_1, \bar{u}_2)| \leq 2|I_1^{(n)}|\delta_n + 2|I_2^{(n)}|\delta_n,$$

where we set $\delta_n = \frac{2^{-n\beta}}{3n}$ for $0 < \beta < \frac{1}{2}$. Thus the total variation distance is bounded by $O\left(2^{-n\beta}\right)$.

**Proof:** Let probability distribution $\mathcal{L}_{U_1, U_2}(\bar{u}_1, \bar{u}_2) \triangleq P_{U_1}(\bar{u}_1)Q_{U_2|U_1}(\bar{u}_2|\bar{u}_1)$. Observe by the triangle inequality for total variation distance,

$$\sum_{\bar{u}_1, \bar{u}_2} |Q(\bar{u}_1, \bar{u}_2) - P(\bar{u}_1, \bar{u}_2)| \leq \sum_{\bar{u}_1, \bar{u}_2} |Q(\bar{u}_1, \bar{u}_2) - \mathcal{L}(\bar{u}_1, \bar{u}_2)|$$

$$+ \sum_{\bar{u}_1, \bar{u}_2} |\mathcal{L}(\bar{u}_1, \bar{u}_2) - P(\bar{u}_1, \bar{u}_2)|.$$

We bound the first term as follows,

$$\sum_{\bar{u}_1, \bar{u}_2} |Q(\bar{u}_1, \bar{u}_2) - \mathcal{L}(\bar{u}_1, \bar{u}_2)|$$

$$= \sum_{\bar{u}_1} \sum_{\bar{u}_2} Q_{U_{1j}}(\bar{u}_1)Q_{U_{2j}|U_{1j}}(\bar{u}_2|\bar{u}_1) - P_{U_1}(\bar{u}_1),$$

$$= \sum_{\bar{u}_1} \sum_{\bar{u}_2} Q_{U_{1j}}(\bar{u}_1) - P_{U_1}(\bar{u}_1) \left[ \sum_{\bar{u}_2} Q_{U_{2j}|U_{1j}}(\bar{u}_2|\bar{u}_1) \right],$$

$$= \sum_{\bar{u}_1} \sum_{\bar{u}_2} Q_{U_{1j}}(\bar{u}_1) - P_{U_1}(\bar{u}_1),$$

$$\leq 2 \sum_{\bar{u}_1} \sum_{j \in L_1^{(n)}} \left| \frac{1}{2} - P_{U_{1j}|U_{11}^{j-1}, U_{1j}}(0|U_{11}^{j-1}) \right|,$$

$$\leq 2|I_1^{(n)}|\delta_n.$$ (15)

Similarly, the second term is bounded.

$$\sum_{\bar{u}_1, \bar{u}_2} |\mathcal{L}(\bar{u}_1, \bar{u}_2) - P(\bar{u}_1, \bar{u}_2)|$$

$$= \sum_{\bar{u}_1} \sum_{\bar{u}_2} Q_{U_{2j}|U_{1j}}(\bar{u}_2|\bar{u}_1) - P_{U_{2j}|U_{21}^{j-1}, U_{2j}}(\bar{u}_1)$$

$$\leq \sum_{\bar{u}_1} \sum_{j \in L_2^{(n)}} \left| \frac{1}{2} - P_{U_{2j}|U_{21}^{j-1}, U_{2j}}(0|U_{21}^{j-1}, U_{1j}) \right|,$$

$$\leq 2|I_2^{(n)}|\delta_n.$$ (16)

The inequalities in Eqns. (15)-(18) are derived in a very similar way to the inequalities in the rate-distortion analysis of polar codes for (single) source coding in [3, Lemma 4, Lemma 7] and will be omitted here.

### IV. POLAR CODING FOR THE DBC

#### A. Encoding and Decoding Steps

Assuming $m = 2$ receivers, the encoder’s task is to map a message $(W_1, W_2)$ uniformly distributed over $2^{nR_1} \times 2^{nR_2}$ to a codeword $\alpha^n$. The polar encoding strategy is to construct first a pair of $n$-length binary sequences $\bar{u}_1$ and $\bar{u}_2$. The bits $u_{1j}$ with indices $j \in I_1^{(n)}$ are set to the uniformly selected
message bits for the first receiver. Likewise, the bits $u_{2j}$ with indices $j \in \mathcal{I}_2^{(n)}$ are set to the uniformly selected message bits for the second receiver. Note that the message indices $\mathcal{I}_1^{(n)}$ and $\mathcal{I}_2^{(n)}$ defined in Eqn. (8) and Eqn. (9) are computed once during polar code construction, and do not depend on the realization of variables during the encoding process. The non-message indices in $u_1$ and $u_2$ are obtained via a bit-by-bit successive randomized calculation as follows.

$$
\forall j \notin \mathcal{I}_1^{(n)}: u_{1j} = \begin{cases} 0, & \text{w.p. } P_{U_1[u_1]}(1|u_{11}), \\
1, & \text{w.p. } P_{U_1[u_1]}(0|u_{11}).
\end{cases}
$$

$$
\forall j \notin \mathcal{I}_2^{(n)}: u_{2j} = \begin{cases} 0, & \text{w.p. } P_{U_2[u_2]}(1|u_{21}), \\
1, & \text{w.p. } P_{U_2[u_2]}(0|u_{21}).
\end{cases}
$$

The probabilities $P_{U_1[u_1]}(1|u_{11})$ and $P_{U_2[u_2]}(1|u_{21})$ above are efficiently computable in $O(n \log n)$ time due to Arıkan’s recursive polarization decomposition (see [2] for the recursive likelihood formulas). Having formed the binary sequences $u_1$ and $u_2$, the encoder applies the inverse polar transform, $(y_1; y_2) = (u_1; u_2) G_n^{-1}$ where $G_n^{-1} = G_n$ in the binary case. Each channel symbol $x_j$ of the codeword $x^n$ is formed by taking any $x_j \in \cap_{i=\ell}^{n-1} F_i^{-1}(y_{ij})$ if this intersection set is non-empty, otherwise the decoder declares a block error. If the encoder succeeds in transmitting a codeword, each receiver $i \in [m]$ obtains the sequence $y_i$ exactly and applies the $G_n$ transform to recover $u_i$ exactly. Knowing the location of the message indices in $u_i$, each receiver decodes its own message correctly.

B. Error Performance

An error event for the DBC occurs exclusively at the encoder if it is unable to construct a codeword $x^n$. In the following theorem, we assume $m = 2$. $P_{Y_1,Y_2}(y_1,y_2)$ is the induced distribution for arbitrary input distribution $P_X$, and the target rate pair is $(R_1, R_2) = (H(Y_1), H(Y_2|Y_1))$.

**Theorem 4:** (Average Block Error Probability) Fix $R_1 < H(Y_1)$, $R_2 < H(Y_2|Y_1)$, and $0 < \beta < 1$. There exists a sequence of polar codes of length $n$ for which the distribution achieves rates $(R_1, R_2)$, so that under successive randomized encoding, the expected average block error probability is $P_e^{(n)} = O(2^{-n^{\beta}})$.

**Proof:** Theorem 2 established that the message sets have cardinalities $|\mathcal{I}_1^{(n)}| \rightarrow n H(Y_1)$ and $|\mathcal{I}_2^{(n)}| \rightarrow n H(Y_2|Y_1)$. Let $\mathcal{E}$ denote the set of all $2^n$-length binary sequences written as $(y_1^n, y_2^n)$ which are inconsistent, i.e. $\exists j \in [n]$ for which $(y_{1j}, y_{2j})$ is not a valid output of the DBC for any input symbol $x_j$. Let $\mathcal{E}$ denote the set of $2^n$-length binary sequences $(u_1^n, u_2^n)$ that correspond to the inconsistent sequences via the one-to-one polar transform $G_n$. The distribution $P$ is such that $\sum_{(u_1, u_2) \in \mathcal{E}} Q_{U_1[U_1]}(u_1, u_2) = 0$. Due to Theorem 3,

$$
\sum_{(u_1, u_2) \in \mathcal{E}} Q_{U_1[U_1]}(u_1, u_2) = O(2^{-n^{\beta}}).
$$

The successive randomized encoding procedure of Sec. IV-A selects a codeword by effectively drawing from the distribution $Q$ for a uniformly selected message. The expected average block error probability of drawing an inconsistent pair is therefore $P_e^{(n)} = O(2^{-n^{\beta}})$.

V. JOINT/SIMULTANEOUS POLARIZATION METHOD

Let matrix $\bar{U} = [\bar{U}_1; \bar{U}_2]$ where $\bar{U}_j \in F_{2^{n/2}}^{\infty}$ was defined in Eqn. (6). We make a brief remark about joint polarization methods for the DBC. It was shown in [5, Lemma 1] that,

$$
\frac{1}{n} \left| \left\{ j \in [n] : H(U_j[|S|](U_j^{-1}) \in \bigcup_{r=0}^{S} [r - e, r + e], \forall S \subseteq [m] \right\} \rightarrow 1, \right.
$$

where here the new notation $U_j$ now denotes the $j$-th column of matrix $\bar{U}$, $U_j(i)$ the $(i, j)$-entry of matrix $\bar{U}$, $U_j[|S|] = [U_j(i) : i \in S]$, and $U_j = [U_1, \ldots, U_m]$. When comparing Arıkan’s polarization of Theorem 2 with that of Eqn. (19), one sees that the latter involves joint polarization column by column, whereas the former involves successive polarization bit by bit. It can be shown that joint polarization methods allow for DBC coding schemes that achieve (sum rate) points on the dominant face of the polyhedron in $R^m_+$ without rate-splitting.

VI. CONCLUSION

The polar encoding strategy is a non-linear mapping from $m$ messages to one codeword that requires $O(n \log n)$ complexity to compute. We believe polar codes will be applied in other network communication scenarios to address the high complexity of random binning and joint typicality coding.

ACKNOWLEDGMENT

The authors thank S. B. Korada for the helpful discussions regarding the journal paper [3].

REFERENCES


Polar Codes for Coordination in Cascade Networks

Ricardo Blasco-Serrano, Ragnar Thobaben, and Mikael Skoglund
KTH Royal Institute of Technology and ACCESS Linnaeus Centre
SE-100 44, Stockholm, Sweden
E-mail: {ricardo.blasco, ragnar.thobaben, mikael.skoglund}@ee.kth.se

Abstract—We consider coordination in cascade networks and construct sequences of polar codes that achieve any point in a special region of the empirical coordination capacity region. Our design combines elements of source coding to generate actions with the desired type with elements of channel coding to minimize the communication rate. Moreover, we bound the probability of malfunction of a polar code for empirical coordination. Possible generalizations and open problems are discussed.

I. INTRODUCTION

The limits of coordination in networks have been recently studied from a mathematical point of view (e.g. [1], [2]). Simple questions like how to measure coordination or how much communication is necessary to achieve a desired level of coordination have been posed and to some extent answered. The authors of [1] developed elements of a fundamental theory with two notions of coordination. In the first one, empirical coordination, the sequences of actions generated in the network must have a type that is close to a desired probability distribution. Empirical coordination is closely related to rate-distortion theory [3]. In fact any good code for rate-distortion is useful for empirical coordination and vice versa [1]. The second notion is that of strong coordination. Here the sequences of actions generated by the nodes must be statistically indistinguishable from those obtained by sampling a certain distribution.

In this work we construct sequences of polar codes (PCs) that achieve a special region of the empirical coordination capacity region for two-node and three-node cascade networks. PCs were introduced by Ankan as a method to achieve the capacity of any symmetric binary-input discrete memoryless channel (BI-DMC) [4]. Since then they have emerged as a powerful technique to develop achievability results in information theory with structured codes (as opposed to random coding). Korada and Urbanke established in [5] the optimality of PCs for lossy source coding of (symmetric) discrete memoryless sources (DMS) with binary reproduction alphabets. These results were later extended to non-binary channels and reproduction alphabets in [6] and [7], respectively.

Our constructions combine elements of PCs for source compression with PCs for channel coding. In addition, we show that, in line with [1], common randomness is not necessary for implementing PCs for empirical coordination, although it is useful in the proofs. We use the properties of PCs to extend the application of results on the rate of polarization [8] to PCs designed for empirical coordination. Finally, we discuss possible generalizations of our constructions as well as some open problems for future research.

This paper is organized as follows. In Section II we summarize basic results on empirical coordination and polar codes along with the notation. We analyze a simple two-node network in Section III. This serves as a building block for the cascade network, which is addressed in Section IV. We conclude our work in Section V with a discussion on the results of the paper as well as on some open problems.

II. PRELIMINARIES

A. Notation

Scalars are written using normal face x and vectors using bold face x. The i\textsuperscript{th} element of a vector x is denoted by x\textsubscript{i}. For a given set of natural numbers F with size |F|, x\textsubscript{F} is shorthand for the subvector with elements whose positions belong to F. We use upper case letters for random variables (RVs) Y and lower case letters for their realizations y. The joint probability distribution on (X, Y) is denoted by P\textsubscript{XY}(x, y). For convenience we shall alternatively drop the subindices or the arguments whenever they are clear from the context. We follow the standard information-theoretic notation from [3].

B. Empirical Coordination

Consider the three-node network in Fig. 1. Node X observes a sequence of N external actions X chosen independently and identically distributed (i.i.d.) according to P\textsubscript{X}. Communication from Node X to Node Y and from Node Y to Node Z is possible at rates R\textsubscript{1} and R\textsubscript{2} (bits per action, respectively). We use these resources to have Node Y and Node Z generate sequences of actions Y and Z, respectively, with length N and joint type close to a desired probability function P\textsubscript{YZ|X}\textsuperscript{T}.

\[ X \sim P\textsubscript{X} \]

\[ \begin{array}{c}
\text{Node X} \\
\uparrow \\
R\textsubscript{1} \\
\text{Node Y} \\
\downarrow \\
Y \\
\text{Node Z} \\
\downarrow \\
R\textsubscript{2} \\
\text{Node Z} \\
\uparrow \\
Z \\
\end{array} \]

Fig. 1. Cascade network.

A (2\textsuperscript{NR\textsubscript{1}}, 2\textsuperscript{NR\textsubscript{2}}, N) coordination code for the network in Fig. 1 consists of an encoding, a recoding, and two decoding functions (see [1]). All of them may have access to a source of common randomness (CR) independent of the external actions. Each coordination code induces a joint distribution on the actions Q\textsubscript{XYZ}.

In this paper we are interested in the joint type of a tuple of action sequences (x, y, z) which is defined as

\[ T\textsubscript{xyz}(x, y, z) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{ (x_i, y_i, z_i) = (x, y, z) \} \]
for all \((x, y, z) \in X \times Y \times Z\), where \(\{\cdot\}\) is the indicator function. In order to measure the distance between two probability distributions \(P_{X,Y,Z}\) and \(Q_{X,Y,Z}\) we consider their total variation, which is defined as

\[
\|P_{X,Y,Z} - Q_{X,Y,Z}\| = \frac{1}{2} \sum_{x,y,z} |P(x, y, z) - Q(x, y, z)|.
\]

We say that a triple \((R_1, R_2, P_{X,Y,Z})\) is achievable for empirical coordination if for any \(\epsilon > 0\) there exists a sequence of \((2^{NR_1}, 2^{NR_2}, N)\) coordination codes and a choice of CR such that

\[
\text{Pr}(\|P_{X,Y,Z} - T_{X,Y,Z}\| > \epsilon) < \epsilon
\]

for sufficiently large \(N\) under the distribution induced by the codes. The empirical coordination capacity region, denoted by \(C_{EP}\), is the closure of the set of achievable triples \((R_1, R_2, P_{X,Y,Z})\).

Due to the nature of PCs we restrict our attention to binary actions \(Y\) and \(Z\) (although no restriction is placed on \(X\)) and to choices of \(P_{Y,Z|X}\) that induce uniform distributions on \(X\) and \(Z\) for the given \(P_X\). All the results in this paper are restricted to this subset of \(C_{EP}\) that we shall refer to as the symmetrical empirical coordination capacity region and denote by \(C_{EP}^{\text{sym}}\). Possible generalizations are discussed in Section V.

C. Polar Codes

Channel polarization is a method for transforming

\[N\]

identical copies of a BI-DMC \(P_{Y|X}\) into \(N\) distinct BI-DMCs \(P^{(i)}(y, u^{(i-1)}_i|u_i)\) \((i \in \{1, \ldots, N\})\) with extremal properties in the sense that, for sufficiently large \(N\), a fraction \(I(X; Y)\) (for \(X\) uniformly distributed) of these synthetic channels is noise-free while the rest is virtually useless.

1) Channel Coding: Channel polarization leads naturally to a code construction that achieves the capacity of any symmetric BI-DMC. These codes are known as PCs and consist of two elements: an encoding matrix \(G_N\) and a Successive Cancellation (SC) decoding algorithm. \(G_N\) synthesizes the channels with extremal properties. Fixed (i.e. frozen) bits are put into the “bad” channels, which are those in the frozen set \(F\) defined as

\[
F = \{i : Z(P^{(i)}) \geq \delta_N\}
\]

for some \(\delta_N > 0\), where \(Z(P^{(i)})\) denotes the Bhattacharyya parameter (an upper bound on the error probability for uncoded transmission \([4]\)) of the BI-DMC \(P^{(i)}\). Information is transmitted at full rate through the rest of channels, i.e. those in the complement of the frozen set \(F\). The SC decoding algorithm generates sequentially estimates \(\hat{u}_i\) for the information bits using the channel distribution \(P^{(i)}(y, u^{(i-1)}_i|u_i)\) and the previous estimates.

The following observation about PCs designed for degraded channels (as defined in \([3]\)) will turn out to be important in the sequel.

1We restrict our attention to PCs as introduced in \([4]\) and therefore consider vectors of \(N = 2^n\) actions (with \(n \in \mathbb{N}\)).

Lemma 1 (Lemma 21 in \([5]\)). Let \(P_{Y|X}\) and \(\tilde{P}_{Y|X}\) be two BI-DMCs with \(\tilde{P}_{Y|X}\) degraded with respect to \(P_{Y|X}\) (denoted by \(\tilde{P}_{Y|X} \preceq P_{Y|X}\)). Then for all \(i \in \{1, \ldots, N\}\) the synthesized channels satisfy \(\tilde{P}^{(i)} \preceq P^{(i)}\) and \(Z(\tilde{P}^{(i)}) \geq Z(P^{(i)})\).

2) Source Coding: Korada and Urbanke showed in \([5]\) that PCs are also optimal for lossy source compression of a DMS \(X \sim P_X\) into \(Y\). In this context the SC algorithm is used for compression while reconstruction is performed with the matrix \(G_N\). PCs for source compression are designed upon the test channel \(P_{X|Y}\) that solves the minimization defining the (symmetric) rate-distortion function. Their fundamental property is that the distribution they induce \(Q_{X,N}\) is arbitrarily close to the \(N\)-product design distribution \(P_{X,Y} = \prod P_{Y|X} P_X\) as summarized in the following Lemma:

Lemma 2 (Lemmas 2, 5, and Theorem 16 in \([5]\)). For any compression rate \(R > I(Y; X)\) \((Y\) uniformly distributed\) we can choose \(\delta_N > 0\) such that if the frozen set \(F\) is defined as

\[
F = \{i : Z(P^{(i)}) \geq 1 - \delta_N\}
\]

then for any \(0 < \beta < \frac{1}{2}\) we have

\[
\frac{1}{2} \sum_{x,y} |P(x, y) - Q(x, y)| \leq O(2^{-N\beta}).
\]

We will combine this result with the optimal coupling. For any two RVs \(X \sim P_X\) and \(Y \sim Q_Y\) defined on the same space their optimal coupling \(C_{PQ}(x, y)\) is a probability distribution with marginals \(P_X\) and \(Q_Y\) and that satisfies \([5, 9]\)

\[
\text{Pr}(X \neq Y) = \frac{1}{2} \sum_{x,y} |P_X(x) - Q_Y(y)|.
\]

III. Two-Node Network

In this Section we restrict our attention to the two-node network in Fig. 2. A \((2^{NR}, \tilde{P}_{Y|X})\) coordination code for this scenario consists only of an encoding and a decoding function.

\[
X \sim P_X
\]

\[
\text{Node } X \xrightarrow{R} \text{Node } Y
\]

Fig. 2. Two-node network.

The empirical coordination capacity region for the network in Fig. 2 is given by \([1]\)

\[
C_{EP} = \{(R, P_{Y|X}) : R \geq I(Y; X)\}.
\]

Theorem 1. For any pair \((R, P_{Y|X}) \in C_{EP}\) (i.e. considering the aforementioned additional restrictions with respect to \(C_{EP}\)) there exists a sequence of PCs that achieves empirical coordination in the two-node network in Fig. 2.

Proof: Let \(P_{Y|X}\) be the desired distribution and let

\[
P_{X|Y} = \frac{P_{X|Y} P_X}{P_Y}
\]

where \(P_Y\) is the uniform distribution. Construct a PC for source compression upon \(P_{X|Y}\) with the frozen set defined
as in (3). For any compression rate \( R > I(Y;X) \) we know from the results on PCs for source coding in [5] that \( R > \frac{E}{N} \) if \( N \) is sufficiently large. Using the CR both nodes generate \( F \) i.i.d. random bits \( U_F \) according to a uniform distribution. Node \( X \) compresses its observation \( X \) using the SC encoding algorithm with frozen bits set to \( U_F \). The output of the algorithm is the vector \( U_F \) which contains \(|F|\) information bits. Therefore a rate of \( R \) bits per action suffices to communicate them to Node \( Y \). This allows Node \( Y \) to generate the sequence \( Y = \text{UG}_Y \). From Lemma 2 we know that the induced distribution is close to the desired \( N \)-product distribution \( P_{X,Y} = \prod P_{Y|X} P_X \):

\[
\|P_{X,Y} - Q_{X,Y}\| = \frac{1}{2} \sum_{x,y} |P(x,y) - Q(x,y)| \leq O(2^{-N^2})
\]

for any \( 0 < \beta < \frac{1}{2} \).

We now verify that this implies empirical coordination. Let \( C_{PQ}(x_P, y_P, x_Q, y_Q) \) denote the optimal coupling between \( P_{X,Y}(x_P, y_P) \) and \( Q_{X,Y}(x_Q, y_Q) \). Consider the events

\[
\mathcal{E} = \{ \|P_{X,Y} - T_{X,Y} Q_{X,Y}\| > \beta \},
\]

\[
\mathcal{E}_{XY} = \{ (X_P, Y_P) \neq (X_Q, Y_Q) \},
\]

and their complements \( \mathcal{E}^c \) and \( \mathcal{E}_{XY}^c \). Since \( C_{PQ} \) has marginal \( Q_{X,Y}(x_Q, y_Q) \), \( \mathcal{E} \) evaluated over \( C_{PQ} \) is our event of interest. We upper bound its probability using the properties of the optimal coupling (see (4)). For any \( 0 < \beta < \frac{1}{2} \) we have that

\[
\text{Pr}(\mathcal{E}) = \text{Pr}(\mathcal{E}|\mathcal{E}_{XY}) \text{Pr}(\mathcal{E}_{XY}) + \text{Pr}(\mathcal{E}|\mathcal{E}_{XY}^c) \text{Pr}(\mathcal{E}_{XY}^c) \
\leq \text{Pr}(\mathcal{E}_{XY}) + \text{Pr}(\mathcal{E}|\mathcal{E}_{XY}^c) \leq O(2^{-N^2}) \quad (5)
\]

The second term in (5) is upper bounded by the probability that a tuple \( (X, Y) \sim P_{X,Y} \) has a type with total variation with respect to \( P_{X,Y} \) larger than \( \epsilon \). From the properties of the Asymptotic Equipartition Property (AEP) (see [1], [3]) we know that this probability goes to zero exponentially fast with \( N \) Therefore the bound for the first term in (5), which is due to the properties of the optimal coupling, dominates. Hence we conclude that the sequence of PCs achieves empirical coordination.

The above proof requires both nodes to generate the frozen bits using CR. The following corollary shows that there exists a fixed choice of the values of these bits such that the sequence of PCs achieves empirical cooperation.

**Corollary 1.** For any pair \((R, P_{Y|X}) \in C_{P_{X,Y}}\) there exists a sequence of PCs and a choice of frozen bits \( U_F \) that achieves empirical coordination in the two-node network in Fig. 2. □

**Proof:** The proof is given in the Appendix.

### IV. CASCADE NETWORK

Consider again the network given in Fig. 1. Its empirical coordination capacity region is given by [1]

\[
C_{P_X} \triangleq \begin{cases} (R_1, R_2, P_{Y|X}) : R_1, R_2 \geq I(Y; Z; X) \\ R_1 \geq I(Y; Z; X) \end{cases}
\]

**Theorem 2.** For any triple \((R_1, R_2, P_{Y|X}) \in C_{P_{X,Y}}\) (i.e. considering the aforementioned additional restrictions with respect to \( C_{P_{X,Y}} \)) there exists a sequence of PCs that achieves empirical coordination in the cascade network in Fig. 1. □

**Proof:** Let \( P_{Y|X,Z} \) be the desired distribution. Design a PC for source compression based on \( P_{X|Z} \) obtained from \( P_{Y|X,Z} P_X \) by conditioning and marginalizing. For any compression rate \( R > I(Z; X) \) (for \( Z \) uniformly distributed) and any \( 0 < \beta < \frac{1}{2} \) we know that

\[
\|P_{X,Z} - Q_{X,Z}\| \leq O(2^{-N^2})
\]

where \( P_{X,Z} = \prod P_{X|Z} P_X \) and \( Q_{X,Z} \) is the distribution induced by the code. The values of the frozen bits are chosen i.i.d. according to a uniform distribution using CR so that they are also available at Nodes \( Y \) and \( Z \). The vector resulting from source compression is sent to Node \( Y \) and then forwarded to Node \( Z \). This allows both nodes to generate \( Z \). Therefore

\[
R_1 \geq R_1' + I(Z; X) \quad R_1 \geq I(Z; X)
\]

where \( R_1' \) is the fraction of \( R_1 \) which is not yet used. Clearly we need \( R_1' \) to be arbitrarily close to \( I(Y; Z; X) - I(Y; X) \).

Now assume that \((X, Z) \sim P_{X,Z} \) (instead of \( Q_{X,Z} \)). Design a new PC for compression of \((X, Z)\) into \( Y \) using the conditional probability \( P_{X|Z} \). Let \( \tilde{Q}_{X,Y,Z} \) be the distribution induced by this new PC. We know from Section III that for any compression rate \( R > I(Y; X, Z) \) for \((Y, Z)\) uniformly distributed and any \( 0 < \beta < \frac{1}{2} \) we can choose the frozen set \( F' \) as in (3) so that for \( P_{Y|X,Z} \)

\[
\|P_{Y|X,Z} P_{X,Z} - \tilde{Q}_{X,Y,Z} P_{X,Z}\| \leq O(2^{-N^2}) \quad (7)
\]

Using the triangle inequality it is easy to show that (6) and (7) imply that for the true distribution on \((X, Z)\), i.e. \( Q_{X,Z} \),

\[
\|P_{Y|X,Z} P_{X,Z} - \tilde{Q}_{X,Y,Z} Q_{X,Z}\| \leq O(2^{-N^2}) \quad (8)
\]

for any \( 0 < \beta < \frac{1}{2} \). As in Section III we use this result to build the optimal coupling \( C_{P_{X,Y,Z}} \) between \( P_{X,Y,Z} \) and \( \tilde{Q}_{X,Y,Z} \) which satisfies, for any \( 0 < \beta < \frac{1}{2} \),

\[
\text{Pr}(\mathcal{E}_{XY} Z) \leq \text{Pr}(\mathcal{E}_{XY} Z) \leq O(2^{-N^2})
\]

This allows us to evaluate the probability of the event

\[
\mathcal{E}_1 \triangleq \{ \|P_{X,Y,Z} - T_{X,Y,Z}\| > \epsilon \}
\]

and bound it, for any \( 0 < \beta < \frac{1}{2} \), as

\[
\text{Pr}(\mathcal{E}_1) \leq \text{Pr}(\mathcal{E}_{XY} Z) \leq O(2^{-N^2})
\]

This means that our construction will yield sequences of actions with the desired type with high probability. Now we only need to make \( Y \) available at Node \( Y \). To accomplish this at the desired rate we take advantage of the fact that \( Z \) is already available at Node \( Y \) and that it is indeed correlated with \( Y \). We model this correlation as transmitting \( Y = \text{UG}_Y \) (i.e. a codeword from a PC) through the DMC \( P_{Z|Y} \) and use the SC decoding algorithm to obtain \( Y \), an estimate of \( Y \).
Such a PC for transmission over $P_{Z\mid Y}$ would have a frozen set $F_V$ as defined in (2). To obtain $Y$ the decoder needs to have access to the frozen bits $U_{F_Y}$. Moreover, since due to Lemma 1 we know that $F_Q \subseteq F_V$ for $\delta_N^2 < \delta_N$ and sufficiently large $N$, only the bits in the set $F_V \cap F_Q$ need to be conveyed (see Figure 3). This can be done at any rate

$$R'_1 \geq \frac{|F_V| - |F_Q|}{N} \geq I(Y;X,Z) - I(Y;Z),$$

or equivalently

$$R'_1 \geq I(Y,Z;X) - I(Z;X),$$

as desired.

It is easy to show that this design will also work fine when $(\tilde{Y},Z) \sim Q_{\tilde{Y},Z}$ by considering the optimal coupling between $P_{Y,Z}$ and $Q_{\tilde{Y},Z}$ and the fact that, for any $0 < \beta < \frac{1}{2}$,

$$\frac{1}{2} \sum_{y,z} |P_{Y,Z}(y,z) - Q_{\tilde{Y},Z}(y,z)| \leq O(2^{-N^{\beta}}),$$

which is derived from (8) with the triangle inequality. That is,

$$\Pr(\mathcal{E}_Y) \triangleq \Pr(Y \neq \tilde{Y}) \leq O(2^{-N^{\beta}})$$

for any $0 < \beta < \frac{1}{2}$.

Finally, we see that the probability of the event

$$\mathcal{E} = \{|P_{X,Y,Z} - T_{X,Y,Z}| > \epsilon\},$$

when evaluated over the distribution induced by the code

$$Q_{X,Y,Z} = P_X Q_{Z\mid X} Q_{Y\mid X,Z} Q_{V\mid \tilde{Y}},$$

where $Q_{V\mid \tilde{Y}}$ accounts for the possible errors in reproducing $\tilde{Y}$ at Node $Y$, is arbitrarily low. That is, for any $0 < \beta < \frac{1}{2}$

$$\Pr(\mathcal{E}) \leq \Pr(\mathcal{E}_Y) + \Pr(\mathcal{E}_Y^c) \leq O(2^{-N^{\beta}}).$$

Hence the sequence of PC achieves empirical coordination. 

As for the two-node network, there exists a choice of the values of the frozen bits such that the sequence of PCs achieves empirical coordination.

**Corollary 2.** For any triple $(R_1, R_2, P_{Y,Z\mid X}) \in C_{\mathcal{P}_N}^c$ there exists a sequence of PCs and a choice of frozen bits that achieves empirical coordination in the network in Fig. 1.

**Proof:** The proof is similar to that of Corollary 1. 

V. DISCUSSION

The previous sections show that PCs are suitable for empirical coordination. However, a few questions remain open. First of all, our discussion has been restricted to the symmetrical empirical coordination capacity region, a portion of the region $C_{\mathcal{P}_N}$. Preliminary results show that the usual techniques for extending the applicability of PCs (see [4]–[7]) may also be useful here to make more general statements as one would expect from [1, Section VII]. Moreover, we have only considered empirical coordination. No practical codes for strong coordination have been designed so far. The problem gets more involved due to the need for exploiting common randomness.

**ACKNOWLEDGMENT**

The authors would like to thank one anonymous reviewer for helpful comments.

**APPENDIX**

**Proof of Corollary 1:** Since the total variation is a bounded measure, Theorem 1 implies that there is a sequence of coordination codes such that

$$\lim_{N \to \infty} \mathbb{E}[X,Y] (\{||T_{X,Y} - P_{X,Y}||\}) = 0.$$  

The expectation, which is taken with respect to the induced distribution, can be written as

$$\mathbb{E}_{X,Y} (\{||T_{X,Y} - P_{X,Y}||\}) = \mathbb{E}_{X,u} (\{||T_{X,u} - P_{X,Y}||\})$$

$$= \sum_{u\in\mathcal{F}} Q(x,u) ||T_{X,u} - P_{X,Y}||$$

$$= \sum_{u\in\mathcal{F}} Q(u_f) \sum_{x,u_f} Q(x,u_f) ||T_{X,u} - P_{X,Y}||$$

$$= \mathbb{E}_{u_f} (\{||T_{X,u_f} - P_{X,Y}||\}).$$

In (9) and (10) we have used the fact that $G_N$ is a one-to-one mapping to modify slightly the definitions of $Q_{X,Y}$ and of the type function to depend on $u$ by setting $y = uG_N$. The notation $T_{X,u_f} = T_{X,u}$ in (11) simply makes explicit the fact that $u = (u_f, u_f)$. For the outer expectation in (11) to go to zero with the block length there must be a choice of the values of the frozen bits $u_f$ such that

$$\mathbb{E}_{u_f} (\{||T_{X,u_f} - P_{X,Y}||\}) \leq \mathbb{E}_{X,u} (\{||T_{X,u} - P_{X,Y}||\}).$$

Moreover, since convergence in the first mean implies convergence in probability, we conclude that there exists a sequence of PCs for coordination and a choice of the frozen bits that achieves empirical coordination.

**REFERENCES**


Polar Codes for Degraded Relay Channels

Mohammad Karzand
Ecole Polytechnique Fédérale de Lausanne
CH-1015 Lausanne, Switzerland
mohammad.karzand@epfl.ch

Abstract—Polar coding is a recent channel coding technique invented by Arıkan to achieve the ‘symmetric capacity’ of binary-input memoryless channels. Subsequently it was observed that such codes are also good for lossy channel coding, general memoryless channels without feedback and several multi-terminal settings. In this paper, we apply polar codes to achieve the capacity of binary input symmetric degraded relay channels. The main novelty is to show how polar codes can implement block Markov coding in this degraded setting, and consequently yielding a low complexity method to achieve “decode-and-forward” rates.

I. INTRODUCTION

The relay channel was introduced by van der Meulen, and has been studied for forty years in the community. Some of the well-known capacity theorems for different classes of relay channels is proved by El Gamal and Cover in [1]. In this note, we are going to construct a polar coding scheme to achieve the rate given by [1] for binary stochastically-degraded relay channels. Note that this bound coincides with the upper bound when the relay channel is physically degraded. The important class of binary input Gaussian relay channels are included in this class.

Channel polarization technique was introduced in [2] to construct capacity achieving codes for discrete memoryless channels. The encoding and decoding of polar codes with blocklength $n$ take $O(n \log(n))$ arithmetic operations. It is shown in [5] and [3] that the block error probability of these codes decays roughly like $2^{-\sqrt{n}}$ asymptotically in $n$.

Channel polarization technique was considered for orthogonal relay channels in [4] where the receiver observes the relay’s and transmitter’s signals separately via independent channels. By the orthogonality assumption on the receiver side, the multiple access nature of relay channel at its receiver is removed. The nested polar code structure of [4] makes crucial use of this orthogonality and the code is not extensible to the general relay channel [1]. Here, we do not assume the orthogonality of the channels, and thus need a new technique to be able to use polar codes.

We face two important challenges in order to implement a low complexity scheme for the relay channel using polar codes. One is that the outputs of the relay and the transmitter should be correlated with each other, which is not trivial to achieve using polar codes. The second one is to avoid implementing a list decoder with a list size increasing exponentially with blocklength. After a brief review on the relay channel problem and polar coding, in section III, we state the main theorem with the proof, and provide the solutions for these challenges. In the last section we discuss the results.

II. REVIEW

A discrete memoryless relay channel consists of four finite sets $X, X', Y, Y'$ and a collection of conditional probability mass functions $\Pr(y, \tilde{y}|x, \tilde{x})$ on $Y \times \tilde{Y}$. A $(M, n)$ code consists of an encoder that assigns a codeword $x^n(w)$ to each message $w \in \{1, \ldots, M\}$, a relay that assigns at time $i \in \{1, \ldots, n\}$ a symbol $\tilde{x}_i(y_{n-1}^{i-1})$ to each received sequence $y_{n-1}^{i-1}$, and a decoder that assigns a message $\hat{w}$ to each received sequence $y_n^n$. The channel is memoryless in the sense that the current received symbols and the past symbols are conditionally independent given the current transmitted symbols.

The average probability of error is $P_e = \Pr(W \neq \hat{W})$ with message $W$ chosen uniformly at random. The rate $R = \frac{\log M}{n}$ is achievable if there exists a sequence of $([2^{nR}], n)$ codes with vanishing average probability of error. The capacity $C$ is the supremum of all achievable rates.

The relay channel is physically degraded if $X' \supseteq X$. We present an overview of the achievability here, and for each of them generate a low complexity scheme for the relay channel using polar codes.

Proof: We provide an overview of the achievability here, so we will identify the requirements on decoder that can perform well on this channel.

Consider $B$ blocks of transmission, each of $n$ symbols. A sequence of $B-1$ indices $u^{(i)} \in \{1, \ldots, 2^{nR}\}$ will be transmitted over the channels in $nB$ transmissions. We partition the set $\{1, \ldots, 2^{nR}\}$ through a appropriate hash function $s^{(i)} = \text{hash}(u^{(i-1)})$, $s^{(i)} \in \{1, \ldots, 2^{nR_0}\}$.

A doubly indexed set of codewords is generated using random coding as follows. We fix $\Pr(\hat{x}) \Pr(x|\tilde{x})$, and generate $2^{nR_0}$ i.i.d. $n$-sequences in $X^n$, each drawn according to $\Pr(\hat{x})$, and each of them generate $2^{nR}$ i.i.d. $n$-sequences, each drawn independently according to $\Pr(x|\tilde{x})$. 

59
Let $w^{(i)}$ be the new index to be sent in block $i$ and $s^{(i)}$ be defined as the partition corresponding to $w^{(i-1)}$. The transmitter sends $x(w^{(i)}, s^{(i)})$. The relay has an estimate of $w^{(i-1)}$ and it sends $\tilde{x}(s^{(i)})$ in block $i$.

Decoding is accomplished in three steps:
1) Given $s^{(i)}$ and upon receiving $\tilde{y}^{(i)}$, the relay estimates the transmitted message $w^{(i)}$. This will succeed with low probability of error if $R \leq I(X; \tilde{Y}|X)$ and $n$ is sufficiently large.
2) The receiver estimates $s^{(i)}$ using a jointly typical decoder for the pair $(\tilde{x}^{(i)}(s), y^{(i)})$. This can be done with arbitrarily small probability of error if $R_0 < I(X; \tilde{Y})$.
3) Assuming $s^{(i)}$ is decoded correctly at the receiver, the receiver looks for a unique $w^{(i-1)}$ among the $2^n(R-R_0)$ indices that satisfies $h(w^{(i-1)}) = s^{(i-1)}$ and is jointly typical with $y^{(i-1)}$. This can be done if $n$ is sufficiently large and if $R < I(X; Y|X) + R_0$.

In this paper we aim to introduce the code so that all three decoding steps can be done using the polar decoder and thus with low complexity. Before we proceed to present our result, let us give a brief overview of polar codes.

We briefly recall how polarization is done for the $q$-ary discrete channel, with $q$ prime. The basic idea behind the polarization technique is to take two independent uses of the same channel, and by processing their inputs construct two unequal channels with different mutual informations such that the sum of mutual information remains unchanged. The polarization construction is obtained by repeating this operation $l$ times to obtain from $N = 2^l$ independent uses of the channel $W$ the $N$ synthesized channels. The synthesized channels are denoted by $I(W^n), s \in \{+,-\}^l$. Polarization means that as $l$ gets large, almost all the channels at level $l$ are either almost perfect or almost pure noise. As discussed earlier, the transformation preserves mutual information; putting these two facts together, one concludes that the fraction of almost perfect channels is equal to the symmetric channel capacity. In order to construct a polar code $C_{n}(F, u_F)$ of blocklength $n = 2^l$ based on this, one finds $F^c = \{s \in \{+,-\}^l : I(W^n) > 1 - \delta\}$ for some sufficiently small $\delta > 0$ and sends information through these ‘good’ channels. For the other channels indexed by $F = \{+,-\}^l \setminus F^c$, the decoder and encoder, prior to the communication, agree on a sequence $u_F$ that will be sent over these channels; these values are said to be ‘frozen’.

The relation between $U$ and $X$ is as follows:
\[ X = UH_1, \]
where $H_1$ is the $l$-fold Kronecker product of the matrix $[1 \ 0]$. The encoder represents the message as $u_F$, and concatenates it with a randomly selected $u_F$ to form the vector $u$. $u_F$ is revealed to the decoder.

**Theorem 2** ([2],[5]). Given a $q$-ary discrete channel, with $q$ prime, and a fixed rate $R < I(W)$, where $I(W)$ is the symmetric channel capacity of the channel, there exists a sequence of polar codes $C_{n}(F, u_F)$ of rate $R_n > R$ with low encoding and decoding complexity, and with vanishing error probability.

In the next section, this theorem provides the essential tools to perform the first two decoding steps successfully in the achievability proof of the degraded relay channel. However, we still need another result from the polar coding literature to guarantee the success of the third step.

We say $W_0(y_2|x)$ is a stochastically degraded version of $W_0(y_1|x)$ if there exists a distribution $q(y_2|y_1)$ such that $W_0(y_2|x) = \sum W_0(y_1|x)q(y_2|y_1)$.

Consider further, two polar constructions $C_{n}(F_0, u_{F_0})$ and $C_{n}(F_2, u_{F_2})$ for $W_0(y_2|x)$ and $W_0(y_1|x)$. The following lemma is proven in [6]:

**Lemma 1.** [6] If $W_0(y_2|x)$ is a degraded version of $W_0(y_1|x)$, then for a given $\delta$ the frozen set for the good channel $F_0$ is a subset of the frozen set for the bad channel $F_0$.

Thus, all good channels for $W_0$ are also good channels for $W_0$. Comparing the two polar constructions and considering Lemma 1, we label the channels in three groups. $I(W_0)$ fraction of all channels consists of channels which are noiseless for both $W_0$ and $W_0$, $I(W_0^0) = I(W_0)$ fraction of channels consist of those channels which are noiseless for $W_0$ but completely noisy for $W_0$. The remaining $1 - I(W_0)$ fraction of channels are completely noisy for both $W_0$ and $W_0$. Suppose one uses polar codes to multicast common information with the rate $I(W_0)$ over both these channels. Knowing the frozen bits, the decoder for the channel $W_0$ is able to decode everything, but the decoder for the channel $W_0$ cannot decode anything because the transmission rate is above its capacity. At this point a list decoder for the channel $W_0$ is able to pin down the transmitted codeword to a list of codewords of size $q(n,H_0 - \delta)$. Suppose a genie reveals the value of the information bits in the second group of channels to the decoder at $W_0$ in each block, which results in decreasing the rate of the code to $I(W_0)$. Now, the decoder is able to decode the remaining bits using the polar decoder for $W_0$.

We will see in the following that the concept of coding for the degraded relay channel regarding the transmission from relay to receiver consists in sending only the values corresponding to the intersection of the two frozen sets in order to support the decoding at the receiver side of the codeword that is sent by the transmitter over the degraded channel.

### III. Polar Codes for the Degraded Relay Channel

In what follows, we consider $|\mathcal{X}| = |\tilde{X}| = 2$. We consider the marginal probability distribution of $x$ and $\tilde{x}$ to be uniform and $\Pr(x \neq \tilde{x}) = \alpha, 0 \leq \alpha \leq 1$. We define the symmetric rate of the stochastically degraded relay channel to be $R_n = \min\{I(X; \tilde{X}|Y), I(X; Y|X)\}$.

**Theorem 3** (Polar codes achieve $\sup_\alpha R_n$ of degraded relay channels). Given a binary-input stochastically degraded relay channel $\Pr(y, y_1|x, x_1), 0 \leq \alpha \leq 1$ and $R < R_n$, there exists
a sequence of codes based on the polar encoder and decoder with low encoding and decoding complexity of rate \( R_N > R \) with vanishing error probability.

**Proof:** Since \( \alpha \to R_\alpha \) is continuous, it suffices to consider \( \alpha \) of the form \( \alpha = \frac{k}{n} \) where \( k \in \mathbb{N} \) and \( q \) is prime.

As in [1], we consider \( B \) blocks of transmission, each consisting of \( n \) symbols. A sequence of \( B - 1 \) indices will be transmitted over the channel in \( nB \) transmissions. We fix \( R = \frac{\log M}{n} \) and \( R_0 \) as design parameter.

Let \( w^{(i)} \in \{1, \ldots, M\} \) be the new index to be sent in block \( i \), and let \( s^{(i)} \) be an integer in the set \( \{1, \ldots, 2^{nR_0}\} \) and a hash of \( w^{(i-1)} \). We will define this hash function further into the proof. Note that \( W^{(i)} \) and \( S^{(i)} \) are independent. We assume that at the end of block \( i - 1 \) the receiver knows \( (w^{(1)}, w^{(2)}, \ldots, w^{(i-2)}) \) and \( (s^{(1)}, s^{(2)}, \ldots, s^{(i-1)}) \) and the relay knows \( (w^{(1)}, w^{(2)}, \ldots, w^{(i-1)}) \) and \( (s^{(1)}, s^{(2)}, \ldots, s^{(i)}) \).

As the proof of Theorem 1 suggests, the following three tasks should be completed at the end of block \( i \):

- Both the relay and the transmitter have access to \( s^{(i)} \), so they should both cooperatively transmit \( s^{(i)} \) successfully to the receiver.
- The transmitter should transmit \( w^{(i)} \) successfully to the relay.
- The receiver should decode \( w^{(i-1)} \) knowing \( y^{(i-1)} \), \( s^{(i-1)} \) (that has been obtained in the previous block) and \( s^{(i)} \).

An overview of the system can be found in the next column and as we go through our proof we explain each part of it.

We suggest using the following scheme. Suppose \( \tilde{x} = uH_n \) and \( x = j + \tilde{x} \). For now, to satisfy the condition on the probability distribution, we only assume that \( U \) is a uniform \( n \)-tuple vector which is derived from \( S^{(i)} \) and has independent uniformly distributed components, \( j \) is a uniform \( n \)-tuple vector which is a function of \( w^{(i)} \) and has independent and identical components distributed as Bernoulli with parameter \( \alpha \), and \( H_n \) is the \( n \)-fold Kronecker product of the matrix \( \begin{bmatrix} 1 & 0 \end{bmatrix} \).

Observe that the channel between \( \hat{X} \) and \( Y \) is governed by the following conditional probability distribution:

\[
\Pr(y|\tilde{x}) = \sum_j \Pr(y|\tilde{x}, j) \Pr(j)
\]

\[
= \prod_{i=1}^n \sum_{j_i} \Pr(y_i|\tilde{x}_i, j_i + \tilde{x}_i) \Pr(j_i)
\]

\[
= \prod_{i=1}^n \sum_{j_i} \left( \sum_{\tilde{x}_i} \Pr(y_i|\tilde{x}_i, j_i + \tilde{x}_i) \right) \Pr(j_i),
\]

so that the channel is memoryless. Therefore, we consider \( \hat{X} \) to be a codeword of a binary polar code \( C_n(F, U_P) \) designed for the channel

\[
\Pr(y|\tilde{x}) = \sum_j \left( \sum_{\tilde{y}_i} \Pr(y, \tilde{y}|\tilde{x}, j + \tilde{x}) \Pr(j) \right).
\]

Choosing \( U_P \) uniformly at random makes \( \hat{X} \), the codeword of a binary polar code, have a uniform distribution, and hence \( \hat{X} \) satisfies the constraint on its distribution. By Theorem 2, we know that a sequence of codes with rate \( R_{0N} > R_0 \) exists for

\[
R_0 < I(\hat{X}; Y)
\]

with vanishing error probability as the length increases. With this, our scheme succeeds in performing the first task in the achievability scheme.

Similarly, observe that there exists an equivalent channel between \( J \) and \( \hat{Y} \) which is also memoryless. If we follow the steps of the previous part, we will not be able to construct a codeword which satisfies the condition on the probability
distribution of $J$. In this case, we need to proceed as follows.

Recall $\alpha = \frac{k}{q}$. Consider the operator $L(x_q): \{0, \cdots, q - 1\} \rightarrow \{0, 1\}$ as follows

$$L(x_q) = \begin{cases} 
1 & \text{if } x_q < k, \\
0 & \text{if } x_q \geq k.
\end{cases}$$

Observe that if $x_q$ is uniformly distributed on the set $\{0, \cdots, q - 1\}$, then $L(x)$ has a Bernoulli distribution with parameter $\alpha$. This trick is mentioned in [7]. This enables us to provide arbitrary correlation between the output of the relay and the transmitter.

Now, we consider $J = L(X_q)$, where $X_q = \mathbf{V}H_n$ is a codeword of a $q$-ary polar code $C_n(G, V_G)$ designed to encode $w(i)$ for the auxiliary channel governed by the following transition probability:

$$Pr(\tilde{Y}, \tilde{X}|X_q) = Pr(\tilde{Y}|L(X_q), \tilde{X})Pr(\tilde{X}) = Pr(\tilde{X}) \prod_{i=1}^{n} Pr(\tilde{y}_i|x_i, L(x_q|i) + \tilde{x}_i).$$

Observe that $J$ has Bernoulli distribution with parameter $\alpha$. We can transmit the message to the relay by using a polar encoder to encode $w(i)$ to $J$. As the relay has access to $\tilde{Y}$ and $\tilde{X}$ such a power encoder exits for rates up to $I(J; \tilde{Y}|\tilde{X}) = I(J; \tilde{Y}, \tilde{X})$. By Theorem 2, we know a sequence of codes with rate $R_N > R$ exists for

$$R < I(X_q; \tilde{Y}, \tilde{X}) = I(X; \tilde{Y}|\tilde{X})$$

with vanishing error probability as the length increases. With this, our scheme succeeds in performing the second task in the achievable scheme. Note that it is easy to extend the ordinary polar decoder with the same complexity to this case by using $(Y, \tilde{X})$ as a super alphabet as the output.

Now the only remaining task is to decode $w(i-1)$ at the receiver by using $Y(i-1)$, $\tilde{X}(i-1)$ and $s(i)$. Up to now, we have not chosen the hash function relating $s(i)$ and $w(i-1)$. Now, we should choose the relation between $s(i)$ and $w(i-1)$ so that we help the receiver to decode $w(i-1)$. Consider the two channels

$$X_q(i-1) \rightarrow (Y(i-1), \tilde{X}(i-1))$$

and

$$X_q(i-1) \rightarrow (Y'(i-1), \tilde{X}'(i-1))$$

Note that the second of these is a degraded version of the first one. Further, consider the polar code $C_n(G, V_G)$ used to transmit $w(i-1)$ to the relay. According to Lemma 1 and the discussion that follows, we need to reveal at least $n[I(X_q(i-1), \tilde{X}(i-1)) - I(X_q; X_q, \tilde{X}(i-1))]$ bits to the receiver to ensure a successful decoding of $w(i-1)$ at the receiver. We will use $s(i)$ to reveal these values to the receiver. Observe that $s(i)$ should be the value of those positions that are not in the frozen set $\tilde{G}$ but should be considered frozen and revealed to the decoder for the channel $X_q(i-1) \rightarrow (Y(i-1), \tilde{X}(i-1))$ because of a lower symmetric channel capacity. We denote the frozen set of the polar code designed for the channel $X_q(i-1) \rightarrow (Y'(i-1), \tilde{X}'(i-1))$ to be $G$.

The figure below shows the relationship between $W(i)$, $V(i)$, $U(i+1)$ and $S(i+1)$.

![Fig. 4. Construction of hash function relating $W$ and $S$](image)

By Lemma 1, we know that $\tilde{G} \subset G$. We need to transmit $|G| - |\tilde{G}|$ $q$-ary digits to the receiver through $nR_0$ bits (i.e., $s(i)$) to the receiver, so we should have

$$R_0 > I(X; \tilde{Y}|\tilde{X}) - I(X; Y|\tilde{X}).$$

Putting (1), (2) and (3) together, we have the following conditions on the successful transmission of $w(i-1)$ to the receiver at the end of time $i$:

$$R < I(X; \tilde{Y}|\tilde{X})$$

$$R < I(X; \tilde{X}, Y).$$

By theorem (2), this implies that there exist specific values of the frozen bits so that all error probabilities are small for all decoders.

**IV. Conclusion**

The main novelty is to adapt polar codes to implement the block Markov scheme of [1] and to show that in the degraded case the polar decoder will succeed with low complexity.

**ACKNOWLEDGMENT**

The author wishes to thank Prof. Emre Telatar for fruitful discussions and helpful comments on the paper.

**REFERENCES**


Unambiguous Decoding of Generalized Reed–Solomon Codes Beyond Half the Minimum Distance

Alexander Zeh$^{1,2}$, Antonia Wachter$^{1,3}$, and Martin Bossert$^1$

$^1$ Institute of Telecommunications and Applied Information Theory, Ulm University, Ulm, Germany
$^2$ École Polytechnique, Paris, France
$^3$ Inst. de recherche math. de Rennes (IRMAR), Université de Rennes 1, Rennes, France

{alexander.zeh, antonia.wachter, martin.bossert}@uni-ulm.de

Abstract—The Schmidt–Sidorenko–Bossert scheme extends a low-rate Reed–Solomon code to an Interleaved Reed–Solomon code and achieves the decoding radius of Sudan’s original list decoding algorithm while the decoding result remains unambiguous. We adapt this result to the case of Generalized Reed–Solomon codes and calculate the parameters of the corresponding Interleaved Generalized Reed–Solomon code. Furthermore, the failure probability is derived.

Index Terms—Failure Probability, (Interleaved) Generalized Reed–Solomon Codes, Multi–Sequence Shift–Register Synthesis

I. INTRODUCTION

Two approaches to decode an Reed–Solomon (RS) code of length $n$ and dimension $k$ beyond half the minimum distance $(n−k)/2$ with polynomial time and space complexity exist. The first one — formulated by Guruswami and Sudan (GS) [1], [2] — considers the decoding problem of an RS code as interpolation problem and returns a list of all possible candidates within the increased decoding radius. The second one — invented by Schmidt, Sidorenko and Bossert (SSB) [3] — extends an RS code virtually to an Interleaved RS code [4]–[8] and achieves the same decoding radius as Sudan’s original approach, where a multiplicity one [1] for the interpolation is used. The decoding result is unambiguous and therefore it is interesting to investigate its failure probability. For further parallels between these two schemes see [9], [10].

While the GS approach was formulated for Generalized RS (GRS) codes, the SSB adaption for GRS codes is still missing. We fill the lack in this contribution by giving an explicit expression for the virtually generated Interleaved Generalized RS (IGRS) code. Furthermore, we bound the failure probability for this unique decoding approach.

Our contribution is organized as follows. In Section II, Interleaved Generalized Reed–Solomon (IGRS) codes are defined and basic properties are discussed. The joint decoding of IGRS codes is explained in Section III. The parameters and the technique of a virtual extension of an low-rate GRS to an IGRS codes is explained in Section IV. The failure probability for this particular IGRS case is derived in Section V. Section VI concludes our contribution.

II. INTERLEAVED GENERALIZED REED–SOLOMON CODES

Let $L = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ denote a set containing $n$ non-zero distinct elements (code locators) of the finite field $\mathbb{F} = GF(q)$ of cardinality $n < q$. Let $V' = \{v'_1, v'_2, \ldots, v'_n\}$ be a set of non-zero elements of $\mathbb{F}$. Denote

$$f(V', L) = (v'_1 f(\alpha_1), v'_2 f(\alpha_2), \ldots, v'_n f(\alpha_n))$$

for a given polynomial $f(x)$ over $\mathbb{F}$.

An GRS code $GRS(V'; n, k)$ of length $n$ and dimension $k$ over $\mathbb{F}$ with $n < q$ is given by

$$GRS(V'; n, k) = \{c = f(V', L) : f(x) \in F_k[x]\},$$

where $F_k[x]$ denotes the set of all univariate polynomials with degree less than $k$ and indeterminate $x$. The generator matrix $G$ of the GRS code is:

$$G = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
\alpha_1 & \alpha_2 & \cdots & \alpha_n \\
\alpha_1^2 & \alpha_2^2 & \cdots & \alpha_n^2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^{n-k} & \alpha_2^{n-k} & \cdots & \alpha_n^{n-k}
\end{pmatrix}
\begin{pmatrix}
v'_1 \\
v'_2 \\
\vdots \\
v'_n
\end{pmatrix},$$

and the parity check matrix $H$ is given by:

$$H = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
\alpha_1 & \alpha_2 & \cdots & \alpha_n \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^{n-k} & \alpha_2^{n-k} & \cdots & \alpha_n^{n-k}
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_2 \\
\vdots \\
v_n
\end{pmatrix}.$$
\textbf{Definition 1 (Interleaved GRS (IGRS) Code)} Let the set \( \mathcal{K} = \{ k_1, k_2, \ldots, k_s \} \), consist of \( s \) integers, where all \( k_i < n \). Let \( \mathcal{M} = (V_1', V_2', \ldots, V_s') \), be a tuple of \( s \) sets of \( n \) non-zero elements in \( \mathbb{F} \).

An Interleaved Generalized Reed–Solomon code \( \mathcal{IGRS}(n, K, M, s) \) of interleaving order \( s \) is given by

\[
\mathcal{IGRS}(n, K, M, s) = \left\{ \left( c^{(1)}, c^{(2)}, \ldots, c^{(s)} \right) : f^{(i)}(x) \in \mathbb{F}_k[x] \right\},
\]

where each submatrix \( S^{(i)} \) is a \((n - k_i - t) \times t \) matrix and each \( T^{(i)} \) is a column vector of length \( n - k_i - t \):

\[
S^{(i)} = \begin{pmatrix}
S^{(i)}_{n-k_i} & S^{(i)}_{n-k_i-1} & \cdots & S^{(i)}_{n-k_i-t+1}
\end{pmatrix},
\]

\[
T^{(i)} = \begin{pmatrix}
-\tau^{(i)}_1 \\
-\tau^{(i)}_2 \\
\vdots \\
-\tau^{(i)}_{n-k_i-t}
\end{pmatrix}.
\]

To obtain a unique solution of (5), the number of equations has to be greater than or equal to the number of unknowns. By counting both, we obtain the joint error–correcting capability of a heterogeneous \( \mathcal{IGRS}(n, K, M, s) \) code [7]:

\[
\tau_{IRS} = \left\lfloor \frac{s}{s+1} \left( n - 1 - \frac{s}{s+1} \sum_{i=1}^{s} k_i \right) \right\rfloor.
\]

An efficient solution of the system of equations (5) can be obtained by multi-sequence shift-register synthesis as proposed in [3], [12].

\[\text{IV. VIRTUAL EXTENSION TO AN IGRS CODE}\]

We extend the Schmidt–Siderokno–Bossett scheme [3] to the case of Interleaved Generalized Reed–Solomon codes (see Definition 1) and give the corresponding parameters.

A low-rate GRS code can be virtually extended to an IGRS code. This IGRS code is denoted by \( \mathcal{VIRS}(\mathcal{V}', n, k, s) \), where \( n, k \) and \( \mathcal{V}' \) are the original parameters of the \( \mathcal{GRS}(\mathcal{V}', n, k) \) code. The parameter \( s \) denotes the order of (virtual) interleaving. Let the vector \( c = (c_1, c_2, \ldots, c_n) \) and let \( c^{(i)} \) denote the vector, where each coefficient is raised to the power \( i \), i.e., \( c^{(i)} = (c_1^i, c_2^i, \ldots, c_n^i) \). The virtual IGRS code is defined as follows.

\textbf{Definition 2 (Virtual Extension to an IGRS code)} Let \( \mathcal{GRS}(\mathcal{V}', n, k) \) be an \( \mathcal{GRS} \) code with the evaluation polynomials \( f(x) \) defined in (2). The virtually extended Interleaved Generalized Reed–Solomon code \( \mathcal{VIRS}(\mathcal{V}', n, k, s) \) of extension order \( s \) is given by

\[
\mathcal{VIRS}(\mathcal{V}', n, k, s) = \left\{ \left( \begin{array}{c}
c^{(1)}_1 \\
c^{(1)}_2 \\
\vdots \\
c^{(1)}_s \\
c^{(2)}_1 \\
c^{(2)}_2 \\
\vdots \\
c^{(2)}_s \\
\vdots \\
\vdots \\
\vdots \\
c^{(s)}_1 \\
c^{(s)}_2 \\
\vdots \\
c^{(s)}_s \\
\end{array} \right) : f^{(i)}(x) \in \mathbb{F}_{(k-1)+1}[x] \right\},
\]

Through element-wise powering of the original received vector \( r^{(1)} \), we obtain a virtually created error in the received word \( r^{(1)} \) with non-zero values at the same positions as the original one. Therefore, we create a virtual burst error. This fact can be used to increase the error-correcting capability of low-rate
GRS codes (see [3]). We obtain a heterogeneous Interleaved Generalized Reed–Solomon code of order \( s \), where \( s \) is chosen s.t. \( s(k - 1) < n \) and the decoding radius is maximized. The parameters of the IGRS code are:

\[
\begin{align*}
\nu_i &= i(k - 1) + 1, \\
\nu_i^{(i)} &= (\nu_i)^i,
\end{align*}
\]

for \( i = 1, \ldots, s \) and \( j = 1, \ldots, n \). The following theorem shows the relation between the column multipliers \( \nu_i^{(i)} \) and \( \nu_j^{(i)} \) of the virtually created GRS code.

**Theorem 1 (Column Multipliers for VIRS Code)** Let an \( \mathcal{GRS}(\mathcal{V}'; n, k) \) code with evaluation polynomial \( f(x) \) as defined in (2) be given and let \( \nu_j, j = 1, \ldots, n, \) be given, such that (3) holds.

Furthermore, let the virtually extended Interleaved Generalized Reed–Solomon code \( \mathcal{VIRS}(\mathcal{V}; n, k, s) \) for a given parameter \( s \) with \( k_i \) and \( \nu_j^{(i)} \) as in (9). Then, the corresponding column multipliers of the \( i \)th subcode \( \{f^{(i)}(\mathcal{L}, \mathcal{V}_i')\} \) are given by

\[
\nu_j^{(i)} = \frac{\nu_j}{(\nu_j')^{i-1}}.
\]

**Proof:** With (3) we have for the \( i \)th subcode \( \{f^{(i)}(\mathcal{L}, \mathcal{V}_i')\} \):

\[
\sum_{j=1}^{n} \frac{\nu_j^{(i)}}{\nu_j'} \alpha_j^\ell = \sum_{j=1}^{n} \frac{\nu_j}{(\nu_j')^{i-1}} \alpha_j^\ell = \sum_{j=1}^{n} \nu_j \nu_j' \alpha_j^\ell = 0, \quad 0 \leq \ell \leq n - 2.
\]

The decoding radius of a \( \mathcal{VIRS}(\mathcal{V}; n, k, s) \) code is the same as in [3]:

\[
\tau_{\mathcal{VIRS}} = \left\lfloor \frac{s}{s+1} \left( n - \left( \frac{s+1}{2} (k - 1) + 1 \right) \right) \right\rfloor,
\]

where we choose \( s \), such that \( \tau_{\mathcal{VIRS}} \) is maximized. Note that this decoding radius results from (8) using an IGRS code with mean dimension

\[
\bar{k} = \frac{1}{s} \sum_{i=1}^{s} k_i = \frac{1}{s} \sum_{i=1}^{s} (i(k - 1) + 1) = \frac{(k - 1)(s + 1)}{2} + 1.
\]

Decoding can be done as for IGRS codes by solving (5) with multi-sequence shift-register synthesis.

**V. Failure Probability**

Since the virtual extension searches a unique solution for the error locator polynomial, a decoding failure occurs if the system of equations (5) has more than one solution. In the following, we derive an upper bound on the failure probability, which is the same as for the virtual extension of the RS codes used in [3] and therefore independent of the column multipliers.

We bound the probability that the \( s(n - \bar{k} - t) \times t \) matrix \( S \) does not have full rank \( t \) and denote the failure probability, if \( t \) errors occurred, by:

\[
P_f(t) \leq P((\text{rank}(S) < t) \mid (|\mathcal{E}| = t)).
\]

The following theorem states the failure probability.

**Theorem 2 (Upper Bound on the Failure Probability)** Let a codeword of an \( \mathcal{GRS}(\mathcal{V}; n, k) \) code be corrupted by an error of weight \( t \leq \tau_{\mathcal{VIRS}} \) (11). Let the GRS code be virtually extended to an IGRS code \( \mathcal{VIRS}(\mathcal{V}; n, k, s) \) of extension order \( s \) with the dimensions \( k_i \) and the set of column multipliers \( \nu_j^{(i)} \), for \( i = 1, \ldots, s \) as in (9).

Assume, for decoding we solve the system of equations from (5). The probability for a decoding failure is upper bounded by:

\[
P_f(t) \leq \left( \frac{q}{q - 1} + \frac{1}{q} \right)^t \cdot \frac{q^{3(\tau_{\mathcal{VIRS}} - t)}}{q - 1}.
\]

**Proof:** As in the proof of [3, Theorem 3], this is equivalent to the case that there exists a vector \( u \neq 0 \) of length \( t \), such that

\[
\exists u \neq 0 : S^{(i)} \cdot u^T = 0, \forall i = 1, \ldots, s.
\]

Each syndrome matrix can be decomposed into five matrices (in [3], the decomposition consists only of four matrices):

\[
S^{(i)} = H^{(i)} \cdot \nu^{(i)} \cdot F^{(i)} \cdot D \cdot V,
\]

where \( D \) and \( V \) are the same full-rank \( t \times t \) matrices as in [3, Proof of Theorem 3] and

\[
H^{(i)} = \left( \alpha^{(\mu-1)} j_\nu \right)_{\mu=1, \ldots, n-k_i-t, \nu=1, \ldots, t},
\]

\[
\nu^{(i)} = \text{diag} \left( \nu^{(i)}_{j_1}, \nu^{(i)}_{j_2}, \ldots, \nu^{(i)}_{j_t} \right),
\]

\[
F^{(i)} = \text{diag} \left( e^{(i)}_{j_1}, e^{(i)}_{j_2}, \ldots, e^{(i)}_{j_t} \right),
\]

where \( \mathcal{E} = \{j_1, j_2, \ldots, j_t\} \) denotes the union of the error positions and \( e^{(i)}_{j_\nu} \) denotes the error value at the \( j_\nu \)th position of the \( i \)th received word. Since \( \nu^{(i)} \) and \( F^{(i)} \) are both diagonal matrices, we can switch their order. Since \( \nu^{(i)} \), \( D \), \( V \) are nonsingular, there is a one-to-one mapping from \( u \) to \( v \), where

\[
v^T = \nu^{(i)} \cdot D \cdot V \cdot u^T = (v_1, v_2, \ldots, v_t)^T.
\]

Hence, (13) is equivalent to

\[
\exists v \neq 0 : H^{(i)} \cdot F^{(i)} \cdot v^T = 0, \forall i = 1, \ldots, s.
\]

This is the same as [3, Proof of Theorem 3, Equation (22)] and using Lemma 1 in the appendix for arbitrary \( q \), the rest of the proof is analog.

Thus, the upper bound on the failure probability is independent of the column multipliers and in particular independent of using GRS codes or the RS codes from [3].

Note that this upper bound generalizes [3, Theorem 3] from RS to GRS codes and also to arbitrary \( q \) (in [3, Theorem 3] it was only shown for \( q = 2 \)). Moreover, the bound does not only hold for \( s = 2 \) (as indicated in [3]), but also upper bounds the
cases with $s > 2$, although the bound is not very tight then. This is due to the fact that we can upper bound the probability that all sub-matrices are singular by the probability that some sub-matrices are singular.

VI. CONCLUSION

We considered unambiguous decoding of low-rate GRS codes beyond half the minimum distance. For this purpose, the received word of a usual GRS codeword was virtually extended to an interleaved GRS code, similar to the Schmidt–Sidorenko–Bossert scheme. We derived the parameters of the extended code and proved that the failure probability is the same as for the RS codes from [3]. Moreover, we generalized the known bound on the failure probability to arbitrary fields.

VII. ACKNOWLEDGMENT

We would like to thank Markus Ulmschneider for implementing the decoding algorithm.

APPENDIX

In the following, we extend [3, Lemma 4] to arbitrary fields.

Lemma 1 Let $(v, e, c)$ be three non-zero elements from a field $\mathbb{F} = \mathbb{GF}(q)$. Then, the set

$$V = \left\{ \frac{v \cdot e}{v \cdot e(2)} : v, e \in \mathbb{F} \setminus \{0\} \right\}$$

with $e^{(2)} = 2c \cdot e + e^2$ forms the set of all full weight vectors of length 2, i.e.

$$V = \{ v \in \mathbb{F}^2 : \text{wt}(v) = 2 \} = \left[ \mathbb{F} \setminus \{0\} \right]^2.$$

Proof: It is sufficient to show that all $(q - 1)^2$ possible vectors $v \in V$ are pairwise different. For any fixed $e$, consider two vectors $v, \tilde{v} \in V$, and assume that $v = \tilde{v}$, then

$$v \cdot e = \tilde{v} \cdot e, \quad v \cdot e^{(2)} = \tilde{v} \cdot e^{(2)}.$$

Dividing (17) by (16) yields

$$\frac{2c \cdot e + e^2}{e} = \frac{e^{(2)}}{e} = \frac{\tilde{e}^{(2)}}{\tilde{e}} = \frac{2c \cdot \tilde{e} + \tilde{e}^2}{\tilde{e}}.$$

Therefore, $2c + e = 2c + \tilde{e}$ and hence, $e = \tilde{e}$. Inserting this into (16), we obtain $v = \tilde{v}$. Thus, for any $c$, two different pairs $(v, e) \neq (\tilde{v}, \tilde{e})$ always result in two different vectors $v, \tilde{v}$.

REFERENCES


Iteration-constrained design of IRA codes

Alex Grant, Ingmar Land and Guangsong Wang
Institute for Telecommunications Research
University of South Australia
{alex.grant,ingmar.land}@unisa.edu.au, wangy012@mymail.unisa.edu.au

Abstract—This paper addresses extrinsic information chart based design of non-systematic irregular repeat-accumulate codes for a given number of decoding iterations. This criterion is of practical importance in many applications where complexity or latency is limited. Our main contribution is a novel formulation of the optimization problem in a particular way that makes its actual evaluation possible. This is achieved by introducing additional optimization variables and constraints which have the effect of “unrolling” the iteration. Our approach enforces the finite iteration condition, while avoiding the need to know arbitrary repeated functional compositions of the component EXIT functions. We restrict attention to the binary erasure channel, for which the EXIT chart approach is exact. Extension to other sparse-graph codes and other communication channels is straightforward.

I. INTRODUCTION

Sparse-graph codes with iterative message-passing decoding represent the state-of-the-art of modern channel coding. The two major classes are low-density parity-check (LDPC) codes, and irregular repeat-accumulate (IRA) codes. For both classes, encoding and decoding complexity are linear in the code length, and there are efficient code design methods [1–3].

The typical design goal in the literature is capacity-approaching codes. These codes that have maximum rate while the probability of error diminishes with increasing code length and increasing number of iterations. The predominant approaches are density evolution and extrinsic information transfer (EXIT) charts [1, 4, 5]. While density evolution provides exact results for the decoding threshold, the EXIT chart method employs a Gaussian approximation. The advantage of the latter, however, is that code design can often be formulated in terms of a convex optimization problem, allowing convenient numerical evaluation. Tradeoffs between the two methods for the design of IRA codes were discussed in [6]. For erasure channels, the two methods are equivalent.

Two important parameters characterizing implementation complexity are the parity-check matrix density and the number of decoding iterations. The relationship between parity-check density and performance in terms of the gap to capacity have been analyzed in [7, 8], the required number of iterations for successful decoding and corresponding bounds have been addressed in [9]. As opposed to LDPC codes, non-systematic IRA codes can achieve capacity on the binary erasure channel with bounded complexity per information bit [7].

Whereas the above approaches assume infinite iteration, the number of iterations required to achieve a target performance is of practical interest. In [10] a method for the design of irregular LDPC codes was proposed, which minimizes the number of iterations for a given target error probability. The key concept is to approximate the number of iterations by a continuous function, and use this as the objective function in the optimization. Starting from a capacity-achieving code, the variable-node degree distribution is iteratively adapted to improve this objective function.

We take a different approach, seeking a code that is optimized for a fixed number of iterations. We restrict ourselves to non-systematic IRA codes over the binary erasure channel (BEC). Extension to other spares-graph codes and to other communication channels (given the usual assumptions and approximations with EXIT charts) is possible.

II. IRA CODES AND EXIT CHARTS

We assume familiarity with low-density parity-check codes, their graphical representation, iterative decoding, and EXIT chart based design [1–3, 5]. We focus on non-systematic irregular repeat accumulate codes, where information bits are repeated, interleaved, passed through parity-checks, and finally scrambled by an accumulator [6, 11]. Only the code bits produced by the accumulator are transmitted over the channel. We restrict our analysis to the binary erasure channel, for which the EXIT chart approach is exact [12].

The encoder maps $k$ information bits onto $n$ code bits, the code rate is $R = k/n$. The code is defined by the edge perspective degree polynomials, $\lambda(z) = \sum_{i=1}^{\infty} \lambda_i z^{-i}$ for the variable-nodes and $\rho(z) = \sum_{i=1}^{\infty} \rho_i z^{-i}$ for the check-nodes, where $\lambda_i, \rho_i, i = 1, 2, \ldots$ denotes proportion of edges connected to variable nodes of degree $i$, and $\rho_i, i = 1, 2, \ldots$ denotes the proportion of edges connected to check nodes of degree $i$.

According to common practice in the literature, $\rho_i$ does not count the connection to the accumulator. The design rate is

$$R_d = \frac{\sum_i \lambda_i / i}{\sum_i \rho_i / i} = \int_0^1 \frac{\lambda(z)dz}{\int_0^1 \rho(z)dz} \leq R,$$ \hspace{1cm} (1)

The code bits are transmitted over a binary erasure channel with erasure probability $\delta$ and capacity $C = 1 - \delta$.

The iterative decoder consists of three components: the variable-node decoder, the check-node decoder, and the accumulator decoder. Only the accumulator obtains observations from the channel. In every iteration the component decoders are activated in the order variable-check-accumulator-check.

In the EXIT chart method, the mutual information transfer between the component decoders is modeled and used to analyze convergence behavior [5]. For the following analysis,
we combine the check-node decoder and the accumulator decoder into a single unit, henceforth referred to as the check-accumulator decoder. Since we have assumed a binary erasure channel, all virtual channels are also erasure channels and EXIT chart analysis is exact [12].

As shown in Figure 1, we orient the extrinsic information transfer chart with the $x$-axis representing the variable-node extrinsic information $I_{Ev}$ and the check-accumulator a-priori information $I_{Av}$ of the variable node decoder after iteration $\ell$. The $y$-axis carries the variable-node a-priori information $I_{Av}$ and check-accumulator extrinsic information, $I_{Eca}$. In order to avoid proliferation of subscripts, where it will not cause confusion, we simply use $x$ and $y$ to represent these quantities. The variable-node EXIT function $f(y)$ maps from $y = I_{Av}$ onto $x = I_{Ev}$. Conversely, the check-accumulator function $g(x)$ maps from $x = I_{Eca}$ onto $y = I_{Eva}$.

![EXIT chart](image)

Fig. 1. Example EXIT chart of a non-systematic IRA code: variable-node EXIT function $f(y)$ and check-accumulator EXIT function $g(x)$.

The variable node EXIT function is given by [12]

$$f(y) = \sum_i \lambda_i f_i(y)$$  \hspace{1cm} (2)

$$f_i(y) = 1 - (1 - y)^{i-1}, \quad i = 1, 2, \ldots$$  \hspace{1cm} (3)

where $f_i$ is the transfer function for a degree-$i$ variable node. The check-accumulator EXIT function is [12]

$$g(x) = \sum_i \rho_i g_i(x),$$  \hspace{1cm} (4)

$$g_i(x) = x^{i-1} \left( \frac{C}{1 - (1 - C)x^i} \right)^2, \quad i = 1, 2, \ldots$$  \hspace{1cm} (5)

where $g_i$ is the transfer function for a degree-$i$ check node combined with the accumulator. We can show that $f(y)$ and $g(x)$ are linear in $\lambda$ and $\rho$ and convex in $y$ and $x$, respectively.

Denote $x^{(\ell)} = I_{Ev}^{(\ell)}$ the extrinsic information at the output of the variable node decoder after iteration $\ell$, and let $y^{(\ell)} = I_{Av}^{(\ell)}$ be the corresponding a-priori information. With reference to Figure 1, the corresponding values of the trajectory for $L$ iterations are given recursively for $\ell = 1, 2, \ldots$ by

$$x^{(\ell)} = f(y^{(\ell-1)}), \quad y^{(\ell)} = g(x^{(\ell)}).$$  \hspace{1cm} (6)

The initial values are $x^{(0)} = 0$ and $y^{(0)} = 0$, resulting from the fact that the variable nodes are not directly connected to the communication channel.

We can eliminate the $y^{(\ell)}$ from (6) to obtain $x^{(\ell)} = f(g(x^{(\ell-1)}))$. Going a step further, we may also substitute these equalities into each other to express the variable node extrinsic information after $L$ iterations $I_{Ev}^{(L)}$ as

$$x^{(L)} = h_L(0)$$  \hspace{1cm} (7)

where $h(x) = f(g(x))$ and $h_L$ denotes $L$-fold functional composition.

The areas under the variable node curve (with the orientation indicated in Figure 1) and under the check-accumulator curve are given by

$$1 - \int_0^1 f(y) \, dy = \sum_i \lambda_i = \int_0^1 \lambda(z) \, dz,$$  \hspace{1cm} (8)

$$\int_0^1 g(x) \, dx = \sum_j \rho_j = \int_0^1 \rho(z) \, dz,$$  \hspace{1cm} (9)

respectively. Substituting from (1) we have the following theorem, first proved for low-density parity-check codes in [12].

**Theorem 1 (Area Theorem):**

$$C - R_c = \frac{A}{\int_0^1 \rho(z) \, dz},$$  \hspace{1cm} (10)

$$A = \int_0^1 g(x) \, dx - \left(1 - \int_0^1 f(y) \, dy \right).$$  \hspace{1cm} (11)

where $A$ is the area between the check-accumulator curve $g$ and the variable node curve $f$.

Thus, finding the code with the smallest rate gap from capacity corresponds to minimizing the area between $f$ and $g$. Note a minor difference from the original LDPC result, where $\int \rho$ sets a “scaling factor”. We do not require the inner check-accumulator code to have rate 1.

For code analysis and design, it is useful to express the (information) bit error rate of the decoder as a function of the mutual information exchanged during the iterative process.

**Theorem 2 (Bit Error Rate):** The bit error rate $P_b$ of a non-systematic IRA code transmitted over a BEC, with variable node degree distribution $\lambda$ and a-priori information $I_{Av}$ (to the variable nodes) is given by

$$P_b = \frac{1 - I_v}{2}$$  \hspace{1cm} (12)

$$I_v = \left( \sum_j \frac{\lambda_j}{j} \right) \sum_i \frac{1}{i} \left(1 - (1 - I_{Av})^i\right).$$  \hspace{1cm} (13)

The value $I_v$ denotes the overall variable node information.
Note that $P_b$ is monotonically increasing in $I_{N_i}$. For fixed design rate $R_i$,

$$I_{N_i} = \left( R_i \sum_j \frac{\rho_j}{\lambda_j} \right)^{-1} \sum_i \frac{\lambda_i}{\lambda_j} \left( 1 - (1 - I_{N_i})^i \right)$$  \hspace{1cm} (14)

which is linear in the $\lambda_i$, convex in the $\rho_j$ and concave in $I_{N_i}$.

With sufficiently many iterations, the decoder will converge to the fixed point $x = f(g(x)) = h(x)$, which for a well-designed code will occur at a value of $x$ very close to 1, corresponding to a low bit error rate. To prevent an undesirable error floor, we may wish to enforce a condition that the iteration does not get stuck at a fixed point prior to some error floor, we may wish to enforce a condition that the corresponding to a low bit error rate. To prevent an undesirable error floor, we may wish to enforce a condition that the iteration does not get stuck at a fixed point prior to some error floor, we may wish to enforce a condition that the corresponding to a low bit error rate. To prevent an undesirable error floor, we may wish to enforce a condition that the iteration does not get stuck at a fixed point prior to some error floor, we may wish to enforce a condition that the corresponding to a low bit error rate. To prevent an undesirable error floor, we may wish to enforce a condition that the iteration does not get stuck at a fixed point prior to some error floor, we may wish to enforce a condition that the corresponding to a low bit error rate.

According to Theorem 1, this corresponds to a curve fitting distribution which results in the minimal rate loss from capacity.

Note that $I_{N_i}$ is a multinomial of degree $v_i$ with $\lambda_i$, $\rho_i$ and $x_i$. Otherwise it is equivalent to

$$\sum_j j \rho_j < 1 + \lambda_2 \frac{2 - C}{C}. \hspace{1cm} (16)$$

The prevailing approach to code design finds the degree distribution which results in the minimal rate loss from capacity. According to Theorem 1, this corresponds to a curve fitting distribution. Equivalently, for fixed $\rho_i$, the maximum design rate is achieved by maximization of $\lambda_i$ of $\sum_i \lambda_i i$. This results in the following linear program.

Problem 1 (Capacity Approaching Code): Given the channel capacity $C$ and check-node degree distribution $\rho_i$, find the variable-node degree distribution $\lambda^*$ as the solution to the following linear program:

$$\max_{\lambda} \sum_i \frac{\lambda_i}{i} \hspace{1cm} (17)$$

s.t. $\sum_i \lambda_i f_i(y) > g^{-1}(y), \quad y \in [0, 1]$  \hspace{1cm} (18)

$$(1 + \lambda_2) \frac{C}{2 - C} - \lambda_2 \sum_j j \rho_j > 0 \hspace{1cm} (19)$$

$\lambda_i \geq 0, i = 1, 2, \ldots \hspace{1cm} (20)$$

$$\sum_i \lambda_i = 1. \hspace{1cm} (21)$$

The constraints (18) ensure that there are no sub-optimal fixed points and (19) is the stability condition.

In many situations of practical interest, the implementation complexity of the decoder is limited, and it is therefore of interest to consider the design of codes that reduce decoder complexity, even if this is at the expense of reduced code rate.

Complexity constraints may be introduced into the optimization problem (17) to a certain extent via (a) fixing the maximal variable node degree (which needs to be done anyhow in order to have a finite number of variables in Problem 1), or (b) introducing other linear constraints on the $\lambda_i$, e.g. corresponding to the cost of high degree nodes.

For a code optimized according to (17), operating at rates close to capacity, it may take many iterations to achieve the target bit error rate. It is therefore of great practical interest to consider the design of codes that achieve the best possible performance after a fixed number of iterations, denoted by $L$. One obvious approach would be to replace the objective function in (17) with the bit error rate at iteration $L$. From (12) and (13), we see that this is equivalent to maximizing the variable node information at iteration $L$,

$$I_{N_i}^{(L)} = \left( \sum_i \frac{\lambda_i}{\lambda_j} \right)^{-1} \sum_i \frac{\lambda_i}{\lambda_j} \left( 1 - g(h^{L-1}(0))^i \right) \hspace{1cm} (22)$$

$$h^{L-1}(x) = \sum_{y_1} f(g_{y_1} g_{y_2} \ldots g_{y_L}(x)) \hspace{1cm} (23)$$

Note that $I_{N_i}^{(L)}$ is a multinomial of degree $L$ times the square of the maximal variable node degree. This could potentially present numerical difficulties for optimization. Furthermore, this formulation requires us to compute $h^{L}$ for each value of $L$ of interest. In the case of IRA codes and the BEC, we have analytical expressions for $f_i(y)$ and $g_i(x)$. In more general settings, we may only have numerical samples of these functions.

Motivated by (6), the key novel step in this paper is to introduce dummy variables $x_i, y_i, \ell = 0, 1, \ldots, L$. With reference to Figure 1, these are the values of the variable node extrinsic and a-priori information at iteration $\ell$. For given target rate $R$, we formulate our design problem as follows (where we have used (14)).

Problem 2 (Minimum bit error rate after $L$ iterations): Given $R < C$ determine the variable node degree distribution $\lambda^*$ and check node degree distribution $\rho^*$ that minimize the bit error rate after $L$ iterations as the solution to:

$$\max_{\lambda, \rho, x, y} \left( R_i \sum_i \rho_i \right)^{-1} \sum_i \frac{\lambda_i}{\lambda_j} \left( 1 - (1 - y^{L-1})^i \right) \hspace{1cm} (22)$$

s.t. $\sum_i \rho_i = 1. \hspace{1cm} (24)$

$$x_0 = 0, y_0 = g(x_0) \hspace{1cm} (25)$$

$$x_l = \sum_i \lambda_i f_i(y_{l-1}), \quad l = 1, 2, \ldots, L \hspace{1cm} (26)$$

$$y_l = \sum_j \rho_j g_j(x_l), \quad l = 1, 2, \ldots, L \hspace{1cm} (27)$$

In this setup, the target rate $R$ would be chosen to be less than capacity, sacrificing some code rate for potential performance improvements after $L$ iterations.
The new constraints (25), (26), (27) ensure consistency of the decoding trajectory according to (6). The trajectory constraints are convex, however as discussed earlier, the objective function is neither convex nor concave.

We have empirically observed that there is very little bit error rate penalty incurred by replacing the bit error rate objective function (22) with $y_L$ or $x_L$ (i.e. maximize the information passed between the decoders). In fact, this may be of interest if the code is an inner component of some other outer iteration, e.g. a multi-user decoder. This results in the following optimization problem:

**Problem 3 (Maximal information after L iterations):**

Given $R < C$ determine the variable node degree distribution $\lambda^{\ast L}$ and check node degree distribution $\rho^{\ast L}$ that maximizes the variable node extrinsic information after $L$ iterations as the solution to the following convex optimization problem:

$$\max_{\lambda, \rho, x, y} x_L$$

s.t. (1), (19), (20), (21), (23), (24), (25), (26), (27)

Note that by “unrolling” the iteration, we obtain constraints that are independent of the maximum number of iterations $L$. Furthermore, we do not need to know the $L$-fold functional composition $f(g(f(\ldots)))$. Instead, we only need to know $f(y)$ and $g(x)$ individually. This could be an additional advantage if these characteristics are only available numerically (e.g. for other kinds of codes).

**IV. Numerical Example**

We seek for a code of rate $R/5 = 0.5$ for transmission over a BEC with erasure rate $\delta = 0.3$, that achieves a low bit error probability with $L = 10$ decoding iterations. We consider variable node degrees from the set \{2, 3, 4, 5, 6, 13, 20, 30\}, and (for simplicity) the fixed check-node distribution $\rho_1 = 0.2$, $\rho_3 = 0.8$. (Note that degree-1 check nodes are required for non-systematic IRA codes.) Solving Problem 3 for maximal information, we obtain the variable node distribution $\lambda^{(1)}$, see Table I, and the EXIT function $f^{(1)}$, depicted in Fig. 2.

As an alternative approach, we designed a capacity-approaching code, according to Problem 1, for a BEC with a higher erasure rate, namely $\delta = 0.47$, to obtain a code rate of $R = 0.5$. We then use this code with $L = 10$ decoding iterations for the given BEC with $\delta = 0.3$. The resulting variable node distribution $\lambda^{(2)}$ is given in Table I, and the EXIT function is depicted in Fig. 2.

For the maximal-information code, the tunnel between the EXIT functions is clearly more open towards the end, and for a given number of iterations a better performance can be achieved. Resulting from that, the bit error rate after $L = 10$

**Table I**

<table>
<thead>
<tr>
<th>$t$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>13</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda^{(1)}$</td>
<td>0</td>
<td>0.11</td>
<td>0.66</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda^{(2)}$</td>
<td>0.16</td>
<td>0.15</td>
<td>0.0</td>
<td>0.54</td>
<td>0.14</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 2. EXIT chart and decoding trajectories: check-accumulator EXIT function $g$; variable-node EXIT functions for the maximal-information code (Problem 3), $f^{(1)}$, and the 'capacity-achieving' code (Problem 1), $f^{(2)}$.

iterations is $10^{-11}$ for the maximal-information code and only $10^{-3}$ for the ‘capacity-achieving’ code.

**ACKNOWLEDGMENT**

This work has been supported in parts by the Australian Research Council under the ARC Discovery Grant DP0986089.

**REFERENCES**


The Relation Between Block Length and Reliability for a Cascade of AWGN Links

Ramanan Subramanian
Institute for Telecommunications Research, University of South Australia, Mawson Lakes, SA 5095, Australia
Email: Ramanan.Subramanian@unisa.edu.au

Abstract—This paper presents a fundamental information-theoretic scaling law for a heterogeneous cascade of links corrupted by Gaussian noise. For any set of fixed-rate encoding/decoding strategies applied at the links, it is claimed that a scaling of $\Theta(\log n)$ in block length is both sufficient and necessary for reliable communication. A proof of this claim is provided using tools from the convergence theory for inhomogeneous Markov chains.

I. INTRODUCTION

Consider a line network in which a message originating at a certain node is conveyed to the intended destination through a series of hops. The reception at every hop is impaired by zero-mean additive white Gaussian noise (AWGN) of a certain variance. The originating node as well as the intermediate nodes are subject to an average transmit-power constraint $P_0$ for the duration of transmission of any message. The messages, drawn from a certain set of size $M$, are encoded and transmitted as codewords in $\mathbb{R}^N$ that satisfy the transmit power constraint. Every intermediate hop decodes (possibly erroneously) the original message from its noisy observation and re-encodes it before transmitting to the next hop. One can form the (row-stochastic) probability transition matrix $P_i$ for the $i$th hop in which each row gives the conditional probability distribution of the message decoded at hop $i$, given a certain message sent by the previous hop:

$$P_i = \begin{pmatrix}
    p_i(1|1) & p_i(1|2) & \cdots & p_i(1|M) \\
p_i(2|1) & p_i(2|2) & \cdots & p_i(2|M) \\
    \vdots & \vdots & \ddots & \vdots \\
p_i(M|1) & p_i(M|2) & \cdots & p_i(M|M)
\end{pmatrix}.$$ 

Assume that a network of the type outlined above consists of $n$ hops. If the encoder-decoder pairs are identical at the hops, and if the noise variance at every hop is $\sigma^2$, then the matrices $P_i$ will be identical to, say $P$ for all $i$. Then, the rows of $P^n$ will give the conditional probability distributions of the message $W_n$ decoded at the final destination, given a certain message $W$ at the originating node. For decisions made on signals corrupted by Gaussian noise, $P$ has the property that any column is either all-zero or contains only non-zero entries. For such a matrix $P$, the rows of $P^n$ tend to be identical for large $n$ (see Theorem 4.9 in [1]), if $M$, $N$, $P_0$, and $\sigma$ are independent of $n$. In other words, the probability distribution of $W_n$ will be almost independent of the original message $W$. This implies that the mutual information $I(W; W_n)$ between the original message and the decoded message tends to zero with $n$. Hence, communication in the network is highly unreliable if $N$ and $M$ do not vary with the number of hops $n$.

On the other hand, if $N$ is large and if one chooses $M = 2^{NR}$ for any $R < \frac{1}{2} \log (1 + \frac{P_0}{\sigma^2})$, random-coding theorem states that there exists a channel code of block length $N$ such that the probability of error during decoding at any link is bounded above by $e^{-N E(R; P/\sigma^2)}$, where $E$ is the random coding exponent for the corresponding AWGN channel [2], [3]. By employing such a code for each link in the line network model, choosing any $f(n) \in \Theta(1)$, and letting $N$ vary with $n$ as $\frac{f(n) \log(1 + f(n))}{\log(1 + f(n))}$, one obtains the following union bound on the probability of communication failure in the network:

$$P \left[ W_n \neq W \right] \leq n P_0 \leq n e^{-N E(R; P/\sigma^2)} = e^{-f(n)} \rightarrow 0.$$ 

Hence, it is possible to transmit messages with very high reliability in the network for a fixed code rate $R$, if the code length $N$ satisfies $N(n) \in \Theta(\log n)^{\dagger}$. On the other hand, the previous discussion about diminishing $I(W; W_n)$ implies that any fixed-rate fixed-length coding scheme is not scalable for the line network. This leads to the natural question whether the scaling $\Theta(\log n)$ for $N$ is indeed optimal. In this paper, we show that for any scaling of $N$ satisfying $N(n) \in o(\log n)^{\dagger}$, the mutual information $I(W; W_n)$ diminishes to zero with $n$ for any set of fixed-rate codes. This implies that the scaling order $\Theta(\log n)$ for $N$ is necessary for reliable communication in the network. Hence, the sufficiency criterion for $N$ given by the union bound in Eqn. (1) is indeed optimal up to a constant factor.

While this scaling problem has been solved partially in the past for cascades of homogeneous DMCs by Niesen et al. in [4], it has also been pointed out in the same work that the tools developed there apply neither to non-homogeneous cascades nor to cascades of continuous-input channels (as is the case in the current paper). In contrast, the results developed here are applicable to continuous-input AWGN links where the links can be non-identical.

\textsuperscript{1}For any $f(n) > 0$ and $g(n) > 0$, $f(n) \in \omega(g(n))$ iff $\lim_{n \to \infty} \frac{f(n)}{g(n)} = \infty$, $f(n) \in o(g(n))$ iff $\lim_{n \to \infty} \frac{f(n)}{g(n)} = 0$, $f(n) \in \Theta(g(n))$ iff $\exists c > 0$ s.t. $\lim_{n \to \infty} \frac{f(n)}{g(n)} = c$.

This work has been supported by the Australian Research Council under the ARC Discovery Grant DP0986089.
II. NOTATIONS AND DEFINITIONS

The set of all real numbers is denoted by $\mathbb{R}$ and the set of all natural numbers by $\mathbb{N}$. Natural logarithms are assumed in the notations, unless the base is specified. The notation $\| \cdot \|$ represents $\mathbb{L}^2$ norm throughout.

We employ the following terminologies in this paper. Let $N \in \mathbb{N}$, called the code length or block length of the transmission scheme. A code rate $R > 0$ is a real number such that $2^{NR}$ is an integer. Let $\mathcal{A} \triangleq \{ 1, 2, 3, \ldots, 2^{NR} \}$, called the message alphabet. Let $P_0 > 0$ be the input power constraint for transmission.

Definition 1: A rate-$R$ length-$N$ code $\mathcal{C}$ with a power constraint of $P_0$ is an ordering of $M = 2^{NR}$ elements from $\mathbb{R}^N$, called code words that satisfies

$$\mathcal{C} = \{ x_1, x_2, x_3, \ldots, x_M \} \text{s.t.} \forall w \in \mathcal{A}, \frac{1}{N} \| x_w \|^2 \leq P_0. \quad (2)$$

Definition 2: A rate-$R$ length-$N$ decision rule $\mathcal{R}$ is an ordering of $M = 2^{NR}$ sub-sets of $\mathbb{R}^N$, called decision regions, spanning $\mathbb{R}^N$ and pairwise disjoint:

$$\mathcal{R} = (R_1, R_2, \ldots, R_M) \text{s.t.} \bigcup_{w \in \mathcal{A}} R_w = \mathbb{R}^N, \text{ and } \bigcap_{w \neq w'} R_{w_1} \cap R_{w_2} = \emptyset. \quad (3)$$

Definition 3: The encoding function $\text{ENC}_{\mathcal{C}} : \mathcal{A} \rightarrow \mathbb{R}^N$ for a code $\mathcal{C}$ is defined by:

$$\text{ENC}_{\mathcal{C}}(w) = x_w, \quad (4)$$

where $x_w$ is the $w$-th code word in $\mathcal{C}$.

Definition 4: The decoding function $\text{DEC}_{\mathcal{R}} : \mathbb{R}^N \rightarrow \mathcal{A}$ for a decision rule $\mathcal{R}$ is defined by:

$$\text{DEC}_{\mathcal{R}}(y) = \sum_{w \in \mathcal{A}} w \mathbb{1}_{R_w}(y), \quad (5)$$

where $\mathbb{1}_{R_w}$ is the indicator function and $R_w$ is the $w$-th subset in $\mathcal{R}$.

III. NETWORK MODEL

The line network model to be considered is described in Fig. 1. There are $n + 1$ nodes in the network identified by the indices $\{ 0, 1, 2, \ldots, n \}$. The $n$ hops in the network are each associated with noise variances $\sigma_i^2, 1 \leq i \leq n$. For encoding, nodes 0, 1, $\ldots$, $n - 1$ choose $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_{n-1}$ respectively to transmit, each code being rate-$R$ length-$N$ with a power constraint of $P_0$. For decoding, nodes 1, 2, $\ldots$, $n$ choose rate-$R$ length-$N$ decision rules $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n$ for reception. From here on, we write $\text{ENC}_{\mathcal{C}_i}$ and $\text{DEC}_{\mathcal{R}_i}$ in place of $\text{ENC}_{\mathcal{C}_i}$ and $\text{DEC}_{\mathcal{R}_i}$ respectively, for simplicity. Node 0 generates a random message $W \in \mathcal{A}$ with probability distribution $p_W(w)$ and intends to convey the same to node $n$ through the cascade of noisy links in a multihop fashion. Each of the $n - 1$ intermediate nodes estimates the message (as sent by the node in the previous hop) from its own noisy observation, re-encodes the decoded message, and transmits the resulting codeword to the next hop. The codeword transmitted by node $i$, for any $0 \leq i \leq n - 1$ is given by

$$X_i = \text{ENC}_{\mathcal{C}_i}(\hat{W}_i), \quad (6)$$

where $\hat{W}_i$ is the estimate of the message at node $i$ after decoding (Note that $\hat{W}_0 = W$ in this notation). The observation received by node $i$, for any $1 \leq i \leq n$ is given by $Y_i$, which follows a conditional density function that depends on the codeword $X_{i-1}$ sent by the previous node:

$$p_{Y_i|X_{i-1}}(y|x) = \frac{1}{(2\pi\sigma_i^2)^{\frac{N}{2}}} e^{-\frac{\| x - y \|^2}{2\sigma_i^2}}. \quad (7)$$

The above density function follows from the assumptions of AWGN noise and memoryless-ness in the channel. The message $\hat{W}_i$ decoded by node $i$ is given by

$$\hat{W}_i = \text{DEC}_{\mathcal{R}_i}(Y_i). \quad (8)$$

Note that the random variable $\hat{W}_n$ represents the message decoded by the final destination, as per the above notation.

IV. AN UPPER BOUND ON THE RELIABILITY OF THE NETWORK

The key result presented here is summarized by the following theorem and its corollary:

Theorem 1: In a line network consisting of a cascade of $n$ AWGN links, let for $m \leq n$ of the links the noise variances satisfy $\sigma_m^2 \geq \sigma_0^2 > 0$. Then, for any choice of codes $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_{n-1}$ and decision rules $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_m$ with rate $R$ length $N$, and for any $\epsilon > 0$, the following holds for sufficiently large $m$:

$$I(W; \hat{W}_n) \leq 2^{NR} \left( 1 - \frac{(\frac{N}{2})^{\frac{m}{2}}}{\Gamma(\frac{N}{2} + 1)} e^{-N\epsilon(P_0/\sigma_0^2)} \right)^{m(1-\epsilon)}, \quad (9)$$

where

$$E'(S) = \exp\left\{ \cosh^{-1}\left( \frac{S}{\frac{n}{2} + 1} \right) \right\} + \cosh^{-1}\left( \frac{S}{\frac{n}{2} + 1} \right).$$

The corollary to the above theorem leads to the upper bound on the reliability vs. block-length scaling as claimed in the introductory section:

Corollary 1: Let $R > 0$, $\sigma_0 > 0$, and $1 \geq r > 0$ be independent of $n$. In a line network of $n$ links, let for $rn$ of the hops the noise variance be at least $\sigma_0^2 > 0$. Then, for any $N(n) \in o(\log n)$, $\lim_{n \rightarrow \infty} I(W; \hat{W}_n) = 0$.

Proof of Theorem 1: At any hop $i$, the process of channel encoding, noisy reception, followed by decoding induces the conditional probabilities $p_{W_i|W_{i-1}}(k|j), \forall j, k \in \mathcal{A}$. For every hop $i$, let $P_i$ be the $M \times M$ row-stochastic matrix whose entry in row $j$ and column $k$ is $p_{W_i|W_{i-1}}(k|j)$. Note that the $j$th row in $P_i$ gives the conditional probability mass function on the estimate of message $W$ at hop $i$ (i.e., $\hat{W}_i$) given that the message sent by hop $i - 1$ was $j$. Let

$$P = \prod_{i=1}^{n} P_i. \quad (9)$$

Then, $P$ clearly represents the row-stochastic probability transition matrix between the original message $W$ and the message decoded at the final destination, $\hat{W}_n$. The transition matrix $P$
along with $p_W$, the probability mass function of the original message $W$ together induce a joint distribution between $W$ and $W_n$. Our goal is to find an upper bound on $\mathcal{I}(W;\hat{W}_n)$, with the constraints as given in the statement of Theorem 1.

For any $M \times M$ row-stochastic matrix $Q = \{Q_{jk}\}$, let us consider the following measures of deviation:

$$
\delta(Q) \triangleq \max_{j,k} \max_{i} |Q_{jk} - Q_{ij}|,
$$

$$
\lambda(Q) \triangleq 1 - \min_{j,k} \sum_{i=1}^{M} \min(Q_{jk} Q_{ij}).
$$

Both the above deviations measure the degree to which the rows of $Q$ differ. The measure $\lambda(Q)$ is known as Dobrushin’s coefficient of ergodicity and plays a role in the convergence of non-homogeneous stochastic processes [5]. Moreover, for any stochastic matrix $Q$,

$$
0 \leq \lambda(Q) \leq 1.
$$

Now consider the total variational distance between the joint distribution of $W$ and $\hat{W}_n$ and the product of the respective marginal distributions:

$$
\tau(W;\hat{W}_n) \triangleq \sum_{w_1, \hat{w}_n \in \mathcal{W}} |p_W(w_1)p_{\hat{W}_n}(\hat{w}_n) - p_W(w_1)\gamma_{\hat{W}_n}(\hat{w}_n)|.
$$

In the above expression, substituting $p_W(w_1)p_{\hat{W}_n}(\hat{w}_n)$ for $p_W(w_1, \hat{w}_n)$, $\sum_{\hat{w}_n \in \mathcal{W}} p_{\hat{W}_n}(\hat{w}_n)\gamma_{\hat{W}_n}(\hat{w}_n)$ for $p_W(w_1, \hat{w}_n)$ and by using the triangle inequality, one obtains the following bound:

$$
\tau(W;\hat{W}_n) \leq \sum_{w_1, \hat{w}_n \in \mathcal{W}} p_W(w_1) p_{\hat{W}_n}(\hat{w}_n) \delta(P) = M\delta(P) = 2^{N\lambda(P)}.
$$

Recalling the definition of $P$ in our case, and from Theorem 2 in [6] (equivalently, Lemma 2 in [7]), we have:

$$
\delta(P) = \delta\left(\prod_{i=1}^{n} P_i\right) \leq \prod_{i=1}^{n} \lambda(P_i)
$$

$$
\leq \prod_{i=1}^{n} \lambda(P_i).
$$

The second inequality in the above follows by applying Eqn. (12) for those hops for which $\sigma_i < \sigma_0$. Note that as per the statement of the theorem, we have at least $m$ terms in the product in the last expression above. Now consider $\lambda(P_i)$ for any hop $i$ such that $\sigma_i \geq \sigma_0$. Suppose $P_i$ has $L$ distinct columns with indices $k_1, k_2, \ldots, k_L$ such that

$$
\forall j, p_{j,k_1} = p_{W_j|W_{i-1}}(k_1|j) \geq p_1
$$

$$
\vdots
$$

$$
\forall j, p_{j,k_L} = p_{W_j|W_{i-1}}(k_L|j) \geq p_L.
$$

Then from the definition in Eqn. (11), it follows that

$$
1 - \lambda(P_i) \geq \sum_{l=1}^{L} p_l.
$$

We now show that the conditions given by Eqn. (15)-(16) hold true in our case, so that we can apply Eqn. (17) to our bound. Let $\alpha > 0$. Consider one particular hop $i$ satisfying $\sigma_i \geq \sigma_0$. Define the spherical regions

$$
S_1 = \left\{ y \in \mathbb{R}^N : \|y\| \leq \sqrt{N\alpha}\right\}
$$

$$
S_\alpha = \left\{ y \in \mathbb{R}^N : \|y\| \leq \alpha\sqrt{N\alpha}\right\}.
$$

Clearly, any code word in $\mathcal{C}_i$ (transmitted by the previous hop) has to be in $S_1$. Next, observe that $\mathcal{F}_i$ has to contain decision regions, say $L$ in number, spanning $S_\alpha$ entirely. Let us designate them as $\mathcal{R}_{k_1} \times \mathcal{R}_{k_2} \times \cdots \times \mathcal{R}_{k_L}$. Also define for each $1 \leq \ell \leq L$, $V_\ell \triangleq \mathcal{R}_{k_\ell} \cap S_\alpha$. Hence, we have:

$$
\bigcup_{\ell=1}^{L} V_\ell = S_\alpha, \text{ and } V_{\ell_1} \cap V_{\ell_2} = \emptyset.
$$

Let $\hat{W}_{i-1} = j$ be the message as decoded by the previous hop. Hence, the code word $x_i$ from $\mathcal{C}_i$ was transmitted by the node $i - 1$. The probability that this was decoded as message $k_\ell$ at hop $i$ is given by:

$$
p_{\hat{W}_i|\hat{W}_{i-1}}(k_\ell|j) = \int_{y \in \mathcal{R}_{k_\ell}} p_{\hat{W}_i|X_i}(y|x_i) dy \geq \int_{y \in V_\ell} p_{\hat{W}_i|X_i}(y|x_i) dy
$$

$$
= (2\pi\sigma_i^2)^{-\frac{N}{2}} \int_{y \in V_\ell} e^{-\frac{\|y-x_i\|^2}{2\sigma_i^2}} dy \geq \int_{y \in V_\ell} e^{-\frac{(\|y-x_i\|^2)}{2\sigma_i^2}} dy
$$

$$
= e^{-\frac{(\|x_i\|^2)}{2\sigma_i^2} |V_\ell|}.
$$
where $|V_i|$ denotes the $N$-dimensional volume of $V_i$. In the above series of inequalities, step (a) follows from Eqn. (7), (b) from the triangle inequality, and (c) from the fact that any $y \in V_i$ must satisfy $\|y\| \leq \alpha \sqrt{NP_0}$, since $V_i \subseteq S_i$, and from the fact that $x_j$, being a code word in $G_i$, naturally satisfies the power constraint criterion given by $\|x_j\| \leq \sqrt{NP_0}$. Note that the lower bound given by Eqn. (19) is independent of the message $j$ decoded by the previous hop. Hence, the conditions given by Eqns. (15)-(16) are satisfied for

$$p_t = \frac{1}{(2\pi \sigma^2)^\frac{N}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} |V_i|. \quad (20)$$

Hence, the bound given by Eqn. (17) holds true. In our case, it gives:

$$1 - \lambda(P_t) \geq 1 - \sum_{l=1}^L p_t = \frac{1}{(2\pi \sigma^2)^\frac{N}{2}} \sum_{l=1}^L |V_i| = \frac{1}{(2\pi \sigma^2)^\frac{N}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left( \frac{\pi}{\Gamma(\frac{N}{2}+1)} \right) (\alpha^2 NP_0)^\frac{N}{2} \quad (21)$$

The last equality follows from the fact that $V_i$ are mutually disjoint and span the entire $N$-dimensional sphere $S_i$ (Eqn. (18)).

The volume of $S_i$ is computed as:

$$|S_i| = \frac{\pi^\frac{N}{2}}{\Gamma(\frac{N}{2}+1)} (\alpha \sqrt{NP_0})^N. \quad (22)$$

The ideas in the above argument are outlined in Figs 2(a) and 2(b).

Since at least $m$ hops satisfy the minimum noise variance criterion $\sigma_i \geq \sigma_0$, we can now combine Eqn. (24) with Eqn. (14) to obtain

$$\delta(P) \leq \left( 1 - \frac{\left( \frac{\Gamma}{\Gamma(\frac{N}{2}+1)} \right)^N (NP_0 \sigma_0^2)^m}{1 - N^E(P_0/\sigma_0^2) e^{2N^E(P_0/\sigma_0^2)}} \right). \quad (25)$$

Applying the above to Eqn. (13), we will have

$$\tau(W; \hat{W}_n) \leq 2^{NR} \left( 1 - \frac{\left( \frac{\Gamma}{\Gamma(\frac{N}{2}+1)} \right)^N (NP_0 \sigma_0^2)^m}{1 - N^E(P_0/\sigma_0^2) e^{2N^E(P_0/\sigma_0^2)}} \right). \quad (26)$$

We can finally bound $I(W; \hat{W}_n)$ using Lemma 2.7 of Chapter 1 in [8] by:

$$I(W; \hat{W}_n) \leq \tau(W; \hat{W}_n) \log \frac{2^{NR}}{\tau(W; \hat{W}_n)}. \quad (27)$$

The bound given by the statement of the theorem then follows from combining the above with Eqn. (26) and by observing that for sufficiently large $m$, $\tau(W; \hat{W}_n) \leq e^{-1}$, that $-x \log x$ is non-decreasing for $x \leq e^{-1}$, and finally from the fact that for any $\epsilon > 0$, $x^2 \geq \log x$ for sufficiently large $x$.

**Proof of Corollary 1**: For the corollary, we have $m = nr$. Moreover, $\Gamma(\frac{N}{2} + 1) \leq (\frac{\Gamma}{\Gamma(\frac{N}{2}+1)})(\frac{N}{2})^N$. Hence, the bound given by Theorem 1 reduces to:

$$I(W; \hat{W}_n) \leq 2^{NR} \left( 1 - e^{-N^E(P_0/\sigma_0^2)} \right)^{nr(1-\epsilon)} \quad (28)$$

$$= 2^{NR} \epsilon e^{-nr(1-\epsilon)} e^{-N^E(P_0/\sigma_0^2)} \quad (29)$$

$$\leq e^{NR \log 2} e^{-nr(1-\epsilon) - N^E(P_0/\sigma_0^2)} \quad (30)$$

The last expression diminishes to 0 with $n \to \infty$ if $N = N(n) = o(\log n)$. Hence, we will have $I(W; \hat{W}_n) \to 0$ for any $N = o(\log n)$.

**REFERENCES**


Secure Multiplex Coding Over Interference Channel with Confidential Messages

Xiaolin Li
Dept. of Electrical and Computer Engineering
Hong Kong University of Science and Technology
darwinlx1@ust.hk

Ryutaroh Matsumoto
Dept. of Communications and Integrated Systems
Tokyo Institute of Technology, 152–8550 Japan, and Dept. of Mathematical Sciences, Aalborg University, Denmark.
ryutaroh@it.ss.titech.ac.jp

Abstract—In this paper, an inner bound on the capacity region of two-user interference channels with two confidential messages has been proposed. With secure multiplex coding, we show that the Han-Kobayashi region can be achieved with strong security requirement. The equivocation rate for a collection of secret messages has also been evaluated, when the length of the message is finite or the information rate is high, our result provides a good approximation for bounding the worst case equivocation rate.

Index Terms—Information theoretic security, capacity region, interference channel, secure multiplex coding, strong security.

I. INTRODUCTION

In [12] the authors proposed the celebrated Han-Kobayashi region which provided the best inner bound on the channel capacity of the interference channel known to now. In the coding scheme proposed in that paper, the information from each transmitter was divided into two parts, the first part was for only one receiver (we say this part of the information is sent over the “private channel”), and the other part could be decoded by both receivers (we say this part of the information is sent over the “common channel”). A natural question raises: what is the secrecy transmission rate if only confidential messages are sent? In [2], the authors proposed an inner bound of secure transmission over interference channel by modifying the coding scheme in [11], and information is only sent over the “private channel”, which is a natural solution. But in this paper, we show somewhat surprisingly that even if we transmit over the “common channel”, confidentiality can also be guaranteed, thus we propose a larger achievable security rate region.

Also in [2], outer and inner bounds have been provided, but are under the weak secrecy requirement [3], which requires that the mutual information divided by the length of the codeword goes to zero as the codeword length goes to infinity. But this requirement is not strong enough for some applications [8] [9], because even if this rate goes to zero asymptotically, vital information bits can still be easily leaked to an illegitimate receiver. Moreover, secrecy is achieved by adding dummy random bits into the transmitted signal, which inevitably decreases the information rate.

The authors in [2] did not evaluate the equivocation rate when the information rates of the secret messages are large or the length of the message is finite. This means that their results are only valid for the cases where secrecy can be asymptotically achieved, but if the secrecy requirement is not achieved, they are not able to evaluate how much information may be leaked out.

In [4], the authors proposed the secure multiplex coding scheme for wiretap channels, the goal of which is to remove the rate loss incurred by the random dummy message. In [5] and [7], the authors applied the secure multiplex coding in different scenarios: broadcast channels with a common message and secure network coding. They showed that secure multiplex coding can not only remove the information rate loss, but can also achieve strong security within the capacity region.

In this paper, the model of interference channel with confidential messages as in [2] is considered. By applying the technique of secure multiplex coding, we propose an inner bound that can achieve the Han-Kobayashi region with the strong security requirement. Moreover, we give the dominating term approximation for a lower bound on the equivocation rate with finite message length.

This paper is organized as the following: in Section II, the system model and preliminaries shall be introduced. In Section III, we present an inner bound on the secure capacity of the interference channel. Section IV concludes the paper.

II. SYSTEM MODEL AND PRELIMINARIES

A. System Model

We adopt the same channel model as in [2]. Consider a discrete memoryless interference channel with finite input alphabets $X_1, X_2$, finite output alphabets $Y_1, Y_2$, and the channel transition probability distribution $P_{Y_1,Y_2|X_1,X_2}$. Two transmitters wish to send independent, confidential messages to their respective receivers.

The definition of the capacity region for the interference channel with secure multiplex coding is given as follow:

Definition 1: The rate tuple $(R_1, \ldots, R_1, R_2, \ldots, R_2, R_2)$ and the equivocation rate tuple $\{R_1, R_1, R_2, R_2\} \not\in \{1, \ldots, T_1\}$, $\not\in \{1, \ldots, T_2\}$ are said to be achievable for the secure multiplex coding with $T_1$ secret messages for sender 1 and $T_2$ messages for sender 2, if there exists a sequence of stochastic encoders for sender 1 denoted as $C_1^n : C_1^n \times \cdots \times C_1^n \rightarrow X_1^n$, and for sender 2 denoted as $C_2^n : C_2^n \times \cdots \times C_2^n \rightarrow X_2^n$, and for receiver 1 deterministic
receivers are denoted as $I_j$, and for receiver 2 determinist decoder $C_2$ collections of random variables: instead of making the mutual information between tuples.

The main idea behind the multiplex coding is that throughout the whole paper we use natural log.

Let $R_{HK}(P^*_t)$ be the set of nonnegative rate-tuples $(R_1, R_2)$ that satisfy

$$R_1 \leq I(V_1; Y_1 | W_1 U)$$
$$R_1 \leq I(V_1; Y_1 | W_1 U) + I(V_2 W_1; Y_2 | W_1 U)$$

Note in the above we abuse the notation a little by writing $R_{HK}(P^*_t)$, we can write this because if $P(u, w_1, w_2, v_1, v_2, x, y_1, y_2) \in P^*_2$, then the marginal distribution $P(u, w_1, w_2, v_1, v_2, x, y_1, y_2) \in P^*_1$.

An inner bound for the interference channels with secure multiplex coding is
\[ R_{in} = \bigcup_{P^*_2} F_P R_{in}(P^*_2) \]

**Proof:**

**Part 1: Random Coding Scheme**

Before presenting the random coding scheme, some notations are introduced: let \( (c^1_{t,1}, \ldots, c^T_{t,1}) \in \mathbb{C}^{T \times 1} \) be the secret messages for transmitter \( t \), and denote \( C^t = \prod_{T=1}^{T} C^T_{t,1} \) where \( C^T_{t,1} \) is the alphabet of randomness used by the stochastic encoder, and \( n \) denotes the code length. Let \( F^n_t \) be the set of all linear bijective maps from \( C^n_t \) to itself.

We modify the random coding scheme proposed in [11], and apply the secure multiplex coding techniques. The new scheme is described in detail as follows:

Fix the distribution of \( P(u), P(w_t, v_t | u) \) and \( P(x_t | v_t) \), also since the channel distribution \( P(y_1, y_2 | x_1, x_2) \) is given, all the distributions in (20) are now fixed.

1. **Codebook Generation:** Sender \( t \) and receiver \( t \) fix and agree on the choice of a bijective function \( f^n_t \in F^n_t \). Given \( T_t \) secret messages \( (c^1_{t,1}, \ldots, c^T_{t,1}) \in \mathbb{C}^{T \times 1} \), uniformly choose \( c^T_{t,T+1} \in C^n_{t,T+1} \) let \( \alpha = (f^n_t)^{-1}(c^1_{t,1}, \ldots, c^T_{t,T+1}) \). Here the message \( c^T_{t,T+1} \) is used by the stochastic encoder to increase the randomness in the secret message. In order to use multiplex coding, we write

\[ C^n_t = ([L^t_1]^{-1} * [C^n_{t,1}]_t, \ldots, [C^n_{t,T+1}]_t)^T \]

Note that if the length of \( C^n_{t,1} \) is \( k_{1,t} \), then \( L_t \) is a nonsingular matrix of size \( l_t \times l_t \) with \( l_t = \sum_{T=1}^{T} k_{T,t} \). Randomly generate a sequence \( \mathbf{u} \) with probability \( P(\mathbf{u}) = P_1 P(\mathbf{u}_1) \cdots P(\mathbf{u}_T) \), and assume that both transmitters and receivers know the time-sharing sequence \( \mathbf{u} \).

For transmitter \( t \), generate \( 2^{nR_t} \) independent sequences \( S_t \) is the information rate over the “common channel”) \( w_t \) each with probability \( P(\mathbf{w}_t | \mathbf{u}) = P_1 P(\mathbf{w}_t_1 | \mathbf{u}_1) \). Then generate \( 2^{nR_t} \) \((R_t = S_t + T_t, and T_t \) is the rate of information over “private channel”) independent sequences \( w_t \) each with probability \( P(\mathbf{w}_t | \mathbf{u}) = P_1 P(\mathbf{w}_t_1 | \mathbf{u}_1) \).

2. **Encoding:** Encode \( c_i \) with encoder \( t1 \) and \( t2 \) (here \( t = 1 \) or 2, refer to Fig. 1), according to the codebook generated in the previous step and obtain the codeword \( w^n_t \). Then the transmitters generate the channel input sequences based on respective mappings \( P_{X_1|V_1} \) and \( P_{X_2|V_2} \).

The structure of the encoder is illustrated in Fig. 1. The codebook for encoder \( i,j \) is denoted in Fig. 1 (i, j = 1, 2), for the simplicity of notations, let \( \Lambda = (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}) \), and \( \Lambda = (\Lambda_{11}, \Lambda_{12}, \Lambda_{21}, \Lambda_{22}) \).

3. **Decoding:** Without loss of generality, we consider for receiver 1. Let \( A^{(n)}(W_1, V_1, W_2, Y_{1}, U) \) denote the set of jointly typical sequences defined in [13]. Since receiver 1 is supposed to know the time sharing sequence \( U^n_t \), after receiving \( Y^n_t \), the receiver will try to find \( W^n_1, V^n_1, W^n_2 \) such that \( (W^n_1, V^n_1, W^n_2, Y^n_1, U^n) \in A^{(n)}(W_1, V_1, W_2, Y_1, U) \). When such choice exists and is unique then the decoding is successful; otherwise declare error.

**Part 2: Evaluation of Equivocation Rate**

If we compare our scheme with the original one in [11], it can be observed that the decoding error probability in our scheme is at least as good as the original one. Hence we only need to show the existence of a tuple \((u, \lambda, f^n_1, f^n_2)\) such that the strong security requirement can be fulfilled. Without loss of generality, we just need to consider the information leaked to receiver 2 from sender 1. In [6], it has been proved that if \( F^n_1 \) is an uniform random variable on \( F^n_1 \) and \( \alpha_{1,2} \) is the projection from \( C^n_1 \) to \( \prod_{T \in I} C^n_{T,1} \), then \( \alpha_{1,2} \circ F^n_1 \) is a family of two-universal hashing functions [10]. The projection is simply \( \alpha_{1,2}(C^n_1) = \{C^n_{1} \mid i \in I \} \). With a little abuse of notations, we also write \( C^n_1 = (F^n_1)^{-1}(C^n_{1,1}, \ldots, C^n_{T,1,T+1}) \), by the uniformity of the distribution it can be seen \( C^n_1 \) and \( F^n_1 \) are statistically independent.

In the following, we evaluate the mutual information \( (I(\alpha_{1,2}(F^n_1); Y^n_2 | F^n_1, \Lambda, U^n)) \) and \( \rho \) is a real constant with \( 0 < \rho < 1 \). For the limitation of the pages, we omit the details which can be found in [14]. With some derivations,

\[ E^{t_{1,2}}_{\lambda},\lambda,\rho \exp \left( \rho I(C_1;Y_2^n | F^n_1 = f^n_1, \Lambda = \lambda, U^n = u) \right) \leq 1 + \exp \left( n \rho (R_{X_2} - R_{F_1} + I(V_1; Y_2^n | U_2) + \epsilon(\rho)) \right), \]

where \( \epsilon(\rho) \rightarrow 0 \), and \( R_{F_1} = \sum_{t \in I} R_{F_1} \).

From (27), if \( R_{X_2} - R_{F_1} + I(V_1; Y_2^n | U_2) < 0 \), we can always choose large enough \( n \), such that

\[ I(C_1;Y_2^n | F^n_1, \Lambda, U^n) \leq \epsilon_n \]  

Hence we can see that the strong security is achieved.

In case strong secrecy can not be achieved or \( n \) is finite, by almost the same argument in [7], there exists at least one tuple \((f^n_1, \lambda, u)\) such that

\[ I(C_1;Y_2^n | F^n_1, \Lambda, U^n) < 2^{T+2} \epsilon_n \] (if \( R_{X_2} = R_{x_{1,2}}(\rho) \)),

\[ \exp(\rho I(C_1;Y_2^n | F^n_1 = f^n_1, \Lambda = \lambda, U^n = u)) \leq 2^{T+2} \left[ 1 + \exp \left( n \rho (R_{X_2} - R_{F_1} + I(V_1; Y_2^n | U_2) + \epsilon(\rho)) \right) \right]. \]
Decoding error probability $\leq 2^{T+2} \delta_0$.

In the above expressions, $\epsilon(\rho)$ is a constant that only depends on $\rho$, and $\lim_{\rho \to 0} \epsilon(\rho) = 0$.

By Eq. (29) we can see
\[
I(C_1; Z^n|F^n_1 = f^n_1, \Lambda = \lambda, U^n = u)) \leq 1 + \log(2^{T+2}) \frac{n}{n\rho} + R_2 - R_p + I(V_1; Y_2|U_2) + \epsilon(\rho).
\]

For $R_2 - R_p + I(V; Z|U) + \epsilon(\rho) \geq 0$. By Eq. (30) we can see that the equivocation rate $H(C_1; Y^n_2, F^n_1 = f^n_1, \Lambda = \lambda)/n$ becomes larger than the required value for sufficiently large $n$. This completes the analysis of the equivocation rates and the mutual information for all $\emptyset \neq I \subseteq \{1, \ldots, T_1\}$.

\textbf{Remark 5:} This inner bound of secret capacity over interference channel given above shows that the whole Han-Kobayashi region can be achieved. Moreover from (27) and (28), it can be observed that strong security can be achieved. Note that (28) – (30) also provides an upper bound for the leakage information, which is not analyzed in [2].

\section{V. Conclusions}

In this paper an inner bound that achieves Han-Kobayashi region can be achieved. Moreover from the University Grants Committee of the Hong Kong Special Administration Region, China, and also by the MEXT Grant-in-Aid for Young Scientists (B) No. 22760267, Grant-in-Aid for Scientific Research (A) No. 23246071, and the Vellum Foundation through their VELUX Visiting Professor Programme 2011–2012.

\section{Acknowledgement}

This work was partly supported by the AoE Grant E-02/08 from the University Grants Committee of the Hong Kong Special Administration Region, China, and also by the MEXT Grant-in-Aid for Young Scientists (B) No. 22760267, Grant-in-Aid for Scientific Research (A) No. 23246071, and the Vellum Foundation through their VELUX Visiting Professor Programme 2011–2012.

\textbf{REFERENCES}

Polymer Expansions for Cycle LDPC Codes

Nicolas Macris and Marc Vuffray
LTHC, IC, EPFL,
CH-1015 Lausanne, Switzerland
nicolas.macris@epfl.ch, marc.vuffray@epfl.ch

Abstract—We prove that the Bethe expression for the conditional input-output entropy of cycle LDPC codes on binary symmetric channels above the MAP threshold is exact in the large block length limit. The analysis relies on methods from statistical physics. The finite size corrections to the Bethe expression are expressed through a polymer expansion which is controlled thanks to expander and counting arguments.

I. INTRODUCTION

A few years ago Cherkov and Chernyak [1] devised a loop series which represents the partition function of a general vertex model as the product of the Bethe mean field expression and a residual partition function over a system of loops. In this representation all quantities are entirely expressible in terms of Belief Propagation (BP) marginals or messages. However it has not been clear so far if this representation leads to a controlled series expansions for the log-partition, in other words the free energy. If this is the case it should hopefully allow to control the difference between the true free energy and the Bethe free energy.

The loop expansion has a potential interest in coding theory since Low-Density-Parity-Check (LDPC) and Low-Density-Generator-Matrix (LDGM) codes on general binary-input memoryless symmetric (BMS) channels fit in the framework of (generalized) vertex models. In this context free energy is just another name for conditional input-output Shannon entropy. For these models it is believed that the Bethe formula for the conditional entropy/free energy is exact. However there is no general proof, except for the cases of the binary erasure channel [2], LDGM codes for high noise, and in special situations for LDPC codes at low noise [3].

We consider cycle LDPC codes for high noise (above the MAP threshold) on the binary symmetric channel (BSC). We show that, under the assumption that there exists a fixed point for the BP equations, the average conditional entropy/free energy is given by the Bethe expression. The novelty of the approach is to turn the loop expansion into a rigorous tool allowing to derive provably convergent polymer expansions [4]. Controlling the loop expansion is a non-trivial task because in most situations of interest the number of loops proliferates. For example, this is the case (for the system of fundamental cycles) in capacity approaching codes even under MAP decoding [5].

II. LOOP AND POLYMER REPRESENTATIONS

Let $\Gamma = (V,E)$ be a graph with vertices $a \in V$ of regular degree $d$ and edges $ab \in E$. The symbol $\partial a$ denotes the set of $d$ neighbors of $a$. In vertex models the degrees of freedom are spins $\sigma_{ab} \in \{-1,+1\}$ attached to each edge. At each function node $a \in V$ we attach a non-negative function $f_a(\sigma_{ba})$ depending only on neighboring variables $\sigma_{ba} \equiv (\sigma_{ab})_{b \in \partial a}$. We study probability distributions which can be factorized as $\mu_V(\sigma) = \frac{1}{Z_V} \prod_{a \in V} f_a(\sigma_{\partial a})$, $Z_V = \sum_{\sigma} \prod_{a \in V} f_a(\sigma_{\partial a})$, and their associated free energy $f_a \equiv \frac{1}{n} \ln Z_V$.

For each edge $ab \in E$ we introduce two directed “messages” $\eta_{a \rightarrow b}$ and $\eta_{b \rightarrow a}$. For the moment these variables are arbitrary and are collectively denoted by $\vec{\eta}$. One has the identity [1]

$$f_a = \frac{1}{n} \ln Z_{\text{Bethe}}(\vec{\eta}) + \frac{1}{n} \ln Z_{\text{corr}}(\vec{\eta}).$$

The first term is the Bethe free energy functional,

$$\ln Z_{\text{Bethe}}(\vec{\eta}) = \sum_{a \in V} \ln \left( \sum_{\sigma} f_a(\sigma_{\partial a}) \prod_{b \in \partial a} e^{\eta_{b \rightarrow a} \sigma_{ab}} \right) - \sum_{ab \in E} \ln (2 \cosh(\eta_{a \rightarrow b} + \eta_{b \rightarrow a})).$$

The “partition function” in the second term can be expressed as a sum over all subgraphs (not necessarily connected) $g \subseteq \Gamma$

$$Z_{\text{corr}}(\vec{\eta}) = \sum_{g \subseteq \Gamma} K(g)$$

and $K(g) = \prod_{a \in g} K_a$ with

$$K_a = \sum_{\sigma} p_a(\sigma_{\partial a}) \prod_{b \in \partial a} e^{\eta_{b \rightarrow a} \sigma_{ab}} \prod_{c \in \partial a \setminus b} e^{\eta_{c \rightarrow a} \sigma_{ab}}$$

$$p_a(\sigma_{\partial a}) = \frac{f_a(\sigma_{\partial a}) \prod_{b \in \partial a} e^{\eta_{b \rightarrow a} \sigma_{ab}}}{\sum_{\sigma} f_a(\sigma_{\partial a}) \prod_{b \in \partial a} e^{\eta_{b \rightarrow a} \sigma_{ab}}}.$$ It is well known that the stationary points of (3) satisfy the BP fixed point equations,

$$\eta_{a \rightarrow b} = \frac{\sum_{\delta_{ab}} \sigma_{ab} f_a(\sigma_{\partial a}) \prod_{c \in \partial a \setminus b} e^{\eta_{c \rightarrow a} \sigma_{ab}}}{\sum_{\delta_{ab}} f_a(\sigma_{\partial a}) \prod_{c \in \partial a \setminus b} e^{\eta_{c \rightarrow a} \sigma_{ab}}}.$$ (5)

The novelty here is to prove that the Bethe expression is exact in the large block length limit. This is achieved by expressing the partition function as a sum over all subgraphs (not necessarily connected) $g \subseteq \Gamma$ and by proving that the average conditional entropy/free energy is given by the Bethe expression. The novelty of the approach is to turn the loop expansion into a rigorous tool allowing to derive provably convergent polymer expansions [4]. Controlling the loop expansion is a non-trivial task because in most situations of interest the number of loops proliferates. For example, this is the case (for the system of fundamental cycles) in capacity approaching codes even under MAP decoding [5].

Remarkably, for any solution of (5), $K(g) = 0$ if $g$ contains a degree one node. Thus if $\vec{\eta}$ is a fixed point of the BP equations then $Z_{\text{corr}}(\vec{\eta})$ is given by the sum in (4) over $g \subseteq \Gamma$ with no degree one nodes. Such graphs are called loops (see figure 1).

One can recognize that $Z_{\text{corr}}$ can be interpreted as the partition function of a system of polymers. A loop $g \subseteq \Gamma$ can be decomposed into its connected parts in a unique way as
each polymer $\gamma$ we define a weight (also called activity), $K(\gamma) = \prod_{a \in \gamma} K_a$. Let $g = \bigcup_{i=1}^M \gamma_i$. Since the $\gamma_i$ are disjoint, $\prod_{a \in \gamma} K_a = \prod_{i=1}^M K(\gamma_i)$. Thus equation (4) can be cast in the form
\[
Z_{\text{corr}}(\vec{\eta}) = \frac{1}{M!} \sum_{\gamma_1,..,\gamma_M} K(\gamma) \prod_{i<j} \delta(\gamma_i \cap \gamma_j = \emptyset).
\]
(6)

In this sum each $\gamma_i$ runs over all connected subgraphs with no dangling edges of the underlying graph $\Gamma$. The sum over the number of polymers $M$ has a finite number of terms because the polymers cannot intersect.

In the next paragraphs $\Gamma$ is a random $d$-regular graph. We denote by $P$ and $E$ the relevant probability and expectation over this ensemble.

III. POLYMER EXPANSION

We wish to compute the correction to the Bethe free energy in (2), namely $f_{\text{corr}}(\vec{\eta}) = \frac{1}{n} \ln Z_{\text{corr}}(\vec{\eta})$ when $\vec{\eta}$ is a BP fixed point. Using (6) the logarithm can be expanded as a power series in $K(\gamma_i)$'s. This yields the polymer (or Mayer) expansion [4]
\[
f_{\text{corr}}(\vec{\eta}) = \frac{1}{n} \sum_{M=1}^{+\infty} \frac{1}{M!} \sum_{\gamma_1,..,\gamma_M} K(\gamma) \prod_{i<j} \delta(\gamma_i \cap \gamma_j = \emptyset) \prod_{G \subseteq \Gamma M} \prod_{(i,j) \in G} \left\{ -\delta(\gamma_i \cap \gamma_j \neq \emptyset) \right\}. \tag{7}
\]

The third sum is over the set $G_M$ of all connected graphs with $M$ vertices labeled by $\gamma_1,..,\gamma_M$, and at most one edge between each pair of vertices. The product of indicator functions is over edges $(i,j) \in G$. It constrains the set of polymers $\gamma_1,..,\gamma_M$ to intersect according to the structure of $G$. In this expansion one sums over an infinite number of terms so it is important to address the question of convergence. A criterion which ensures the convergence of the expansion uniformly in system size $n$ (and thus ensures convergence in the infinite size limit) is
\[
\sum_{t=0}^{+\infty} \frac{1}{t!} \sup_{a \in V} \sum_{\gamma \ni a} |\gamma|^t |K(\gamma)| < 1 \tag{8}
\]

To illustrate the use of the polymer expansion in a simple case, consider a vertex model at high temperature defined by
\[
f_a(\sigma_{ab}) = \frac{1}{2} (1 + \tanh J_a \sum_{b \in \partial a} \sigma_{ab} e^{\frac{1}{2} h_{ab} \sigma_{ab}}\tag{9}
\]
where $J_a$ and $h_{ab}$ are $\in \mathbb{R}$ with $\sup_{a \in V} J_a \equiv J << 1$ and $\sup_{b \in \partial a} h_{ab} < h < +\infty$. For $J$ is small enough the BP (5) equations have a unique fixed point solution [6]. We call $\vec{\eta}_n$ this fixed point. The subscript $n$ indicates (with some abuse of notation) that this fixed point depends on the finite instance, that is, the graph $\Gamma$, and $\vec{h}_n$. For the activities of the polymers, computed at the fixed point, we have the bounds $|K(\gamma)| \leq (2J)^{|\gamma|}$. Moreover the number of polymers $\gamma \ni a$ is (for each $a$) at most $e^{c_d |\gamma|}$ with $c_d \geq 0$ a numerical constant depending only on $d$. Using also that the smallest polymer must have $|\gamma| \geq 3$, it is then easily shown that the left hand side of (8) is $O(J^3) << 1$. By standard methods [4] one can then estimate the sum over $M$ in (7) term by term, which yields
\[
|f_{\text{corr}}(\vec{\eta}_n)| \leq (1 + O(J^3)) \frac{1}{n} \sum_{a \in V} \sum_{\gamma \ni a} (2J)^{|\gamma|} e^{h_{ab} \sigma_{ab}} \tag{9}
\]

Proposition 3.1: For $J < J_0(h)$ small enough, we have
\[
\lim_{n \to +\infty} \mathbb{E}[f_{\text{corr}}(\vec{\eta}_n)] = 0.
\]
Proof idea: From (9) $\mathbb{E}[f_{\text{corr}}(\vec{\eta}_n)] \leq (1 + O(J^3))\mathbb{E}[(\sum_{\gamma \ni a} (2J)^{|\gamma|} e^{h_{ab} \sigma_{ab}})]$ where $o$ is any specified node in the graph. In order to conclude it suffices to use the fact that on a random $d$-regular graph, with probability $1 - o(1)$, polymers have a size $|\gamma| \geq a_d \ln n$, where $a_d \geq 0$ a positive numerical constant.

IV. CYCLE LDPC CODES OVER THE BSC

Random $d$-regular graphs are equivalent to the LDPC($2,d$) ensemble of cycle codes. Code bits $x_{ab} = 0,1$ are attached to the edges $ab \in E$. In the spin language bits are $\sigma_{ab} = \pm 1$ and the parity check constraints are $\prod_{b \in \partial a} \sigma_{ab} = 1$. For definiteness we assume transmission over the BSC($p$), $p \in [0,\frac{1}{2}]$. Without loss of generality one can assume that the transmitted word is (1,...,1) so that MAP decoding is based on the posterior distribution (1) with
\[
f_a(\sigma_{ab}) = \frac{1}{2} (1 + \prod_{b \in \partial a} \sigma_{ab}) \prod_{b \in \partial a} e^{\frac{1}{2} h_{ab} \sigma_{ab}}. \tag{10}
\]
Here $h_{ab}$ is the half-log-likelihood for the bit $\sigma_{ab} = \pm 1$, based on the channel output. The Shannon conditional input-output entropy and free energy are essentially equivalent, and related by the simple formula
\[
\frac{1}{n} H(\vec{X}|\vec{Y}) = E[\hat{n}(\vec{h} \vec{X})] - \frac{1}{2} \ln \frac{1}{p} - \frac{1}{2} \ln \frac{1}{1-p} \tag{11}
\]

where $E_{\mathbf{h}}$ is the average over channel outputs (or the log-likelihood vector).

We interested in the high noise regime where $p$ is close to $1/2$. Therefore we seek solutions of the BP equations such that $\sup_{u \in E} [h_{ab}] \leq h$ where $h > 0$ is a fixed small number. We assumed that for $h$ small enough there exists a fixed point of the BP equations for each finite instance\(^1\). We denote it $\bar{\eta}_{\mathbf{h}}$ as before. Note that assuming its unicity is not needed.

**Proposition 4.1:** Assuming the existence of a fixed point $\bar{\eta}_{\mathbf{h}}$ of the BP equations for $h$ small enough, we have

$$\lim_{n \to +\infty} \mathbb{E}[\frac{1}{n} \ln Z_{\text{corr}}(\bar{\eta}_{\mathbf{h}})] = 0.$$  

In view of (2), (11) the proposition implies that the average conditional entropy is given by the average of the Bethe expression computed at the fixed point (for $|p-\frac{1}{2}| < \varepsilon$).

In order to prove the proposition we will use the identity

$$\ln Z_{\text{corr}}(\bar{\eta}) = \ln Z_p(\bar{\eta}) + \left\{ 1 + \sum_{\gamma \subseteq \Gamma} \left( \frac{|\gamma| > n/2}{\gamma} Z_p(\bar{\eta} | \gamma) \frac{Z_p(\bar{\eta})}{Z_p(\bar{\eta} | \gamma)} \right) \right\}$$  

(12)

where

$$Z_p(\bar{\eta} | \gamma) = \sum_{M \geq 0} \frac{1}{M!} \sum_{\gamma_1, \ldots, \gamma_M} \prod_{i=1}^M K(\gamma_i) \prod_{i < j} \prod_{\gamma_i \cap \gamma_j = \emptyset} \prod_{i=1}^M K(\gamma_i) \prod_{i < j} \prod_{\gamma_i \cap \gamma_j = \emptyset} \prod_{i=1}^M K(\gamma_i) \prod_{i < j} \prod_{\gamma_i \cap \gamma_j = \emptyset}$$

(13)

and $Z_p(\bar{\eta}) \equiv Z_p(\bar{\eta} \emptyset)$. This identity is derived by splitting the sum over $\gamma_1, \ldots, \gamma_M$ in (6), into a sum where all polymers are small ($\forall i, |\gamma_i| < n/2$), and a sum where there exists at least one large polymer ($\exists i, |\gamma_i| \geq n/2$); and by noting that when there exists a large polymer it has to be unique.

We will need three lemmas.

**Lemma 4.2:** For $h$ small enough we have

$$\lim_{n \to +\infty} \mathbb{E}[\frac{1}{n} \ln Z_{\text{corr}}(\bar{\eta}_{\mathbf{h}})] = 0.$$  

**Sketch of Proof:** It is possible to estimate the activities computed at the fixed point,

$$|K(\gamma)| \leq (1 - \alpha d^{-\frac{1}{2}} h^2)^{n_i(\gamma)} \prod_{i=1}^{d-1} (\alpha_i h^{d-1-i})^{n_i(\gamma)}.$$  

(14)

Here $0 < \alpha_i < 1$, and $\alpha_i > 1$, $i = 2, \ldots, d - 1$ are fixed numerical constants (that we can take close to 1). The $n_i(\gamma)$ denotes the number of nodes of degree $i$ in the polymer $\gamma$.

Estimate (14) is essentially optimal for $h$ small, as can be checked by Taylor expanding $K(\gamma)$ in powers of $h_{ab}$. Hard constraints manifest themselves in the factor $(1 - \alpha d^{-\frac{1}{2}} h^2)^{n_i(\gamma)}$ which is not small enough to compensate the entropic term $e^{c_\gamma h^2}$ in the convergence criterion. However for polymers of size $|\gamma| < \frac{n}{2}$ we can use expanded arguments to circumvent this problem. Let $c(\gamma)$ the set of edges in $E$ connecting $g$ to $\Gamma \setminus g$. We say that $\Gamma$ is an expander if for all $g \subseteq \Gamma$ with $|g| \leq \frac{n}{2}$ we have $|c(\gamma)| \geq k|g|$. For all $d \geq 3$ [8]

$$P|\Gamma| \text{ is an expander with } k = 0.18 |d| = 1 - o_n(1)$$  

(15)

\(^1\)This assumption can be relaxed by softening the hard constraint in (10) and using existence results [7]. Indeed all our estimates are uniform in the softening parameter. We omit this discussion here due to lack of space.

Now note that for polymers $c(\gamma) \leq \sum_{i=2}^{d-1} (d - i) n_i(\gamma) \leq d \sum_{i=2}^{d-1} n_i(\gamma)$. Therefore we deduce thanks to (15) that with high probability $\sum_{i=2}^{d-1} n_i(\gamma) \geq 0.18 |\gamma| - 2\varepsilon_0(1)$ for $|\gamma| < \frac{n}{2}$. This is sufficient to control the convergence criterion, and achieve the proof of this lemma by methods similarly to the high temperature case.

**Lemma 4.3:** Fix $\varepsilon > 0$. Then

$$\mathbb{P}[|\gamma| \in \Gamma : e^{-2n\varepsilon} \leq \frac{Z_p(\bar{\eta} | \gamma)}{Z_p(\bar{\eta})} \leq e^{2n\varepsilon}] \geq 1 - \frac{1}{\varepsilon} o_n(1).$$

The proof uses rather trivial bounds on the partition functions. We omit the details.

**Lemma 4.4:** Fix $\delta > 0$. There exists a numerical constant $C > 0$ and such that for $h$ small enough

$$\mathbb{P} \left[ \sum_{g \subseteq \Gamma} |K(g)| \geq \delta \right] \leq C \delta e^{-n_2 \frac{\delta}{4} h^2}.$$  

(16)

This inequality is a fortiori valid for $g$ fixed by $\gamma$’s in the sum.

**Sketch of Proof:** We denote by $K_{\mathbf{h}}$ the complete graph with $n$ vertices. By Markov’s inequality,

$$\mathbb{P} \left[ \sum_{g \subseteq \Gamma} |K(g)| \geq \delta \right] \leq \frac{1}{\delta} \sum_{g \subseteq K_n} \mathbb{E}[|K(g)| | g \subseteq \Gamma]$$

(17)

Consider graphs $g$ with $n_i(g)$, $i = 2, \ldots, d$ fixed. Mackay [9] provides a bound for the probability $\mathbb{P}[g \subseteq \Gamma]$ of finding a particular subgraph into a regular graph $\Gamma$. Namely for $\frac{1}{2} \sum_{i=2}^{d} n_i(g) + 2d^2 \leq \frac{\alpha d}{\pi}$

$$\mathbb{P}[g \subseteq \Gamma] \leq \frac{\prod_{i=2}^{d} (\frac{n}{2} - 2d^2)^{\frac{\alpha d}{\pi}}}{\prod_{i=2}^{d} n_i(g)}.$$  

(18)

The number of subgraphs of $K_n$ with given $n_i(g)$ is

$$\frac{n!}{(n - \sum_{i=2}^{d} n_i(g))! \prod_{i=2}^{d} n_i(g)!} \times \left( \frac{1}{2} \sum_{i=2}^{d} n_i(g) \right)^2 \sum_{i=2}^{d} n^{n_i(g)} \prod_{i=2}^{d} (\frac{\alpha d}{\pi} - 2d^2)^{\frac{\alpha d}{\pi}} n^{\frac{\alpha d}{\pi} - 2d^2}.$$  

(19)

Replacing (18) in (17), using (19), setting $x_i = \frac{n_i(g)}{n}$, and performing an asymptotic calculation for $n$ large, we show (here $\tilde{x} \equiv (x_2, \ldots, x_d)$ and $\Delta \equiv \{ h_{ab} \leq \sum_{i=2}^{d} x_i \leq 1 \}$)

$$\mathbb{P} \left[ \sum_{g \subseteq \Gamma} |K(g)| \geq \delta \right] \leq \frac{1}{\delta} \int_{\Delta} d^{2d} \tilde{x} g_n(\tilde{x}) \exp \left( \frac{n}{2} f_n(\tilde{x}) \right) + x_d \ln(1 - \alpha \frac{d}{2} h^2) + \sum_{i=2}^{d-1} x_i \ln(\alpha_i h^{d-1-i})$$  

(20)
The large $n$ behavior of the integral asymptotic is controlled by $f_n(\vec{x})$, and $g_n(\vec{x})$ gives sub-dominant contributions that do not concern us here. We have

$$f_n(\vec{x}) = \frac{1}{2} \sum_{i=2}^{d} i x_i \ln \left( \frac{1}{2} \sum_{i=2}^{d} i x_i - \frac{1}{2} \sum_{i=2}^{d} x_i \ln \frac{x_i}{\vec{x}} \right)$$

$$(1 - \frac{1}{2} \sum_{i=2}^{d} x_i) \ln (1 - \frac{1}{2} \sum_{i=2}^{d} x_i - \frac{2r^2}{n} \ln (\frac{r}{2} - \frac{2r^2}{n}) + (\frac{r}{2} - 1) \sum_{i=2}^{d} i x_i - \frac{2r^2}{n} \ln (\frac{r}{2} - \frac{2r^2}{n})$$

For $h$ small enough, in the domain $\Delta$, the exponent in (20) is negative in the domain $\Delta$. Its maximum at $x_2 = 0, x_3 = 1$ is $O(h^2)$.

strictly negative and attains its maximum at the corner point $x_2 = \cdots = x_{d-1} = 0, x_d = 1$. At this point it is equal to $\ln(1 - \alpha_d h^2)$ which allows to conclude (16).

We are now in a position to prove proposition 4.1.

**Proof of proposition 4.1:** In view of (12), we must show that for $h$ small enough,

$$\frac{1}{n} \mathbb{E} \left[ \left| \ln \left( 1 + \sum_{\gamma \in \Gamma} |K(\gamma)| \frac{Z_p(\vec{\eta}) | \gamma \rangle}{Z_p(\vec{\eta})} \right) \right| \right] = o_n(1).$$

Call $I_\gamma$ the event

$$\left| \sum_{\gamma \in \Gamma} |K(\gamma)| \frac{Z_p(\vec{\eta}) | \gamma \rangle}{Z_p(\vec{\eta})} \right| < \zeta$$

where $\zeta$ is a positive constant that will be adjusted later on. We split the expectation in two terms $A + B$ by conditioning over $I_\gamma$ and its complement $I_\gamma^c$, and estimate each contribution. For the first contribution, using $|\ln(1 + x)| \leq |\ln(1 - |x|)|$ for $|x| < 1$, $A \leq \frac{1}{n} \mathbb{E} \left[ \left| \ln(1 - \zeta) \right| P[I_\gamma^c] \right] \leq \frac{1}{n} \mathbb{E} \left[ \left| \ln(1 - \zeta) \right| \right]$. For the second contribution we have to estimate $P[I_\gamma]$. The events

$$\left\{ \forall \gamma \in \Gamma : e^{-2\epsilon x} \leq Z_p(\vec{\eta}) | \gamma \rangle \leq e^{2\epsilon x} \right\}$$

and $|\sum_{\gamma \in \Gamma} |K(\gamma)| \leq \delta$ imply $I_{\delta_2x < n}$. Therefore $I_{\delta_2x < n}$ implies the union of the complementary events, so that applying the union bound together with lemmas 4.3 and 4.4,

$$P[I_{\delta_2x < n}] \leq \frac{C}{\delta} e^{-\alpha_d h^2} + 1 \epsilon o_n(1).$$

Now suppose for a moment that there exist a positive constant independent of $n$ such that

$$\frac{1}{n} \mathbb{E} \left[ \ln \left( 1 + \sum_{\gamma \in \Gamma} |K(\gamma)| \frac{Z_p(\vec{\eta}) | \gamma \rangle}{Z_p(\vec{\eta})} \right) \right] \leq C'$$

Then $B \leq C'P[I_\gamma^c]$. Setting $\zeta = 2e^{-2\epsilon h^2}$, the above arguments imply

$$\mathbb{E} \left[ \frac{1}{n} \mathbb{E} \left[ \ln \left( 1 + \sum_{\gamma \in \Gamma} |K(\gamma)| \frac{Z_p(\vec{\eta}) | \gamma \rangle}{Z_p(\vec{\eta})} \right) \right] \right] \equiv A + B$$

$$\leq \frac{1}{n} \mathbb{E} \left[ \ln(1 - 2e^{-2\epsilon h^2}) + C \frac{C}{\delta} e^{-\alpha_d h^2} + 1 \epsilon o_n(1) \right].$$

We are free to choose $\delta = e^{-\alpha_d h^2}$ and $\epsilon = o_n(h^2)$ (lemmas 4.3, 4.4 hold) and this choice $A + B = o_n(1)$, which proves (22).

It remains to justify (23). From the convergence of the polymer expansion we deduce that $\frac{1}{n} \ln Z_p(\vec{1}p) \leq \ln 2 + \frac{1}{2} h$. In the high noise regime the BP messages are bounded so that from (3) we deduce that $\frac{1}{n} \ln Z_{\text{Bethe}}(\vec{1}p)$ is bounded by a constant independent of $n$. Finally the triangle inequality implies that $\frac{1}{n} \ln Z_{\text{Bet}}(\vec{1}p) = \ln Z_{\text{Bethe}}(\vec{1}p) - \ln Z_p(\vec{1}p)$ is bounded uniformly in $n$. This is precisely the statement (23).

V. Conclusion

The approach is quite general and can hopefully be generalized to standard irregular LDPC codes with bounded degrees and binary-input memoryless output-symmetric channels with bounded log-likelihood variables. This will be the subject of future work.

ACKNOWLEDGMENT

M.V. acknowledges support by the Swiss National Foundation for Scientific Research, Grant no 2000-121903. N.M. benefited from discussions with M. Cherftkov and R. Urbanke.

REFERENCES

Factoring linear trellises

David Conti
Claude Shannon Institute and
School of Mathematical Sciences
University College Dublin, Ireland
Email: david.conti@ucdconnect.ie

Nigel Boston
Departments of Mathematics and ECE
University of Wisconsin, USA
Email: boston@math.wisc.edu

Abstract—Koetter/Vardy proved in [9] the remarkable result that any linear trellis can be factored into elementary trellises. We prove that the spans (and their multiplicities) of the elementary factors of a linear trellis are uniquely determined, extending a known result for minimal trellises ([4], [8], [11]). In doing so, we give a graphical method to compute all the spans and their multiplicities. We show also how to determine all the possible edge-labelings of the factors of a linear trellis, and give exact conditions for their uniqueness. These results can help to compute and classify all trellises for a given code.

I. INTRODUCTION

Linear trellises are amongst the most important graph representations of linear codes, both for theoretical and practical reasons. Towards the end of last millennium, due to the observation of striking low complexity achieved by nonconventional trellises ([11]), the interest expanded to the whole class of tail-biting trellises (of which conventional trellises can be seen as a particular case). As a consequence, in the recent years, linear tail-biting trellises has been the subject of major research within trellis theory ([9], [8], [4], [5], [10], [7]). The most important achievement is probably the factorization theorem proven in [9], which states that every linear trellis is a product of elementary trellises. This fundamental theorem opened the road to other central works concerning minimal linear trellis representations of codes ([8], [4]).

In this paper we investigate further linear trellis factorizations and give graphical/combinatorial methods to compute such factorizations. We first prove that if we forget about edge-labels then unique factorization holds, which means that all the spans of factors and their multiplicities are uniquely determined. Our proof yields at the same time a method for computing all the spans and multiplicities. This result was proven before for minimal trellises ([11]), and the striking fact is that this holds true for any linear trellis and that edge-labels play no role. We then show how to find all the possible labelings of the elementary factors of a linear trellis, once the spans’ multiplicities are known. As a corollary, we get precise conditions under which such labelings are uniquely determined. Finally, we mention how these results help us to compute and classify all trellises representing a fixed code, which is important for the search of trellis representations that perform well under iterative or LP decoding.

A. Basics on Trellises

All trellises considered will be over a fixed finite field \( \mathbb{F} \equiv \mathbb{F}_q \) and have a fixed length \( n \). Also, we want to do modular arithmetic on the coordinate indices of \( \mathbb{F}^n \), so \( \mathbb{Z}_n \equiv \mathbb{Z}/n\mathbb{Z} \) will be the index set of coordinates. A trellis (over \( \mathbb{F} \), and of length \( n \)) is a directed graph \( T = (V, E) \) together with a partition \( V = \sqcup_{i \in \mathbb{Z}_n} V_i \) indexed over \( \mathbb{Z}_n \), such that the edges are \( \mathbb{F} \)-labeled and can only go from \( V_i \) to \( V_{i+1} \), so that \( E \) has also a partition \( E = \sqcup_{i \in \mathbb{Z}_n} E_i \), where \( E_i \subseteq V_i \times \mathbb{F} \times V_{i+1} \). For formal reasons we also allow multiple edges in a trellis, i.e. each edge \( e \in E \) also has a multiplicity \( m(e) \geq 1 \). A trellis is conventional if \( |V_0| = 1 \).

Two trellises \( T = (V, E), T' = (V', E') \) are isomorphic if there exists a bijection \( f : V \rightarrow V' \) such that \( e = (v, a, w) \in E \) if and only if \( f(v) = (f(v), f(a), f(w)) \in E' \), \( m(e) = m(f(e)) \) for all \( e \in E \), and \( f(V_0) = V'_0 \). If by forgetting their labels \( T \) and \( T' \) are isomorphic as directed multigraphs, then we say that \( T \) and \( T' \) are structurally isomorphic (again, we require \( V_0 \) to be sent to \( V'_0 \)) and write \( T \cong T' \).

By a cycle of a trellis \( T \) we mean a closed directed path in \( T \) of length \( n \) starting and ending at \( V_0 \). The subset \( C(T) := \{ \text{edge-label sequences of cycles of } T \} \subseteq \mathbb{F}^n \) is the code represented by \( T \). A trellis is reduced if every edge and every vertex belongs to a cycle. A trellis is linear if it is reduced, the multiplicity function \( m : E \rightarrow \mathbb{N} \) on each \( E_i \) is a constant of the form \( q^i \) for some \( j \geq 0 \), and all \( V_i \)'s have an \( \mathbb{F} \)-vector space structure such that \( E_i \) is an \( \mathbb{F} \)-subspace of \( V_i \times \mathbb{F} \times V_{i+1} \) for all \( i \). If \( T \) and \( T' \) are isomorphic then \( T \) is linear if and only if \( T' \) is linear. Also, if \( T \) is linear then so is \( C(T) \). In this paper we will consider only linear trellises, so, to simplify the terminology, from now on by a trellis we will always mean a linear trellis.

We will denote by \( \mathbb{T} \) the set of trellises up to isomorphism, and by \( \mathbb{T} \) the set of unlabeled trellises, i.e. trellises up to structural isomorphism. Clearly \( \mathbb{T} \) can be identified with the subset of all trellises having all edge-labels equal to 0, so, we see \( \mathbb{T} \subseteq \mathbb{T} \). We have a projection map \( \mathbb{T} \rightarrow \mathbb{T} \), which to each trellis \( T \) associates its underlying unlabeled trellis \( \mathbb{T} \) (equivalently, putting all edge-labels of \( T \) equal to 0).

For \( 0 \leq h \leq n \) and \( i \in \mathbb{Z}_n \), \( (i, i+h] := \{ i+1, i+2, \ldots, i+h \} \subseteq \mathbb{Z}_n \) is a span of length \( h \) of \( v \in \mathbb{F}^n \) if \( \text{supp}(v) \subseteq [i, i+h] := (i) \cup (i, i+h] \). If \( h < n \) we say that the span \( (i, i+h] \) is proper, and that it starts at \( i \) and ends at \( j = i+h \).
The degenerate span \( Z_n \) is a span of any \( v \in \mathbb{F}^n \). Now, let \( s \) be a span of \( v \in \mathbb{F}^n \). Let \( V_t := \mathbb{F} \) for all \( t \in s \) and else \( V_t := 0 \). Choose \( \omega_t \in V_t \setminus \{0\} \) for all \( t \notin s \), while for all \( t \notin \omega \) put \( \omega_t := 0 \). Let then \( E_t := \mathbb{F}(\omega_t, v_t, w_{t+1}) \), i.e. \( E_t \) is the subspace of \( V_t \times V_t \times V_{t+1} \) generated by the vector \((\omega_t, v_t, w_{t+1})\). We also put \( m(e) := 1 \) for all \( e \in E \), except in the case \( v[s] = 0 \{i, i\} \) where we put \( m(e) = q \) for the (only) edge \( e = (0, 0, 0) \in E_t \) (i.e. \( e \) is repeated \( q \) times). This gives the element \( e \) (up to trellis isomorphism, it doesn’t depend on the choice of the \( \omega_t \)). It is clearly a (linear) trellis, and it satisfies \( v[s] = \lambda v[0] \) for \( \lambda \neq 0 \). Now, let \( E(v) := \mathbb{F}_v \). Notice also that \( v[s] = v[t][s] \) if and only if \( s = s' \), and that since \( \mathbb{F} \) is finite, the set \( T_{E(s)} \) of elementary trellises is finite too.

B. Trellis products

\( T \) benefits of a natural monoid structure \((T, \otimes)\) given as follows. Let \( T, T' \in T \) be two trellises. For \( e = (v_t, a, v_{t+1}) \in E_t', e' = (v'_t, a', v'_{t+1}) \in E'_t \) define \( e \otimes e' := ((v_t, v'_t), a + a', (v_{t+1}, v'_{t+1})) \). For each \( i \) let \( V'_t := V_t \times V'_t \) and let \( E'_{0} := \{ e \otimes e' | e \in E_t, e' \in E'_t \} \). Then the trellis \( T \times T' := (\bigcup V'_t, \bigcup E'_0) \) is called the trellis product of \( T \) and \( T' \). This is the usual definition of trellis product as given for example in [8], [9]. Since we also allow multiple edges in a trellis, we naturally define the multiplicity of \( e'' \in E'' \) as \( m(e'') := \sum_{e \in E''} m(e)m(e') \). It’s easy to see that if \( T \) is a product of elementary trellises then \( T \) has no multiple edges if and only if no factor of span length 0 is repeated and no factor is of the form \( 0[i, i] \).

The operation \( \otimes \) is associative, commutative, and has an identity, the zero trellis \( 0 \in T \), i.e. \((T, \otimes)\) is a commutative monoid. Note that the only invertible element of \((T, \otimes)\) is the zero trellis, that the subset \( \overline{T} \) is a submonoid of \((T, \otimes)\), and \( \overline{T} \overline{T} = \overline{T} \).

Also, a fundamental property of the trellis product is the identity \( C(T \otimes T') = C(T) + C(T') \), from which follows that any linear code can be represented by a product of elementary trellises. That any trellis is a product of elementary trellises is the following famous factorization theorem (see [9]):

**Theorem 1 (KV).** Any (linear) trellis is a product of elementary trellises.

II. UNIQUE FACTORIZATION IN \( T \)

Knowing that every trellis can be factored into elementary trellises, a natural following question is whether the factorization must be unique or not. The answer is that in general it does not. For example, \( 111[(0, 2] \otimes 010[(1, 1] = 101[(0, 2] \otimes 010[(1, 1] \), but the two factorizations are different, as \( 111[(0, 2] \neq 101[(0, 2] \). Nevertheless, one sees here that the two factorizations have precisely the same set of paths, i.e. up to labels the two factorizations are equal. This is no coincidence, as it is a well known result that different factorizations of the minimal conventional trellis for a linear code must have the same set of spans (for details on minimal and conventional trellises see [11], [8]), a result that is proved only as a byproduct of the minimality assumption. More recently the same result was also proven for minimal nonconventional trellises (see [8], [4]), again by exploiting the minimality assumption. We show that this holds actually for any trellis, and that the edge-labels really play no role in determining the spans of factors. The structure alone determines the spans. Here is our main result of this section:

**Theorem 2.** In \((T, \otimes)\) unique factorization holds, that is, if \( T_1 \otimes T_2 \otimes \ldots \otimes T_r = T'_1 \otimes T'_2 \otimes \ldots \otimes T'_r \), with \( T_i, T'_i \in T_{E(s)} \), then \( r = r' \) and, up to order, \( T_i = T'_i \) for all \( i \). In other words, the spans (along with their multiplicities) appearing in a factorization of a trellis \( T \) are uniquely determined.

It is easy to see that the multiplicity of the degenerate span \( Z_n \) is uniquely determined and is equal to the number of connected components of \( T \), so the real problem is understanding what happens with proper spans. In what follows we will tackle this problem by a graph-theoretical approach, which yields also a method to compute the structural factorization. Before going on to prove Theorem 2, we also inform the reader that we were recently told by Heide Gluesing-Luerssen that her PhD student Elizabeth Weaver has independently obtained uniqueness of spans of factors in the particular case of one-to-one trellises.

A. Graph-theoretical approach

The graph-theoretical approach involves studying the earliest intersections of paths starting along different edges from a fixed vertex. Not only does this enable to prove unique factorization in \((T, \otimes)\), but from the same intersection data one is also able to determine precisely all the spans of the factors.

Now, let us give some notation. Given a multiset \( S \), we write \( m(s, x) \) for the multiplicity of \( x \) in \( S \) (in particular \( m(x, S) > 0 \) if and only if \( x \in S \)). If \( T \) is a trellis and \( e = (v, a, v') \) is an edge of \( T \), then \( h(e) := v' \) is the head of \( e \), and \( t(e) := v \) is its tail. A path \( p = p_0 p_1 \ldots p_r \) in \( T \) is a sequence of edges \( p_0, p_1, \ldots, p_r \in T \) such that \( h(p_i) = t(p_{i+1}) \), \( i = 0, \ldots, r-1 \).

Now, let \( e \neq e' \) be two different edges of \( T \) such that \( t(e) = t(e') \). Define \( I(e, e') \) to be the smallest \( r \geq 0 \) such there exist two directed paths \( p = p_0 \ldots p_r = p_0' \ldots p_r' \) in \( T \) satisfying \( p_0 = e, p_0' = e' \), and \( h(p_i) = h(p_i') \). If there exist no directed paths satisfying those conditions then we put \( I(e, e') = \infty \). Define then the multiset \( I(e) := \{I(e, e') | e' \neq e, t(e') = t(e)\} \).

**Example 1.** Consider the unlabeled trellis

```
\[e\rightarrow e'\]
```

Then \( I(e) = \{1\} \), \( I(e') = \{2\} \), and \( I(e'') = \{1, 1\} \).

Now, let \( T \) be a trellis and let \( e, e' \) be two edges with same tail. A priori, to compute \( I(e, e') \) one needs to compute the set of all possible couple of paths \( p, p' \) starting along
e, e', and look for the earliest possible intersection. But in
fact, by linearity (we are only considering trellises with linear
structure) one can fix p and let only p' vary:

**Observation 1.** Let T be a trellis and let e, e' be two edges
such that t(e) = t(e'). Fix a path p such that p_0 = e. Then
I(e, e') = \min\{r | \exists p' = p'_0 \ldots p'_r such that p'_0 = e', h(p'_r) = h(p_0)\}.

The highly symmetrical structure of (linear) trellises is
further reflected in the following fundamental lemma:

**Lemma 1.** Let T be a trellis of length n. Then:
- I(e, e') \leq n - 1 for any edges e, e'.
- I(e) = I(e') if t(e), t(e') \in V_i for some i.

In sight of this result, for a trellis T and i ∈ Z_n it is
legitimate to define I_i(T) := I(e), where e is any edge such
that t(e) ∈ V_i.

The next lemma tells us that I_i(T) is determined only by
those factors of T whose spans start at i.

**Lemma 2.** Let T be a trellis, and let T' ∈ T_{Ei} be an
elementary trellis with span not starting at i or degenerate
span. Then I_i(T ⊗ T') = I_i(T).

The above lemma is crucial for the following theorem,
which is the key result:

**Theorem 3.** Let T = T_1 \oplus \ldots \oplus T_r, with T_1, \ldots, T_r ∈ T_{Ei}. Fix
i ∈ Z_n, and let L_i(T) be the multiset of span lengths of those
T_j's with (proper) span starting at i. Then for l = 0, \ldots, n - 1

log_2 \left(1 + \sum_{j=0}^{l} m(j, I_i(T))\right) = \sum_{j=1}^{l} m(0, L_i(T))

The above equations can be recursively solved for
m(j, L_i(T)), j = 0, \ldots, n - 1, and the solution is unique.
Therefore, since the left-hand side depends only on T, we
conclude that Theorem 2 holds. Moreover, by computing all
I_i(T)'s we are able to determine all (proper) spans and their
multiplicities in any factorization of T.

**Example 2.** Let T be the unlabeled connected trellis

This trellis is F_2-linear, its linear structure given by the shown
F_2-labels of vertices. Then one easily computes:
- I_0(T) = I(e_0) = \emptyset
- I_1(T) = I(e_1) = \{2\}
- I_2(T) = I(e_2) = \emptyset
- I_3(T) = I(e_3) = \{3, 4, 4\}

So, by Theorem 3, we get that L_0(T) = L_2(T) = L_4(T) = \emptyset,
L_1(T) = \\{\{3\}\}, L_3(T) = \\{\{4, 5\}\}. Therefore T = \{\{1, 3\} \oplus \{\{3, 1\} \oplus \{3, 2\}\}

III. FACTORIZATION IN T

We have already pointed out that in (T, \otimes) unique
factorization doesn't hold, since the labels are not necessarily
uniquely determined. Nevertheless, we give a method to find
a legitimate labeling of a structural factorization of a trellis.
We will also deduce precise conditions for a trellis T under
which unique factorization in (T, \otimes) holds.

Let T ∈ T, and let \mathcal{S}(T) be the span multiplet of T, which
by Theorem 2 is uniquely determined. If (i, j) ∈ \mathcal{S}(T), i.e.
m(i, j, \mathcal{S}(T)) \geq 1, we say that (i, j) is a span of T. Also,
we put s := [i, j] for a proper span s = [i, j], and \mathcal{Z}_n = Z_n.

By linearity there exists a zero vertex z_i ∈ V_i for all i. Let
(i, j) be a proper span of T, and let C_{(i,j)}(T) denote the code
of all label-sequences of paths starting at z_i and ending at z_j.
This is a code of length equal to the span length of (i, j),
but it is naturally a code in F^n by setting the coordinates
outside the span (i, j) equal to 0. It is not hard to see that it
is linear, since T is so. Also, for the degenerate span \mathcal{Z}_n we
put C_{\mathcal{Z}_n}(T) := C(T). Then we have the following:

**Theorem 4.** Let T ∈ T, and let s be a span of T of multiplicity
r = m(s, \mathcal{S}(T)). Then \mathcal{F}_v[s \oplus \ldots \oplus \mathcal{F}_v]s divides T if and only if

\mathcal{F}_{v_1} + \ldots + \mathcal{F}_{v_r} + \sum_{s' \in \mathcal{S}(T) | s' \subset s} C_{s'}(T) = C_s(T)

The following corollary extends the well known result that if
there are no containments between different atomic spans of a
code then the associated minimal conventional trellis factorizes
uniquely, since all atomic generators are uniquely determined
(up to scalar multiplication).

**Corollary 1.** Let T = v_1[i_1, j_1] \oplus \ldots \oplus v_r[i_r, j_r]. Then
v_1, \ldots, v_r are uniquely determined up to scalar multiplication
and ordering (i.e. the factorization of T is unique) if and only
if there are no containments between different spans of T and
all their multiplicities are equal to 1.

**Example 3.** Label the edges of T from example 2 as follows
(by using dashed lines for 0’s and full lines for 1’s):

Thus labeled, T is again F_2-linear. Then one gets:
- C_{(3,1)}(T) = \{00000, 01010\}
\[ C_{[3,1]}(T) = \{00000, 01011\} \]
\[ C_{[3,2]}(T) = \{00000, 01011\} \]

Therefore by Theorem 4 and its corollary \( T \) has two (and only two) distinct factorizations,

\[
T = 01010([3,1] \otimes 0101)|([3,1] \otimes 0101)|([3,2] \\
T = 01010([3,1] \otimes 0101)|([3,1] \otimes 0000)|([3,2] 
\]

Note also that \( T' = 01010([3,1] \otimes 0000)|([3,1] \otimes 0111)|([3,2] \) (depicted below) is a different trellis from \( T \), according to Theorem 4. In fact, in \( T' \) it is possible to go out from \( 0 \in V_3 \) along two paths with all edge-labels equal to 0, and that meet at \( V_2 \), while this is not possible in \( T \).

![Trellis Diagram](image)

IV. APPLICATIONS

The complexity of trellis-based decoding is directly proportional to the size of the chosen representing trellis, so, to achieve low complexity it is important to search for the smallest trellis representations of codes. If trellis size is all what one is interested in, then obviously the way labels are arranged on a trellis doesn’t matter (as long as the trellis represents the wanted code), which implies that it is sufficient to classify all the trellis representations up to structural isomorphism. In [8] Koetter and Vardy show that up to structural isomorphism all the minimal trellises for a given code are encoded into a single matrix (characteristic matrix), and show how to efficiently compute that matrix (see also [4]).

On the other hand, the performance of iterative or LP decoding algorithms on a trellis is affected by so-called pseudocodewords, which makes it important to sort trellises also with respect to their pseudocodewords, as one wants to find trellises that yield few bad pseudocodewords ([3], [6]). In [2] we have given an example which shows that by changing labels in a trellis it is possible to change the set of yielded pseudocodewords and yet represent the same code. This means that to sort all the trellis representations of a given code with respect to pseudocodewords it is not sufficient to do a classification of trellises up to structural isomorphism. We need a complete classification of trellises. Theorem 2 and Theorem 4 help in that direction. Indeed they yield an exact method to check whether two trellises are isomorphic or not.

To see this recall that (by Theorem 1) to define a trellis we need to specify a span multiset and a list of vectors that is compatible with that span multiset. Theorem 2 tells us first of all that two different span multisets always yield two different trellises (since they are not even structurally isomorphic!). Once a span multiset (and so by Theorem 2 the unlabeled graph structure of the trellis) is chosen, Theorem 4 tells us the precise relations that hold among all the possible lists of vertices that give the same trellis. More precisely, if the chosen spans are \( s_1, \ldots, s_\ell \) with respective multiplicities \( m_1, \ldots, m_\ell \), and \( \{v_{ij}\}_{1\leq i \leq \ell, 1\leq j \leq m_i}, \{w_{ij}\}_{1\leq i \leq \ell, 1\leq j \leq m_i} \) are lists of vectors associated to the spans, then it is easy to see from Theorem 4 that the corresponding trellises are isomorphic if and only if for all \( i = 1, \ldots, \ell \) the equation

\[
\sum_{h \in \tilde{s}_i \subseteq \tilde{s}_i} \sum_{1 \leq j \leq m_h} F_{h,j} = \sum_{1 \leq j \leq m_h} \sum_{h \in \tilde{s}_i \subseteq \tilde{s}_i} F_{h,j}
\]

holds.

Remark 1. It is easy to show that if a trellis \( T \) for a given code is optimal with respect to iterative or LP decoding (in the sense that it yields few bad pseudocodewords) then it must be one-to-one, i.e. the list of generating vectors of \( T \) must be a basis of the code. So, for iterative or LP decoding purposes, it is actually sufficient to do a complete classification of one-to-one trellises.

ACKNOWLEDGMENTS

This work is part of the Ph.D. research of D. Conti supported by the Science Foundation Ireland grant 06/MI/006, and supervised by N. Boston.

REFERENCES

Abstract—In the splitting model, information theoretic authentication codes allow non-deterministic encoding, that is, several messages can be used to communicate a particular plaintext. Certain applications require that the aspect of secrecy should hold simultaneously. Ogata–Kurosawa–Stinson–Saido (2004) have constructed optimal splitting authentication codes achieving perfect secrecy for the special case when the number of keys equals the number of messages. In this paper, we establish a construction method for optimal splitting authentication codes with perfect secrecy in the more general case when the number of keys may differ from the number of messages. To the best knowledge, this is the first result of this type.

I. INTRODUCTION

The development of quantum computer resistant cryptograhic schemes and security technologies is of crucial importance for maintaining cryptographic long-term security and/or confidentiality of digital data, as classical cryptographic primitives such as RSA, DSA, or ECC would be easily broken by future quantum computing based attacks (e.g., [1], [2]). Application areas where cryptographic long-term security and/or confidentiality is strongly required include archiving official documents, notarial contracts, court records, medical data, state secrets, copyright protection as well as further areas concerning e-government, e-health, e-publication, et cetera.

To this end, one promising approach is the design of information theoretic authentication and secrecy systems (e.g., [3], [4]). The information theoretic, or unconditional, security model does not depend on any complexity assumptions and hence cannot be broken given unlimited computational resources. This guarantees not only resistance against future quantum computing based attacks but also perfect security in the classical world.

This paper considers authentication and secrecy codes in the splitting model. Splitting is of importance, for instance, in the context of authentication with arbitration (i.e., protection against insider attacks in addition to outsider attacks). Ogata–Kurosawa–Stinson–Saido [6] have constructed optimal splitting authentication codes with perfect secrecy for the special case when the number of keys equals the number of messages. In this work, we develop a construction method for optimal splitting authentication codes with perfect secrecy in the more general case when the number of keys may differ from the number of messages. To the best knowledge, this is the first result of this type. Our simple yet powerful approach is based on the notion of cyclic splitting designs and establishes an efficient method to construct optimal splitting authentication codes with perfect secrecy.

II. THE SPLITTING MODEL

We rely on the information theoretical, or unconditional secure, authentication model developed by Simmons (e.g., [7], [8]). Our notation follows [6], [9], [10]. In this model, three participants are involved: a transmitter, a receiver, and an opponent. The transmitter wants to communicate information to the receiver via a public communications channel. The receiver in return would like to be confident that any received information actually came from the transmitter and not from some opponent (integrity of information). The transmitter and the receiver are assumed to trust each other. An authentication code is sometimes called, for short, an A-code.

Let $S$ denote a finite set of source states (or plaintexts), $M$ a finite set of messages (or ciphertexts), and $E$ a finite set of encoding rules (or keys). Using an encoding rule $e \in E$, the transmitter encrypts a source state $s \in S$ to obtain the message $m = e(s)$ to be sent over the channel. The encoding rule is communicated to the receiver via a secure channel prior to any messages being sent. When it is possible that more than one message can be used to communicate a particular source state $s \in S$ under the same encoding rule $e \in E$, then the authentication code is said to have splitting. In this case, a message $m \in M$ is computed as $m = e(s,r)$, where $r$ denotes a random number chosen from some specified finite set $R$. If we define

$$e(s) := \{ m \in M : m = e(s,r) \text{ for some } r \in R \}$$

for each encoding rule $e \in E$ and each source state $s \in S$, then splitting means that $|e(s)| > 1$ for some $e \in E$ and some $s \in S$. In order to ensure that the receiver can decrypt the message being sent, it is required for any $e \in E$ that $e(s) \cap e(s') = \emptyset$ if $s \neq s'$. For a given encoding rule $e \in E$, let

$$M(e) := \bigcup_{s \in S} e(s)$$

denote the set of valid messages. A received message $m$ will be accepted by the receiver as being authentic if and only if
For every encoding rule \( e \in E \) and every source state \( s \in S \), the receiver and transmitter also choose a splitting design with \( |S| = s \). Then the number of blocks containing each element of \( S \) as per Definition 1 is given by

\[
\lambda_s = \frac{\lambda_{(v-s)}^{(u-s)}}{c^{(t-s)}}.
\]

In particular, for \( t \geq 2 \), a \( t-(v, b, l = cu, \lambda) \) splitting design is also an \( s-(v, b, l = cu, \lambda_s) \) splitting design.

Proposition 2: Let \( D = (X, B) \) be a \( t-(v, b, l = cu, \lambda) \) splitting design. Let \( r := \lambda_1 \) denote the number of blocks containing a given point. Then the following holds:

(a) \( bl = vr \).

(b) \( \left( \frac{v}{t} \right)^{\lambda} = b \left( \frac{u}{t} \right)^{\lambda} \).

(c) \( r e^{t-1}(u-1) = \lambda_2(v-1) \) for \( t \geq 2 \).

Proposition 3: Let \( D = (X, B) \) be a \( t-(v, b, l = cu, \lambda) \) splitting design. Then

\[
\lambda \left( \frac{v-s}{t-s} \right) \equiv 0 \mod c^{t-s} \left( \frac{u-s}{t-s} \right)
\]

for every positive integer \( s \leq t \).

Proposition 4: If \( D = (X, B) \) is a \( t-(v, b, l = cu, \lambda) \) splitting design with \( t \geq 2 \), then

\[
b \geq \frac{v}{u}.
\]

IV. OPTIMAL SPLITTING AUTHENTICATION CODES

We state lower bounds on cheating probabilities for splitting authentication codes (cf. [12], [13]).

Theorem 1: In a splitting authentication code, for every \( 0 \leq i \leq t \), the deception probabilities are bounded below by

\[
P_{di} \geq \min_{e \in E} \frac{|M(e)| - i \cdot \max_{e \in S} |e(s)|}{|M| - i}.
\]

A splitting authentication code is called \( t \)-fold secure against spoofing if

\[
P_{di} = \min_{e \in E} \frac{|M(e)| - i \cdot \max_{e \in S} |e(s)|}{|M| - i}
\]

for all \( 0 \leq i \leq t \).

We indicate a lower bound on the size of encoding rules for splitting authentication codes (see [10], and [14], [15] for the case \( t = 2 \)).

Theorem 2: If a splitting authentication code is \( (t-1) \)-fold secure against spoofing, then the number of encoding rules is bounded below by

\[
|E| \geq \prod_{i=0}^{t-1} \frac{|M| - i \cdot \max_{e \in S} |e(s)|}{|M| - i}.
\]

A splitting authentication code is called optimal if the number of encoding rules meets the lower bound with equality.
Corollary 1: In a $c$-splitting authentication code,
\[ P_d \geq \frac{c(|S| - i)}{|M| - i} \]
for every $0 \leq i \leq t$.

Corollary 2: If a $c$-splitting authentication code is $(t - 1)$-fold secure against spoofing, then
\[ |E| \geq \frac{|M|}{c^t(|S| - 1)} \]

Optimal splitting authentication codes can be characterized in terms of splitting designs (see [10], and [6] for the case $t = 2$) as follows.

Theorem 3: Suppose there is a $t$-$(v, b, l = cu, 1)$ splitting design with $t \geq 2$. Then there is an optimal $c$-splitting authentication code for $u$ equiprobable source states, having $v$ messages and $\binom{v}{t}$ encoding rules, that is $(t - 1)$-fold secure against spoofing. Conversely, if there is an optimal $c$-splitting authentication code for $u$ source states, having $v$ messages and $\binom{v}{t}$ encoding rules, that is $(t - 1)$-fold secure against spoofing, then there is a $t$-$(v, b, l = cu, 1)$ splitting design.

V. OPTIMAL SPLITTING AUTHENTICATION CODES WITH PERFECT SECRECY

In what follows, we are interested in optimal splitting authentication codes that simultaneously achieve perfect secrecy. According to Shannon [16], an authentication code is said to have perfect secrecy if
\[ p_E(s|m) = p_E(s) \]
for every source state $s \in S$ and every message $m \in M$, that is, the a posteriori probability that the source state is $s$, given that the message $m$ is observed, is identical to the a priori probability that the source state is $s$.

By introducing the notion of an external difference family (EDF) (which yields a certain type of a splitting design), Ogata–Kuroswa–Stinson–Saido [6, Thm. 3.4] have given a construction scheme for optimal splitting authentication codes with perfect secrecy in the special case when the number of keys equals the number of messages.

Theorem 4: Suppose there exists a $(v, c, 1)$-EDF over an Abelian group of order $v$, then there is an optimal $c$-splitting authentication code for $u$ equiprobable source states, having $v$ messages and $v$ encoding rules, that is one-fold secure against spoofing and simultaneously achieves perfect secrecy.

An example is as follows (cf. [6, Exs. 2.3 & 5.2]).

Example 1: An optimal 2-splitting authentication code for $u = 2$ equiprobable source states, having $v = 9$ messages and $b = 9$ encoding rules, that is one-fold secure against spoofing and achieves perfect secrecy can be constructed from a 2-$(9, 9, 4 = 2 \times 2, 1)$ splitting design. Each encoding rule is used with probability $1/9$. An encoding matrix is given in Table 1.

<table>
<thead>
<tr>
<th>$s_1$</th>
<th>$s_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>3.5</td>
</tr>
<tr>
<td>2.3</td>
<td>4.6</td>
</tr>
<tr>
<td>3.4</td>
<td>5.7</td>
</tr>
<tr>
<td>4.5</td>
<td>6.8</td>
</tr>
<tr>
<td>5.6</td>
<td>7.9</td>
</tr>
<tr>
<td>6.7</td>
<td>8.1</td>
</tr>
<tr>
<td>7.8</td>
<td>9.2</td>
</tr>
<tr>
<td>8.9</td>
<td>1.3</td>
</tr>
<tr>
<td>9.1</td>
<td>2.4</td>
</tr>
</tbody>
</table>

In the following, we develop a construction method for obtaining optimal splitting authentication codes with perfect secrecy in the more general case when the number of keys may differ from the number of messages:

1. We first introduce the notion of a cyclic splitting design. Let $D = (X, B)$ be a $2$-$(v, b, l = cu, 1)$ splitting design, and let $\sigma$ be a permutation on $X$. For a block $B_i = \{B_{i,1}, \ldots, B_{i,u}\} \in B$ given as in (ii) of Definition 1, define $B^\sigma_i := \{B^\sigma_{i,1}, \ldots, B^\sigma_{i,u}\}$, satisfying
\[ B^\sigma_i = B^\sigma_{i,1} \cup \cdots \cup B^\sigma_{i,u} \]
with $|B^\sigma_{i,1}| = \cdots = |B^\sigma_{i,u}| = c$ and $|B^\sigma_i| = l = cu$. If $B^\sigma := \{B^\sigma_i : B_i \in B, 1 \leq i \leq b = B\}$, then $\sigma$ is called an automorphism of $D$. If there exists an automorphism $\sigma$ of order $v$, then $D$ is called cyclic. In this case, the point-set $X$ can be identified with $\mathbb{Z}_v$, the set of integers modulo $v$, and $\sigma$ can be represented by $\sigma : j \rightarrow j + 1$ (mod $v$). For a block $B_i = \{B_{i,1}, \ldots, B_{i,u}\}$, the set
\[ B_i + j := \{B_{i,1} + j \mod v, \ldots, B_{i,u} + j \mod v\} \]
for $j \in \mathbb{Z}_v$ is called a translate of $B_i$, and the set of all distinct translates of $B_i$ is called the orbit containing $B_i$. If the length of an orbit is $v$, then the orbit is said to be full, otherwise short. A block chosen arbitrarily from an orbit is called a base block (or starter block). For a cyclic 2-$(v, b, l = cu, 1)$ splitting design to exist, a necessary condition is $v \equiv 1 \mod (u(u - 1)c^2)$. When $v \equiv 1 \mod (u(u - 1)c^2)$ all orbits are full.

2. Let us assume that there exists a cyclic 2-$(v, b, l = cu, 1)$ splitting design without short orbit. Then, by Theorem 3, there is an optimal $c$-splitting authentication code for $u$ equiprobable source states, having $v$ messages and $\binom{v}{t}/c^t$ encoding rules, that is one-fold secure against spoofing. Furthermore, when considering the corresponding $b \times u$ encoding matrix, it follows by constructive reasons from the underlying cyclic splitting design without short orbit that the code simultaneously achieves perfect secrecy under the assumption that the encoding rules are used with equal probability.

Hence, we have proved the following theorem.
**Theorem 5:** Suppose there is a cyclic $2-\{v, b, l = cu, 1\}$ splitting design without short orbit (that is, it holds that $v \equiv 1 \pmod{u(u - 1)c^2}$). Then there is an optimal $c$-splitting authentication code for $u$ equidrawable source states, having $v$ messages and $(\binom{v}{c})/[v^2 \binom{c}{2}]$ encoding rules, that is one-fold secure against spoofing and simultaneously achieves perfect secrecy.

Relying on some recent constructions of splitting designs (cf. [17, Sect. 3.2]), we give exemplarily a series of optimal splitting authentication codes with perfect secrecy.

**Example 2:** (i) An optimal $2$-splitting authentication code for $u = 2$ equidrawable source states, having $v = 17$ messages and $b = 34$ encoding rules, that is one-fold secure against spoofing and achieves perfect secrecy can be constructed from a cyclic $2-\{17, 34, 4 = 2 \times 2 \times 1\}$ splitting design with base blocks $\{\{1, 2\}, \{3, 5\}\}$ and $\{\{1, 2\}, \{11, 13\}\}$. Each encoding rule is used with probability $1/34$. An encoding matrix is given in Table II.

(ii) As generalization of (i), an optimal $c$-splitting authentication code for $u = 2$ equidrawable source states, having $v = 2c^2h + 1$ messages and $b = (2c^2h + 1)h$ encoding rules, that is one-fold secure against spoofing and achieves perfect secrecy can be constructed from a cyclic $2-\{2c^2h + 1, (2c^2h + 1)h, l = c \times 2, 1\}$ splitting design with base blocks $\{\{1, 2, \ldots, c\}, (2c^2h - (2c^2 - c) + 1, 2c^2h - (2c^2 - c) + c + 1, \ldots, 2c^2h - (2c^2 - c) + (c - 1) + 1\}\}$ for all $1 \leq h \leq n$.

(iii) Further examples of splitting authentication codes with perfect secrecy, also for $u > 2$, can be obtained in the same way from various further constructions of splitting designs in [17, Sect. 3.2].

**ACKNOWLEDGMENT**

The author gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) via a Heisenberg grant (Hu954/4) and a Heinz Maier-Leibnitz Prize grant (Hu954/5).

**REFERENCES**


Anti-Structure Problems

Ram Zamir
Department of Electrical Engineering-Systems
Tel Aviv University, Israel
zamir@eng.tau.ac.il

Abstract—The recent success of structured solutions for a class of information-theoretic network problems, calls for exploring their limits. We show that sum-product channels resist a solution by structured (as well as random) codes. We conclude that the structured approach fails whenever the channel operations do not commute (or for general functional channels, when the channel function is non decomposable).

Key words: sum-product channels, distributed coding, functional source coding, functional Gelfand-Pinsker problem, non-decomposable functions, commutativity, associativity.

I. INTRODUCTION

Structured codes can be effective, and significantly better than random codes, for various multiuser problems. Prominent examples include the well known Korner-Marton (KM) “modulo-two sum” problem [1], as well as more recent setups such as the “dirty” multiple-access channel (MAC) [2], the noisy linear network (along with the compute & forward relaying technique) [3], and more [4]. The effectiveness of structured codes, in particular linear or lattice codes, is due to a good match between their linear structure and the additive nature of the source or channel involved.

In the Korner-Marton problem, for example, the two components \((X,Y)\) of a doubly-symmetric binary source are encoded separately. The joint decoder is not interested in a full reconstruction of \(X\) and \(Y\), which are viewed as “helper sources”; rather the decoder is interested in their modulo-two sum \(X+Y\) (or Xor). Writing the statistical relation between \(X\) and \(Y\) as a modulo-additive noise channel

\[ Y = X + Z \]  

we can recast the problem as that of lossless reconstruction of the noise \(Z\) from separate encodings of \(X\) and \(Y\).

As shown by Korner and Marton, a linear structured coding scheme, which sends the syndromes of \(X^n\) and \(Y^n\) with respect to a suitable linear binary code, achieves the minimum possible rate of \(H(Z)\) - the entropy of \(Z\) - per each encoder. In contrast, a conventional random coding scheme cannot do better than to encode at a rate equal to the joint entropy of \(X\) and \(Y\). This corresponds to the Slepian-Wolf solution [5], i.e., to a lossless reconstruction of both \(X\) and \(Y\) at the decoder. The resulting sum rate can be therefore significantly higher than \(2H(Z)\) for highly correlated sources.

The binary KM problem can be generalized to a \(q\)-ary field, in which case a linear \(q\)-ary code replaces the linear binary code in the KM solution [6]. And it also has a quadratic-Gaussian version [7].

A dual example with a similar characteristics is that of the “doubly dirty” MAC. This channel extends Costa’s “writing on a dirty paper” problem [8], [9] to a MAC; i.e., an additive-noise channel with two inputs \(X_1\) and \(X_2\) and an output \(Y\) given by

\[ Y = X_1 + X_2 + S_1 + S_2 + \text{noise} \]  

where \(S_1\) and \(S_2\) are two interferences, each known as “side information” to one of the encoders. Addition in (2) is over some group in the discrete channel case, or the usual addition in the continuous case. The problem is made interesting by imposing input constraints upon \(X_1\) and \(X_2\), thus the encoders cannot simply subtract the interferences.

Similarly to the Korner-Marton problem, a linear/lattice pre-coding scheme (which subtracts the interference “modulo the code”) achieves the capacity region of this channel [10], [2].1 And in contrast, the rates achieved by a more conventional random binning scheme vanish in the limit of strong interference signals.

This sharp discrepancy is due to the distributive nature of the side-information; if the knowledge of \(S_1\) and \(S_2\) were centralized - i.e., they were both known to one encoder or to the joint decoder, then random binning could be effective and (nearly) achieve capacity; see [2].

Sometimes structured codes are inferior to random codes. This situation occurs when the linear structure of the code causes ambiguity at the decoder; for example, the symmetric-rates point of the (clean) MAC capacity region, or of the Slepian-Wolf rate region.2

In this short note we focus on another, perhaps obvious weakness of structured codes; they are sensitive to the structure of the channel. Specifically, we show that if the additive channel in (1) or in (2) is replaced by one involving both addition and multiplication, then structured codes - and in fact, any other coding scheme - are not effective.

II. SUM-PRODUCT KORNER-MARTON

Consider a generalization of the KM problem (1), where the statistical relation between the component sources is given by

1In the discrete noiseless case the linear coding scheme is exactly optimal, while in the continuous case it is asymptotically optimal in the continuous high SNR case.

2This problem can be resolved by using two different (linearly independent) linear codes [11].
\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Fig. 1. Functional source coding with side information at the decoder.}
\end{figure}

In the \textit{centralized state} case, the equivalent channel state \( B = B' \times B'' \) is either known to the joint decoder (i.e., one encoder observes \( A \), the other encoder observes \( Y \), and the decoder has access to \( B \)) or to both encoders (i.e., one encoder observes \( A, B \), and the other encoder observes \( Y, B \)). It is not hard to show that in this case the compression rate is \( H(Z) \) per each encoder, independent of the state distribution, as in the classical KM problem.\(^3\)

Our focus is, however, on the \textit{de-centralized state} case, where each of the channel states \( B' \) and \( B'' \) is available to a different encoder. The performance in this case can be bounded by the simplified setup shown in Fig. 1, where one of the helper sources is available (un-coded) as “side information” at the decoder. That is, there is only one encoder which observes \( X = (A, B) \), while

\[ Y = A + B \times C \]

is available at the decoder, who wishes to reconstruct \( Z = C \).

This latter problem is a special case of \textit{functional source coding} \cite{Kobayashi}, where a function \( Z = F(X, Y) \) needs to be reconstructed from separate coded versions of \( X \) and \( Y \). The setup of Fig. 1 corresponds to the case where \( Y \) is encoded as “side information” to the decoder (or it is encoded at a rate greater than or equal to \( H(Y) \)), and where

\[ X = (A, B), \quad Y = A + B \times C, \quad Z = C \]

and \( F(X, Y) = (Y - A)/B \).

A precise definition of functional source coding with side information at the decoder is as follows. The encoding function is \( f : F_q^n \rightarrow \mathcal{M} \), where the size of the message space \( \mathcal{M} \) is \( 2^{nR} \), with \( n \) being the code block length and \( R \) being the coding rate. The decoding function is \( g : \mathcal{M} \times F_q^n \rightarrow F_q^n \). The probability of error \( P_e \) is the probability that \( g(M, Y^n) \) is not equal to the vector \( Z^n \), where \( Z_i = F(X_i, Y_i) \), \( i = 1 \ldots n \), and where \( M = f(X^n) \) is the encoded message. For a given memoryless double source \( (X_1, Y_1, X_2, Y_2, \ldots) \) and a function \( F(\ldots) \), a rate \( R \) is said to be “achievable” if we can make \( P_e \) as small as desired for some functions \( f \) and \( g \), and large enough \( n \). Finally, \( R^* \) denotes the minimum achievable rate.

Clearly, the minimum achievable rate \( R^* \) satisfies

\[ H(X|Y) \geq R^* \geq H(F(X,Y)|Y) \] \hspace{1cm} (5)

where the LHS corresponds to the case where the decoder fully reconstructs \( X \) before computing \( F(X,Y) \), while the RHS corresponds to the case where the encoder also has access to \( Y \), so it can first compute \( F(X,Y) \) and then compress it. In the sum-product case (4), if \( A \) is uniform over \( F_q \), then the bounds (5) become

\[ H(B) + H(C) \geq R^* \geq H(C) \] \hspace{1cm} (6)

Note that in the classical KM problem \( B = 1 \), i.e., \( H(B) = 0 \); thus the bounds coincide, and the coding rate is merely the entropy of the desired variable \( C \).

Han and Kobayashi \cite{Kobayashi} give necessary and sufficient conditions for the LHS of (5) to be tight.\(^4\) These conditions are satisfied in the sum-product case.

\textbf{Lemma 1.} In the sum-product (functional source coding) problem (4), \( R^* = H(X|Y) \). Thus, if \( A \) is uniform over \( F_q \), then \( R^* = H(B) + H(C) \).

\textit{Proof:} Follows since two different lines in \( F_q \) intersect in at most one point, implying the condition in \cite[lem.1]{Kobayashi}.

This result implies that the minimum coding rate \( R^* \) is in general larger than the entropy of the desired variable \( C \), which is the rate in the classical KM setting (1). In fact, the “extra” rate can be as large as \( \log(q - 1) \), for \( B \) which is uniform over \( F_q \setminus \{0\} \).

As a corollary from Lemma 1, it follows that for uniform channel states \( B' \) and \( B'' \) in the distributed coding setup of (3), the rate of each encoder is at least \( \log(q - 1) + H(C) \). The interpretation is that the introduction of the multiplicative state variables breaks the structure of the classical KM problem; the states \( B' \) and \( B'' \) must be fully conveyed to the decoder before the linear structure of the channel can be utilized (by means of a linear “syndrome” coding) to encode the desired source \( C \).

\section{The Sum-Product Dirty MAC}

Consider next a modification of the dirty MAC problem (2), in which the channel output is given by

\[ Y = A' + A'' + B \times C \] \hspace{1cm} (7)

\(^3\)Regarding the former case \((B \text{ available at the decoder}), \text{note that a random parity-check matrix } \mathbf{H} \text{ is “good” with high probability for the classical KM problem (1) (i.e., } Z^n \text{ can be reliably decoded from the syndromes } HX^n \text{ and } HY^n)\); hence, \( H(B') \) is “good” with high probability for the generalized KM problem (3). In the latter case, the encoders simply divide their observations by \( B \), hence get back to the classical KM problem.

\(^4\)They in fact consider a more general case where both \( X \) and \( Y \) are encoded.
where as in (3) all variables belong to a finite field $F_q$ of size $q$, and the $+$ and $\times$ are the sum and product operations over $F_q$. The inputs of this MAC are $A'$ and $A''$ (corresponding to $X_1$ and $X_2$ in (2)), while $B$ and $C$ are the channel state variables (corresponding to $S_1$ and $S_2$ in (2)). There is no additional noise, nor input constraints.

As in the sum-product KM, the centralized state case is easy: if both state variables $B$ and $C$ are known to one encoder, or are known to the joint decoder, then the product $B \times C$ can be simply subtracted; hence the sum capacity is $\log(q)$, as if the channel was noiseless.

The interesting setup is, again, the de-centralized state case. That is, each encoder has access to only one of the channel states, while the decoder is completely ignorant of the states. The capacity in this case is bounded from above by that of the single-user channel shown in Fig. 2, with

$$Y = A + B \times C. \quad (8)$$

Here there is a single encoder that has access to one of the states ($S_1 = B$), while the decoder has access to the second state ($S_2 = C$), where $C$ is independent of both the input $X = A$ and $B$.

A precise definition of encoding and decoding over such a channel is as follows. The encoding function is $f : M \times F_q^n \rightarrow F_q^n$ and the decoding function is $g : F_q^n \times F_q^n \rightarrow M$, where the size of the message space $M$ is $2^nR$, $R$ being the coding rate. The error probability $P_e$ is the probability that $g(Y^n, S^n_2) \neq M$, where $Y^n$ depends on $X^n, S^n_1$ and $S^n_2$, and where $X^n = f(M, S^n_2)$ for $M \in M$. A rate $R$ is said to be “achievable” if we can make $P_e$ as small as desired for some functions $f$ and $g$, and large enough $n$. Finally, the capacity $C$ is the highest achievable rate.

The sum-product channel (8) is, in fact, a deterministic channel, where the output $Y$ is a function of the input $X$, and the two states $S_1$ and $S_2$:

$$Y = F(X, S_1, S_2). \quad (9)$$

There is no additional noise in the channel, beyond the randomness of the two (known) states $S_1$ and $S_2$.

The setup of (9) is an instance of the Gelfand-Pinsker problem [9], i.e., a channel with non-causal side information at the encoder. Hence, it has a single letter solution of the form

$$C = \max I(U; Y|S_2) - I(U; S_1) \quad (10)$$

where the maximization is over a suitable set of auxiliary random variables $U$, and functions $X = X(U, S_1)$.\footnote{A continuous version of this setup may be thought of as a channel with a fading interference [12].}

The structure of the function $F$ in (9) plays a key role in determining the capacity $C$. A favorable case is when $F$ has a composite form, where the dependence on the encoder variables $(X, S_1)$ is separate from the decoder state $S_2$.

**Lemma 2.** If the function $F : F_q \times F_q \times F_q \rightarrow F_q$ can be decomposed into $F(a, b, c) = \tilde{F}(G(a, b), c)$, where $\tilde{F}$ is invertible with respect to the first argument (i.e., the equation $y = \tilde{F}(t, c)$ has a solution $t$ for every $y$ and $c$), then (10) is optimized by $U = G(X, S_1)$. If also $G$ is invertible with respect to the first argument, then the capacity is

$$C = \log(q)$$

and it is achieved by an input $p(x|s_1)$ that makes $G(X, s_1)$ uniform over $F_q$ for all values of $s_1$.

*Proof:* The first part follows from [13, sec. III.F], and the invertibility of $\tilde{F}$. See also [14].

The sum-product channel (8) clearly does not satisfy the first condition of the lemma, as addition and multiplication do not commute. In fact, this channel is much worse. To assess its capacity, we shall first establish a relation to a “minimum entropy” problem.

**Lemma 3.** The capacity of a deterministic two-state channel of the form

$$Y = X + F(S_1, S_2) \quad (11)$$

where $+$ denotes addition in $F_q$ and where $S_1$ and $S_2$ are available at the encoder and the decoder, respectively, is given by

$$C = \log(q) - \inf_1 \frac{1}{n} H(q(S^n_1) + F(S^n_1, S^n_2)|S^n_2) \quad (12)$$

where the second term is the average conditional entropy given $S^n_2$, and where the infimum is over all code block lengths $n$, and functions $g : F_q^n \rightarrow F_q^n$.

*Proof:* Easy and will be omitted.

Although (12) is not a single-letter expression, it is sometimes easier for analysis than the Gelfand-Pinsker solution (10). Specifically, for the sum-product channel (8) we have:

**Theorem 1.** (Shany-Zamir [15]) For the case where $Y = X + S_1 \times S_2$, the minimum average conditional entropy in (12) is bounded by

$$\log \left( \frac{q}{2} \right) \leq H_{\min} \leq \log \left( \frac{q}{2-1/q} \right).$$

The upper bound is achieved (for all $n$) by a quadratic per-letter function $g(s) = s^2$.

As a corollary from this theorem, we conclude that the capacity of the sum-product channel (8) is at most one bit.
for all $q$. This is quite disappointing when compared to the
capacity of $\log(q)$, which is achievable in the centralized-state
case. The same statement is true also for the sum-product dirty
MAC (7); that is, the rate of each user is at most one bit.7

IV. DISCUSSION

The essence of the examples given in this paper is that the
order of performing the sum and product operations matters;
and in fact, they are very “different”. One aspect of this
difference - which is related to the sum-product KM problem -
is that the expression

$$a + b \times c$$ \hspace{1cm} (13)

cannot be decomposed into the form: function function $(a, b), c)$ (not even approximately). A second aspect - related to the sum-product dirty MAC - is
that it is impossible to find a function of $a = a(b)$ such that
(13) would be only a function of $c$ (not even approximately).
In contrast, these two requirements are easily fulfilled if
the expression in (13) is a pure sum $a + b + c$ (by the
associativity of summation), or a pure product $a \times b \times c$ (by the
associativity of multiplication).

It would be interesting to explore further (and perhaps quantify)
the information-theoretic aspects of function decom-
position. Note that this question is almost “distribution free”
(i.e., nearly independent of the probability distributions of
sources and channels). A different aspect of “anti structure”,
which is due to a “bad” noise distribution, can be found in
[16].

ACKNOWLEDGMENT

I thank Yaron Shany for many interesting discussions about
the sum-product channel. Thanks are also due to Aaron
Wagner for an interesting discussion about the sum-product
KM problems, and to Prakash Iswar and Sandeep Pradhan for
pointing out the relation to the work of Han and Kobayashi.

REFERENCES

correlated sources $(X,Y)$ from the viewpoint of the achievable rate
1987.
Jointly Gaussian sources and reconstruction of a linear function,” IEEE
The dirty multiple access channel,” IEEE Trans. Information Theory,
using distributed structure,” in Proc. of ISIT, Aug. 2011, St. Petersburg,
pp. 1003–1007.
with fading dirt,” in Proc. of the Electrical and Electronics Engineers
in Israel (IEEEI), Nov. 2010, Eilat, pp. 000 525–000 529.
embedding and source coding with side information, and some
2003.
thesis, Tel Aviv University, Nov. 2007.
function determined up to a diagonal linear map on $\mathbb{F}_q^n$,” Arxiv, vol.
for writing on dirty paper,” IEEE Trans. Information Theory, vol. IT-54,

7Perhaps even the stronger statement - that the sum rate is at most one bit
is true.
8Note that pure multiplicative versions of the generalized KM and DMAC
problems can be solved using linear codes over a “logarithmic” domain.
Real-Time Coding of Markov Sources over Erasure Channels: When is Binning Optimal?

Ashish Khisti and Farrokh Etezadi
University of Toronto
10 King’s College Road
Toronto, ON, Canada
Email: {akhisti, fetezadi}@comm.utoronto.ca

Mitchell Trott
HP Labs
1501 Page Mill Road
Palo Alto, California
Email: mitchell.trott@hp.com

Abstract—We study sequential coding of a Markov source process under error propagation constraints. The channel can erase up to B packets in a single burst, but reveals all other packets to the destination. The destination is required to reproduce all the (vector) source sequences sequentially, except those that occur in a window of length $B + W$ following the start of the erasure burst. Our earlier work establishes upper and lower bounds on the compression rate as a function of $B$ and $W$. In this work we show that for the class of symmetric sources, if we restrict to a memoryless encoding function, then a binning based scheme is optimal. Our converse involves a drawing connection between the sequential coding problem and a multiuser source coding problem called Zig-Zag source coding with side information.

I. INTRODUCTION

A tradeoff between compression efficiency and error resilience is fundamental to any video compression system. In live video coding, an encoder observes a sequence of correlated video frames and produces a compressed bit-stream that is transmitted to the destination. If the underlying channel is an ideal bit-pipe, it is well known that predictive coding [1] achieves the optimum compression rate. Unfortunately in many emerging video distribution networks, such as peer-to-peer systems and mobile systems, packet losses are unavoidable. Predictive coding is highly sensitive to such packet losses and can lead to a significant amount of error propagation. Various techniques are used to practice prevent such losses. Commonly used video coding techniques use a group of picture (GOP) architecture, where intra-frames are periodically inserted to limit the effect of error propagation. Forward error correction codes can also be applied to compressed bit-streams to recover from missing packets [2], [3]. Modifications to predictive coding, such as leaky-DPCM [4], [5], have been proposed in the literature to deal with packet losses. The robustness of distributed video coding techniques in presence of packet losses has been studied in e.g., [6], [7].

Information theoretic analysis of video coding has received significant attention in recent times, see e.g., [8]–[10] and the references therein. These works focus primarily on the source coding aspects of video. The source process is a sequence of vectors, each of which is spatially i.i.d. and temporally correlated. Each source vector is sequentially compressed into a bit stream. The destination is required to recover the source vectors in a sequential manner as well. However all of these works assume an ideal channel with no packet losses. To our knowledge even the effect of a single isolated packet loss is not fully understood [11].

In an earlier work [12] we build upon [8], [9] and introduce an information theoretic framework to characterize the tradeoff between error propagation and compression rate. An encoder is revealed source vectors in a sequential manner and compresses them sequentially into channel packets that are then transmitted over a channel. An information theoretic notion of error propagation is defined and upper and lower bounds are obtained on the compression rate. The lower bound is based on a careful analysis of information flow during the decoding process whereas the upper bound is based on a binning technique. For a special class of sources, a new technique, prospicient coding is proposed, and shown to be optimal. However the optimal compression rate remains an open problem for a large class of sources including the binary symmetric Markov source.

In this paper we consider a class of symmetric Markov sources. For this class the minimum compression rate has not yet been characterized. Our main result in this paper is that if we restrict to the class of memoryless encoders, then the binning based scheme proposed in [12] is optimal. The converse is established by drawing a connection to a multiuser source coding problem called zig-zag source coding network [13]–[15].

II. PROBLEM STATEMENT

A. Source Model

We consider a semi-infinite stationary vector source process \( \{ s_t \}_{t \geq 0} \) whose symbols (defined over some finite alphabet \( \mathcal{S} \)) are drawn independently across the spatial dimension and from a first-order Markov chain across the temporal dimension, i.e., for each \( t \geq 1 \),

\[
\Pr( s_t = s_0 \mid s_{t-1} = s_{t-1}' ) = \prod_{j=1}^{n} p_{s_j|s_j'}(s_{j} | s_{j-1}'), \quad \forall t \geq 1.
\]

(1)

We assume that the prior distribution \( p_{s_0}(\cdot) \) and \( p_{s_{j+1}|s_j}(\cdot|\cdot) \) are selected such that the underlying random variables \( \{ s_t \}_{t \geq 0} \) constitute a time-invariant and a first-order stationary Markov
chain. Of particular interest in this paper is the class of symmetric sources where the underlying Markov chain is also reversible i.e., the random variables satisfy \((s_0, \ldots, s_t) \rightarrow (s_t, \ldots, s_0)\), where the equality is in the sense of distribution [16]. Of particular interest to us is the following property satisfied for each \(t \geq 1\):

\[
p_{s_{t+1}, s_t}(s_u, s_v) = p_{s_{t+1}, s_t}(s_v, s_u), \quad \forall s_u, s_v \in S
\]  

(2)

i.e., we can “exchange” the source pair \((s_{t+1}, s_t)\) with \((s_t, s_{t+1})\) without affecting the joint distribution. An important class of sources that are symmetric are the binary sources: \(s_i^0 = s_i^0 \oplus s_i^0\), where \(\{z_i\}_{i=1}^B\) is an i.i.d. binary source process (in both temporal and spatial dimensions) with the marginal distribution \(Pr(z_i = 0) = p\), the marginal distribution \(Pr(s_i = 0) = Pr(s_i = 1) = \frac{1}{2}\) and \(\oplus\) denotes modulo-2 addition.

**B. Rate-Recovery Function**

A rate-\(R\) causal encoder maps the sequence \(\{s_{t}^j\}_{j \geq 0}\) to an index \(f_i \in [1, 2^B]\) according to some function

\[
f_i = f_j(s_0^i, \ldots, s_B^i)
\]  

(3)

for each \(i \geq 0\). A memoryless encoder satisfies \(f_j(s_0^i, \ldots, s_B^i) = f_j(s_B^i)\) i.e., the encoder does not use the knowledge of the past sequences.

The channel introduces an erasure burst of size \(B\), i.e. for some particular \(j \geq 0\), it introduces an erasure burst such that \(g_i = 1\) for \(i \notin \{j, j+1, \ldots, j+B-1\}\) and \(g_i = f_i\) otherwise. Upon observing the sequence \(\{g_j\}_{j \geq 0}\) the decoder is required to perfectly recover all the source sequences using decoding functions

\[
s_j^e = g_j(g_0, g_1, \ldots, g_i), \quad i \notin \{j, j+B+W-1\}.
\]  

(4)

It is however not required to produce the source sequences in the window of length \(B+W\) following the start of an erasure burst. We call this period the error propagation window. The setup is shown in Fig. 1.

A rate \(R(B, W)\) is feasible if there exists a sequence of encoding and decoding functions and a sequence \(e_n\) that approaches zero as \(n \rightarrow \infty\) such that, \(Pr(s_i^j \neq \hat{s}_n) \leq e_n\) for all \(i \notin \{j, \ldots, j+B-1\}\). We seek the minimum feasible rate \(R(B, W)\) which we define to be the rate-recovery function. The following upper and lower bounds have been established in [12].

**Theorem 1 ([12]):** For any stationary first-order Markov source process the rate-recovery function satisfies

\[
R^-(B, W) \leq R(B, W) \leq R^+(B, W)
\]  

where

\[
R^-(B, W) = H(s_1|s_0) + \frac{1}{W+1}H(s_W; s_{W+B+1}|s_0).
\]  

(5)

\[
R^+(B, W) = H(s_1|s_0) + \frac{1}{W+1}H(s_B; s_{B+1}|s_0)
\]  

(6)

Notice that the upper and lower bound coincide for \(W = 0\) and \(W \rightarrow \infty\), yielding the rate-recovery function in these cases. The upper bound is obtained via a memoryless binning based scheme. At each time the encoding function \(f_i\) in (3) is obtained as the bin-index of an independent Slepian-Wolf codebook [17]. The rate expression for \(R^+(B, W)\), which is equivalent to [12]

\[
R^+(B, W) = \frac{1}{W+1}H(s_{W+1}, s_{W+2}, \ldots, s_{W+B+1}|s_0)
\]  

(7)

guarantees that the decoder can recover \(s_{j+B+W}\) following an erasure burst between \([j, j+B-1]\) using the \(W+1\) bin indices \(f_{j+B}, f_{j+B+1}\) and the source sequence \(s_{j+B}^e\) before the erasure.

In [12] some counter-examples are provided where the lower bound \(R^-(B, W)\) (c.f. (5)) is tight and binning based upper bound \(R^+(B, W)\) (c.f. (6)) is not optimal in general. Nevertheless such examples require a special structure and do not include many natural source models such as the binary symmetric sources. The optimal rate-recovery function for the class of symmetric sources remains open. In this paper we establish the optimality of binning based scheme if one restricts to the class of memoryless encoders.

**Theorem 2:** For the class of symmetric sources that satisfy (2) the rate-recovery function, restricted to the class of
memoryless encoders, is given by
\[ R(B, W) = \frac{1}{W+1} H(s_{B+1}, s_{B+2}, \ldots, s_{B+W+1}|s_0). \] (8)

Note that the achievability follows immediately from (7). Thus it only remains to show that the lower bound (5) needs to be improved. We have only been able to obtain this improvement for the class of memoryless encoders. For the general encoder structure (3) this remains an open problem. At first glance one may expect that the binning based scheme is always optimal for the class of memoryless encoders. This is however not true. Interestingly the *prosipient encoders* in [12] that improve upon the binning based lower bound are also memoryless. Our proof involves an interesting connection a multi-terminal source coding problem called zig-zag source coding [13]–[15]. In particular we develop a simple approach to lower bound the sum-rate of a zig-zag source coding network with symmetric sources that may be of independent interest.

### III. PROOF OF THEOREM 2

The special case when \( W = 0 \) follows directly from (5). We only need to consider the case when \( W \geq 1 \). For simplicity in exposition we consider the case when \( W = 1 \). Then we need to show that
\[ R(B, W = 2) \geq \frac{1}{2} H(s_{B+1}, s_{B+2}|s_0). \] (9)

The proof for general \( W \geq 1 \) follows along similar lines and will be sketched briefly.

Assume that an erasure-burst spans time indices \( j-B, \ldots, j-1 \). The decoder must recover
\[ s_j^0 = \hat{g}_{j+1}(\epsilon_{j-B}^{j-1}, f_j, f_{j+1}). \] (10)

From Fano’s inequality, we have,
\[ H(s_j^0|f_j) \leq n\varepsilon. \] (11)

Furthermore if there is no erasure until time \( j \) then
\[ s_j^0 = \hat{g}_j(\epsilon_0^j) \] (12)

must hold. Hence from Fano’s Inequality,
\[ H(s_j^0|f_j) \leq n\varepsilon. \] (13)

Our aim is to combine (11) and (13) to establish the following lower bound on the sum-rate
\[ R_j + R_{j+1} \geq H(s_{j+1}|s_j) + H(s_j|s_{j-B-1}). \] (14)

The lower bound then follows since
\[ R \geq \max(R_j, R_{j+1}) \geq \frac{1}{2}(R_j + R_{j+1}) \geq \frac{1}{2}(H(s_{j+1}|s_j) + H(s_j|s_{j-B-1})) \]
\[ \geq \frac{1}{2}(H(s_{j+1}|s_j) + H(s_j|s_{j-B-1})). \] (15)

Thus establishing (9).

To establish (14) we make a connection to a multi-terminal source coding problem in Fig. 2.

### A. Zig-Zag Source Coding

Consider the source coding problem with side information illustrated in Fig. 2(a). In this setup there are four source sequences drawn i.i.d. from a joint distribution \( p(s_j, s_{j-1}, s_{j-B-1}) \). The two encoders \( j \) and \( j+1 \) are revealed source sequences \( s_j^0 \) and \( s_{j+1}^0 \) and the two decoders \( j \) and \( j+1 \) are revealed sources \( s_j^0 \) and \( s_{j+1}^0 \). The encoders operate independently and compress the source sequences to \( f_j \) and \( f_{j+1} \) at rates \( R_j \) and \( R_{j+1} \) respectively. Decoder \( j \) has access to \( (f_j, s_j^0) \) while decoder \( j+1 \) has access to \( (f_{j+1}, s_{j+1}^0) \) and are interested in reproducing,
\[ s_j^0 = \hat{g}_j(f_j, s_{j-1}^0) \]
\[ s_{j+1}^0 = \hat{g}_{j+1}(f_{j+1}^{j+1}, s_{j+1-B}^0) \] (20)

respectively such that \( Pr(s_j^0 \neq s_{j+1}^0) \leq \varepsilon_i \) for \( i = j, j+1 \).

When \( s_{j+B}^0 \) is a constant sequence, the problem has been studied in [13], [15]. A complete single letter characterization involving an auxiliary random variable is obtained. Fortunately in the present case of symmetric sources a simple lower bond can be obtained using the following observation.

**Lemma 1:** The set of all achievable rate-pairs \( (R_j, R_{j+1}) \) for the problem in Fig. 2(a) is identical to the set of all achievable rate-pairs for the problem in Fig. 2(b) where the side information sequence \( s_{j}^0 \) at decoder 1 is replaced by the side information sequence \( s_{j+1}^0 \).

The proof of Lemma 1 follows by observing that the capacity region for the problem in Fig. 2(a) depends on the joint distribution \( p(s_j, s_{j-1}, s_{j-B-1}) \) only via the marginal distributions \( p(s_j, s_{j-1}) \) and \( p(s_j, s_{j-B-1}) \). When the source is symmetric the distributions \( p(s_j, s_{j-1}) \) and \( p(s_j, s_{j-B-1}) \) are identical. The formal proof will be omitted.

This suffices to lower bound the achievable sum rate for the problem in Fig. 2(b). First upon applying the Slepian-Wolf lower bound to encoder \( j+1 \)
\[ nR_{j+1} \geq H(s_{j+1}|s_j^0, s_{j-B-1}, f_j) - n\varepsilon \] (22)

and to bound \( R_j \)
\[ nR_j \geq H(f_j) = I(f_j; s_j^0|s_{j-B-1}) \]
\[ \geq H(s_j^0|s_{j-B-1}) - H(s_j^0|s_{j-B-1}, f_j) \]
\[ \geq nH(s_{j+1}|s_j^0) - H(s_j^0|s_{j-B-1}, f_j) \]
\[ = H(s_j^0|s_{j-B-1}) - I(s_j^0, s_{j+1}^0, s_{j-B-1}, f_j) - n\varepsilon \] (23)

where (23) follows by applying Fano’s inequality to since \( s_j^0 \) can be recovered from \( (s_{j+1}^0, f_j) \) and hence
\[ H(s_j^0|s_{j-B-1}, s_{j+1}^0, f_j) \leq n\varepsilon \] holds and (24) follows from the Markov relation \( s_j^0 \rightarrow f_j \rightarrow (f_{j+1}, s_{j+1-B}^0) \). Observe that (14) follows by summing (22) and (24).

97
B. Connection between Streaming and Zig-Zag Coding Problems

It remains to show that the lower bound on the Zig-Zag coding problem also constitutes a lower bound on the original problem.

Lemma 2: Suppose that the encoding function \( f_j = \mathcal{E}_j(s^n_j) \) is memoryless. Suppose that there exist decoding functions \( \hat{s}^n_j = \mathcal{D}_j(f^n_j) \) and \( \hat{s}_{j+1}^n = \mathcal{D}_j(f^n_j) \) such that \( \Pr(s^n_j \neq \hat{s}^n_j) \) and \( \Pr(s^n_j \neq \hat{s}^n_{j+1}) \) both vanish to zero as \( n \to \infty \). Then

\[
H(s^n_j|s^n_{j-1}, f_j) \leq n \epsilon_n
\]

and

\[
H(s^n_{j+1}|s^n_{j-B-1}, f_j, f_{j+1}) \leq n \epsilon_n
\]

also hold.

We omit the proof due to space constraint. The conditions in (25) and (26) show that any rate that is achievable in the original problem is also achieved in the Zig-Zag source network. Hence a lower bound to the source network also constitutes a lower bound to the original problem.

C. Extension to Arbitrary \( W > 1 \)

Finally we comment on the extension of the above approach to \( W = 2 \). We now consider three encoders \( t \in \{ j, j+1, j+2 \} \). Encoder \( t \) observes a source sources \( s^n_t \) and compresses it into an index \( f_t \in \{ 1, 2^{a_t} \} \). The corresponding decoders are revealed \( s^n_{t-B-1} \) for \( t \in \{ j, j+1 \} \) and the decoder \( j+2 \) is revealed \( s^n_{j-B-1} \). An argument analogous to Lemma 1 the rate region is equivalent to the case when decoders \( j+1 \) and \( j+2 \) are instead revealed \( s^n_{j+1} \) and \( s^n_{j+2} \) respectively. For this new setup it is easy to show that decoder \( j+2 \) must reconstruct \( (s^n_j, s^n_{j+1}, s^n_{j+2}) \) given \( (s^n_{j+1}, s^n_{j+2}, f_{j+1}, f_{j+2}) \). The sum rate must therefore satisfy \( R_j + R_{j+1} + R_{j+2} \geq 2H(s_j, s_{j+1}, s_{j+2}|s^n_{j-B-1}) \). Using an extension of Lemma 2 we can show that the proposed lower bound also continues to hold for the original streaming problem. This completes the proof. The extension to any arbitrary \( W > 1 \) is completely analogous.
SuperNyquist Rateless Coding for Intersymbol Interference Channels

Uri Erez
Tel Aviv University, Ramat Aviv, Israel
Email: uri@eng.tau.ac.il

Gregory W. Wornell
Dept. EECS, MIT, Cambridge, MA
Email: gww@mit.edu

I. INTRODUCTION

In traditional digital communication, achieving high throughput when the channel state allows is accomplished by selecting high-order signal constellations. However, an alternative approach, originally proposed several decades ago [1], exploits super-Nyquist (SNQ) signaling. In SNQ signaling, the symbols are taken from a fixed constellation, typically BPSK or QPSK, independent of the transmission rate. Higher rates are achieved by increasing the signaling rate—i.e., the rate at which the symbols are modulated onto the bandlimited pulse shape—beyond the Nyquist rate. Thus, in SNQ systems, the signaling rate is decoupled from the transmission bandwidth, and can greatly exceed the transmission bandwidth.

Because SNQ modulation introduces ISI, it necessitates the use of equalization, which traditionally made it unappealing for early applications; see, e.g., [2]. In this paper, however, we establish that SNQ signaling has some particularly valuable properties for communication over Gaussian intersymbol interference (ISI) channels where the transmitter knows neither the channel impulse response nor the maximal rate that my be supported by the channel. In particular, we establish the somewhat surprising result that the use of SNQ signaling allows for highly efficient joint design of the physical and link layers. Indeed, from such signaling we develop a rich family of low-complexity, capacity-approaching rateless codes for scalar ISI channels, which have natural extensions to vector ones.

II. SYSTEM AND CHANNEL MODEL

We consider a linear dispersive Gaussian channel for which the complex baseband channel output takes the form

\[ y(t) = h(t) * x(t) + z(t), \]

where \( z(t) \) is additive white Gaussian noise (AWGN) with one-sided power spectral density \( N_0 \), and where \( x(t) \) is the input, which is subject to a power constraint \( \mathbb{E}\{|z(t)|^2\} \leq P \) and bandwidth constraint \( W \). The associated white-input capacity of the channel is

\[ C_{[b/s]} = \int_{-W/2}^{W/2} \log \left( 1 + \frac{P|H(f)|^2}{N_0W} \right) \, df. \]  

We consider pulse-amplitude modulation, whereby

\[ x(t) = \sum_n s[n] g(t - nT), \]

where \( T \) is the symbol duration and \( T_0 = 1/W \) is the Nyquist sampling time. The associated “over-signaling” ratio is thus \( L = T_0/T \). After matched filtering and sampling at the symbol rate, the equivalent discrete-time channel is

\[ y[n] = k[n] * s[n] + z[n], \]

where \( k[n] = k(nT) \), where \( k(t) = h^*(−t) * h(t) * g(−t) * g(t) \), and where

\[ S_{xx}(e^{j2\pi f}) = \frac{N_0}{2} K(e^{j2\pi f}) \]

\[ = \frac{N_0}{2T} \sum_i |H(f/T + i/T)|^2 |G(f/T + i/T)|^2. \]

The pulse shape is required to be limited to system bandwidth \( W \), i.e., \( g(t) \) satisfies \( G(f) = 0 \) for \( |f| > W/2 \). To simplify our development, we largely restrict our attention to the case

\[ g(t) = \sin(t/T_0), \text{ with } \sin(u) \triangleq \sin(\pi u)/\pi u. \]

Taking the symbols \( s[n] \) to be independent identically-distributed (i.i.d.) circularly symmetric complex Gaussian with power \( P/L \) results in a (proper) Gaussian random input signal.
\( x(t) \) with power \( P \). It follows that the capacity of the discrete-time channel (3) is

\[
C_{[b/\text{SNQ symbol}]} = \int_{-1/2}^{1/2} \log \left( 1 + \frac{P}{N_0} |H(f)|^2 \right) df
\]

\[
= t \int_{-1/2T}^{1/2T} \log \left( 1 + \frac{P}{T_0N_0} \sum_{i} |H(f+i/T)|^2 |G(f+i/T)|^2 \right) df
\]

\[
= t \int_{-1/2T}^{1/2T} \log \left( 1 + \frac{|H(f)|^2 |G(f)|^2}{T_0N_0} \right) df,
\tag{5}
\]

where the last equality follows from the fact that \( g(t) \) is bandlimited.

Note that for sinc modulation (4), \( x(t) \) has a flat power spectrum over the bandwidth \( W \), and the modulation achieves the white-input capacity (1) for any \( L \), i.e., (5) specializes to

\[
C_{[b/\text{SNQ symbol}]} = \frac{1}{LW} \int_{-W/2}^{W/2} \log \left( 1 + \frac{P |H(f)|^2}{N_0W} \right) df
\]

\[
= \frac{T}{T_0} \frac{1}{W} C_{[b/\tilde{q}]}.
\tag{6}
\]

### III. LINEAR SNQ RATELESS CODING

Consider now packetized transmission where the packet size is large but otherwise plays no role in the analysis. We consider a simplified model where the channel response experienced throughout transmission of the \( m \)th packet, \( m = 1, \ldots, M \), is linear time-invariant (LTI) but the impulse response, which we denote by \( h_m(t) \), may vary from packet to packet. The channel input-output relation for the transmission of the \( m \)th packet is therefore

\[
y_m[n] = s_m[n] * h_m[n] + z_m[n],
\tag{7}
\]

where \( k_m[n] = k_m[nT] \) and \( h_m(t) = h_m(-t) * h_m(t) + g^*(-t) + g(t) \). Assuming discrete-time white-input transmission for all packets, it follows from (6), that the mutual information (in b/\text{SNQ symbol}) corresponding to each packet is

\[
C_{m[0-\text{SNQ symbol}]} = \int_{-1/2}^{1/2} \log \left( 1 + \frac{PK_m(e^{j2\pi f})}{N_0L} \right) df
\]

\[
= \frac{1}{LW} \int_{-W/2}^{W/2} \log \left( 1 + \frac{P|H_m(f)|^2}{N_0W} \right) df,
\]

where the second equality holds for ideal sinc modulation. Upon receiving a set \( \mathcal{S} \subseteq \{1, \ldots, M\} \) of packets, the aggregate mutual information is thus

\[
C(\mathcal{S}) = \sum_{m \in \mathcal{S}} C_m.
\tag{8}
\]

Our aim is to design a low complexity coding and modulation scheme that (simultaneously) approaches \( C(\mathcal{S}) \) for all sets \( \mathcal{S} \subseteq \{1, \ldots, M\} \) without requiring the transmitter to have knowledge of the capacities \( C_m \). Rather, for any given chosen target rate \( R \), and no knowledge of the channel, transmission should be successful whenever \( C(\mathcal{S}) > R \) holds for the received set of packets \( \mathcal{S} \).

We proceed to describe the proposed linear rateless SNQ construction. All the signals \( s_m[n] \) are obtained from a single coded stream \( s[n] \) according to

\[
s_m[n] = v_m[n] s[n],
\tag{9}
\]

where \( v_m[n] \) are sequences to be specified. The transmitted signal corresponding to packet \( m \) is thus

\[
x_m(t) = \sum_{i} s[n] v_m[n] g(t - iT).
\]

Provided we choose the sequences \( v_m[n] \) so that the transmitted signals \( x_m(t) \) are statistically independent and the \( \tilde{x}_m[n] \) are white circularly-symmetric complex Gaussian processes, the mutual information corresponding to each packet remains \( C_m \) and furthermore the aggregate mutual information from the receipt of multiple packets is the sum of the individual ones.

A simple means to achieve this is by taking \( v_m[n] = e^{-j2\pi mn/L} \). Accordingly, we define

\[
x_m[n] \triangleq x_m(nT) = \sum_{i} s[n] e^{-j2\pi in/L} g[n - i],
\tag{10}
\]

where \( g[n] = g(nT) = \text{sinc}(n/L) \), with \( g(t) \) as in (4). Using (10), we see that the transmit signals “shifted-back in frequency” in this case are

\[
\tilde{x}_m[n] \triangleq e^{j2\pi mn/L} x_m[n] = \sum_{i} s[n] g[n - i],
\]

where \( g_m[n] = e^{j2\pi mn/L} g[n] \).

Clearly, requiring that the signals \( \{x_m(t)\} \) be mutually independent is equivalent to requiring that the associated discrete-time signals \( \{\tilde{x}_m[n]\} \) be. Furthermore, the latter holds if and only if \( \{\tilde{x}_m[n]\} \) are mutually independent. Therefore, it suffices to verify the last condition. Since the signals \( \tilde{x}_m[n] \) are jointly Gaussian and stationary, they are independent if their cross-spectra

\[
S_{\tilde{x}_m, \tilde{x}_m}(e^{j2\pi f}) = S_{\tilde{x}, \tilde{x}}(e^{j2\pi f}) G_m(e^{j2\pi f}) G_m^*(e^{j2\pi f})
\]

\[
= \frac{1}{T^2} S_{\tilde{x}, \tilde{x}}(e^{j2\pi f}) \left( \frac{f + m_1/L}{T} \mod \frac{1}{T} \right) \cdot G^* \left( \frac{f + m_2/L}{T} \mod \frac{1}{T} \right)
\]

vanish. Since \( G(f/T) \) occupies no more than \( 1/L \) of the SNQ frequency band, it follows that there is no overlap between the frequency responses \( G((f + m_1/L)/T \mod 1/T) \) for different values of \( m_1 \), and hence \( S_{\tilde{x}_m, \tilde{x}_m}(e^{j2\pi f}) \) indeed vanishes for \( m_1 \neq m_2 \).

#### A. Receiver Architecture

A low-complexity receiver architecture suffices to approach the associated information-theoretic limits. In particular, specializing (7) to the case (9) with our choice \( v_m[n] = e^{-j2\pi mn/L} \), the equivalent “shifted back” channel model is

\[
\tilde{y}_m[n] \triangleq y_m[n] e^{j2\pi mn/L} = \tilde{k}_m[n] * s[n] + \tilde{z}_m[n],
\tag{11}
\]
where $\tilde{k}_m[n] = k_m[n] e^{j2\pi mn/L}$. Since these channels do not overlap in frequency, they may be added without loss, resulting in the effective scalar ISI channel

$$g[n] = \sum_{m \in S} y_m[n] e^{j2\pi mn/L} = s[n] * \left( \sum_{m \in S} \tilde{k}_m[n] \right) + \sum_{m \in \bar{S}} z_m[n].$$

For such ISI channels, the unbiased MMSE decision-feedback equalizer (DFE) is an information-lossless receiver structure [3], [4]. In particular, to approach capacity, one can use Guess-Varanasi interleaving [3] and a single (fixed-rate) base code designed for an AWGN channel. In essence, every symbol is replaced by a different codeword and thus the DFE decision device acts on codewords rather than symbols.

IV. MIMO-SNQ: EXTENDING SNQ TO MIMO SYSTEMS

A straightforward extension of the preceding architecture to a multi-input multi-output (MIMO) system, which we term MIMO-SNQ, is as follows. The particular channel model of interest is

$$y_m(t) = H_m(t) * x_m(t) + z_m(t),$$

where there are $N_t$ transmit and $N_r$ receive elements. In turn, with input of the form $x_m(t) = \sum_n x_m[n] g(t - nT)$, the associated discrete-time channel, after applying a matrix matched filter, is

$$y_m[n] = K_m[n] * x_m[n] + z_m[n],$$

where $K_m[n] = K_m(nT)$ with $K_m(t) = H_m^\dagger(-t) * H_m(t) * g^\dagger(-t) * g(t)$.

We employ a single stream transmission architecture based on the application of time-varying DFT beamforming to a scalar signal $s[n]$. Specifically, we assume that the transmitted signal corresponding to packet $m$ is formed as $x_m[n] = s[n] v_m[n]$, where $v_m[n] = \psi[n] e^{-j2\pi mn/L}$ with

$$\psi[n] = \begin{bmatrix} 1 & e^{-j2\pi n/N_t} & \cdots & e^{-j2\pi(n-1)n/N_t} \end{bmatrix}^T.$$

The effective received signal for packet $m$ is, after frequency shifting [cf. (11)],

$$\tilde{y}_m[n] = y_m[n] e^{j2\pi mn/L} = \sum_l K_m[l] v[n-l] s[n-l] + z_m[n],$$

where $K_m[n] = K_m[n] e^{j2\pi mn/L}$. Note that the effective channel (12) is a periodically varying MIMO-ISI channel with period $N_t$. We refer to each of the $N_t$ induced subchannels as “phases.” As in the SISO case, we may employ a DFE at the receiver, but due to the time-varying nature of the effective channel, for each phase a different set of $N_t$ feedforward filters is applied to the channel output vector sequence. Thus, the equalizer is also periodic with period $N_t$.

Note the covariance matrix of the transmitted vector for our SNQ modulation is white. We conclude that for this modulation, the transmitted signal is white over all degrees of freedom as long as the oversampling rate satisfies $L \geq N_t M$.

However, this doesn’t guarantee capacity can be achieved. In particular, we associate with each of the $N_t$ “phases” a signal-to-interference-plus-noise ratio (SINR) value corresponding to the associated DFE slicer input. Equivalently, we may associate with each such phase a corresponding capacity. Hence, while the sum of the per-phase capacities equals the white-input capacity of the MIMO channel, the per-phase capacities are in general not equal. Moreover, since the variation is unknown to the transmitter, in a Guess-Varanasi transmission architecture, a fixed code rate is used, and thus the achievable rate is determined by the minimum of the per-phase capacities.

It is worth emphasizing that this SINR variation across phases is analogous to the SINR variation across streams in a V-BLAST system, in which independently coded streams are sent over the antennas [5]. For this reason, V-BLAST serves as a useful benchmark with which to compare the performance of SNQ modulation.

A. Parallel channels

In some cases, MIMO-SNQ is strictly capacity achieving. For example, consider the special case of $N$ parallel ISI channels. This model is essentially equivalent to using SNQ modulation for transmission over a block-varying ISI channel as considered in Section III, with the $N$ parallel channels replacing the $L$ consecutive blocks of the SISO ISI channel.

As we have shown that SNQ modulation is an optimal scheme in such a scenario, it follows that MIMO-SNQ is optimal for the case of parallel channels.

B. Channels Without Temporal ISI

As another class of channels of interest, consider the special case in which there is no temporal ISI and only inter-channel interference (ICI) is present, i.e., $H_m(t) = H_m$. While for diagonal $H_m$ MIMO-SNQ is capacity achieving, there exist other $H_m$ for which MIMO-SNQ achieves zero rate. For example, if $H_m$ is the (rank-one) matrix of all 1’s, there will exist an SNQ Nyquist substream that experiences a zero-capacity channel since a vector of all 1’s is orthogonal to $\psi[n]$ for $n \neq l N_t$, all $l$. Hence, MIMO-SNQ achieves zero rate, while V-BLAST achieves a strictly positive rate. However, for all but such pathological $H_m$, MIMO-SNQ supports a rate that grows with SNR. By contrast, it is well known that for all rank-one channel matrices V-BLAST performance is interference-limited, i.e., is bounded with increasing SNR.

More generally, when the channel matrix $H_m$ is drawn from a random ensemble, its performance is on average never worse that V-BLAST, and has significant advantages, particularly when keeping in mind that in V-BLAST an ordering is forced in the detection process, while MIMO-SNQ requires no such ordering since the scheme is inherently more symmetric.

We consider the average throughput for an ensemble of $H_m$ with i.i.d. circularly symmetric complex Gaussian entries. The resulting average throughput of MIMO-SNQ is depicted in Fig. 1, along with that for both fixed- and optimized-decoding-order V-BLAST. As the plot reflects, the performance
of MIMO-SNQ modulation lies in between the two and approaches the latter at high SNR.

We can also relate MIMO-SNQ performance to that of D-BLAST [6]. In particular, as is well known, D-BLAST can achieve capacity, but to do so requires a base code designed for time-varying channels. In that sense, MIMO-SNQ can also achieve capacity provided practical such base codes exist. However, when we require that a communication architecture has the property that the base code sees an AWGN channel, Fig. 1 reflects that MIMO-SNQ can perform as well as V-BLAST with an optimized decoding order. In contrast, when the same constraint is imposed on D-BLAST, the result is V-BLAST with a fixed decoding order.

C. Spatio-Temporal ISI Channels

When there is also temporal ISI, MIMO-SNQ is even more attractive, as we next illustrate. We now consider a random MIMO channel model where the Nyquist-rate equivalent discrete-time matrix channel impulse response $K_m[n]$ is of finite length and each Nyquist-rate tap is drawn i.i.d. over spatial and time dimensions according to a circularly-symmetric complex Gaussian distribution. Fig. 2 depicts the expected (averaged over the ensemble) capacity of MIMO-SNQ for different channel lengths. We observe that the gap-to-capacity decreases as the channel length grows confirming that SNQ modulation is able to exploit the temporal diversity afforded by the channel.

Moreover, such behavior is not specific to such i.i.d. ensembles. Indeed, Table I describes the performance of MIMO-SNQ for a typical $2 \times 2$ underwater acoustic communication channel realization from the recent KAM-11 experiment.\(^1\) Performance was numerically evaluated for an oversignaling rate of $L = 2$. In this case, the equivalent discrete-time (Nyquist-rate) baseband channel impulse responses are 100 taps long. As the table reflects, MIMO-SNQ is effectively capacity achieving for this channel, with a gap to capacity of less than 0.5 dB.

\(^1\)The authors thank Qing He for providing this sample channel.

\begin{table}[h]
\centering
\caption{MIMO-SNQ spectral efficiency (b/s/Hz) on a sample $2 \times 2$ underwater acoustic channel of length 100 taps}
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
SNR (dB) & 0 & 2 & 4 & 6 & 8 & 10 & 12 & 14 \\
\hline
Capacity & 0.95 & 1.31 & 1.76 & 2.30 & 2.94 & 3.68 & 4.51 & 5.41 \\
SNQ & 0.86 & 1.20 & 1.63 & 2.16 & 2.80 & 3.53 & 4.37 & 5.29 \\
\hline
\end{tabular}
\end{table}

REFERENCES

Abstract—In prior work, we proposed the compute-and-forward framework for sending linear combinations of messages to relays. In this note, we extend the notion of successive interference cancellation to the compute-and-forward setting. We find that once a relay has decoded a linear combination, it can mix it with its channel output to create a new effective channel output. The resulting effective channel can be tuned so that it is more suitable for decoding a second linear combination than the original channel.

I. INTRODUCTION

The classical approach to communication over a wireless relay network treats interference between transmitters as a nuisance to be avoided. Typically, each relay observes a noisy linear combination of the transmitted codewords and attempts to decode one (or more) of them while treating the others as additional noise. Recent efforts have revealed that interference can in fact be exploited if we are willing to expand the set of decoding possibilities to include linear combinations of codewords. One natural approach, sometimes referred to as compute-and-forward, is to employ a lattice codebook so that integer combinations of codewords are themselves codewords [1]–[5]. Relays are then free to select integer coefficients that match the channel coefficients as closely as possible, thus reducing the effective noise and increasing the achievable rates.

Under the classical approach, a relay can employ successive interference cancellation to remove decoded codewords from its channel observation. This decreases the effective noise encountered in the next decoding step. In this paper, we devise an analogous technique for the compute-and-forward framework. After decoding a linear combination, a relay can combine it with its channel observation to obtain a new effective channel that is even better for decoding the next linear combination. For ease of exposition, we will focus on the case where each relay wants to recover just two linear combinations.

Owing to space limitations, we do not attempt a full survey of the literature. We refer interested readers to [2] for additional references pertaining to compute-and-forward and to [6], [7] for surveys of the closely related topic of physical-layer network coding.

II. PROBLEM STATEMENT

Our setting is nearly identical to that of [2] and we reproduce some of the key definitions below. For ease of exposition, we will limit ourselves to real-valued channels and symmetric rates. We will denote addition and summation over $\mathbb{R}$ with $+$ and $\sum$, respectively, and use $\oplus$ and $\oslash$ to denote the same over $\mathbb{F}_p$.

Each transmitter (indexed by $\ell = 1, \ldots, L$) has a length-$k$ message that is drawn independently and uniformly over a prime-sized finite field, $w_\ell \in \mathbb{F}_p^k$. An encoder, $E_\ell : \mathbb{F}_p^k \rightarrow \mathbb{R}^n$, then maps the message into a length-$n$ codeword, $x_\ell = E(w_\ell)$, which must satisfy the usual power constraint $\|x_\ell\|^2 \leq nP$. The message rate is $R = (k/n)\log p$.

Each relay (indexed by $m = 1, \ldots, M$) observes a noisy linear combination of the codewords,

$$y_m = \sum_{\ell=1}^L h_{m\ell}x_\ell + z_m,$$

(1)

where the $h_{m\ell}$ are the channel coefficients and $z_m \sim \mathcal{N}(0, I)$ is i.i.d. Gaussian noise. Let $h_m = [h_{m1} \cdots h_{mL}]^T$ denote the vector of channel coefficients. The goal is for each relay to recover two linear combinations of the messages of the form

$$u_{m(1)} = \bigoplus_{\ell=1}^L q_{m\ell(1)}w_\ell \quad u_{m(2)} = \bigoplus_{\ell=1}^L q_{m\ell(2)}w_\ell$$

(2)

where the $q_{m\ell(1)}$, $q_{m\ell(2)} \in \mathbb{F}_p$ are finite field coefficients. To this end, each relay is equipped with a decoder, $D : \mathbb{R}^n \rightarrow \mathbb{F}_p^k \times \mathbb{Z}_p$ that produces estimates $\hat{u}_{m(1)}$ and $\hat{u}_{m(2)}$ of its desired linear combinations. We will say that the average probability of error is at most $\epsilon$ if

$$P\left(\bigcup_m \{\hat{u}_{m(1)} \neq u_{m(1)}\} \cup \{\hat{u}_{m(2)} \neq u_{m(2)}\}\right) < \epsilon.$$  

(3)

To map between the real-valued linear combination provided by the channel and the desired finite field linear combinations, we will need a bit of additional nomenclature. Specifically, we will refer to $u_{m(1)}$ as the linear combination with coefficient vector $a_{mL} = [a_{m1} \cdots a_{mL}]^T \in \mathbb{Z}_p^L$ if its finite field coefficients satisfy

$$q_{m\ell(1)} \equiv [a_{m\ell}] \mod p.$$  

(4)

1This is a slight abuse of notation. More formally, we should explicitly define a mapping between $\mathbb{F}_p$ and $\{0, 1, \ldots, p-1\}$. See [2, Definition 6] for more details.
Similarly, we will refer to $u^{(2)}_m$ as the linear combination with coefficient vector $b_m = [b_{m1}, \ldots, b_{mL}] \in \mathbb{Z}^L$ if
\[ q^{(2)}_m = b_m \mod p. \] (5)

We will say that the computation rate region $R(h_m, a_m, b_m)$ is achievable if, for any $\epsilon > 0$ and $n$ large enough, there exist encoders and decoders, such that all relays can recover their desired linear combinations with average probability of error $\epsilon$ as long as
\[ R < \min_m R(h_m, a_m, b_m). \] (6)

Note that the relays are free to choose which linear combinations to decode so long as (6) is satisfied.

III. NESTED LATTICE CODES

One key requirement of our scheme is that all integer combinations of codewords must be afforded protection against noise. Nested lattice codes are a natural fit for this purpose. A lattice $\Lambda$ is a discrete subgroup of $\mathbb{R}^n$ with the property that if $t_1, t_2 \in \Lambda$ then $t_1 + t_2 \in \Lambda$. By construction, all of our lattices will contain the zero vector. A pair of lattices $\Lambda, \Lambda_{FINE}$ is nested if $\Lambda \subset \Lambda_{FINE}$.

A lattice quantizer is a function, $Q: \mathbb{R}^n \rightarrow \Lambda$, that maps vectors to the nearest lattice point in Euclidean distance,
\[ Q_{\Lambda}(x) = \arg \min_{t \in \Lambda} \|x - t\| \]. (7)
The fundamental Voronoi region is the subset of points in $\mathbb{R}^n$ that quantize to the zero vector, $\mathcal{V} = \{x : Q_{\Lambda}(x) = 0\}$. The modulo operation returns the quantization error with respect to the lattice,
\[ [x] \mod \Lambda = x - Q_{\Lambda}(x), \] (8) and satisfies the distributive law,
\[ [a[x] \mod \Lambda + b[y] \mod \Lambda] \mod \Lambda = [ax + by] \mod \Lambda, \]
for any $a, b \in \mathbb{Z}$.

A nested lattice code $C$ is created by taking the set of fine lattice points that fall within the fundamental Voronoi region of the coarse lattice, $C = \Lambda_{FINE} \cap \mathcal{V}$. Erez and Zamir have shown that their exist nested lattice codes that can approach the capacity of a point-to-point Gaussian channel [8].

IV. COMPUTE-AND-FORWARD

In [2], we proposed the compute-and-forward framework as a way of communicating linear combinations of messages. Our focus was on the case where each relay decodes a single linear combination with coefficient vector $a_m$. Define
\[ R_{CF}(h, a) \triangleq \frac{1}{2} \log^+ \left( \left( \|a_m\|^2 - \frac{(h_m^T a_m)^2 \|P\}}{1 + P\|h_m\|^2} \right)^{-1} \right) \]
where $\log^+(x) \triangleq \max(\log(x), 0)$.

**Theorem 1 (2, Theorem 2):** For any set of channel vectors $h_m \in \mathbb{R}^L$ and coefficient vectors $a_m \in \mathbb{Z}^L$, the following computation rate region is achievable:
\[ R(h_m, a_m) = R_{CF}(h_m, a_m). \] (9)

We now provide a brief overview of the basic compute-and-forward encoding and decoding functions which will be useful in the proof of our main result. Using Construction A [8], [9], we select a pair of nested lattices $\Lambda \subset \Lambda_{FINE}$ that can approach the capacity of a point-to-point Gaussian channel. It can be shown that there is a one-to-one map $\phi$ between $\mathbb{R}_p$ and the nested lattice code $C$ that preserves linearity (see [2, Lemma 5]). Using this mapping, the decoder chooses a lattice point $t_\ell = \phi(w_\ell)$. It then applies a dither $d_\ell$ that is drawn independent and uniformly over $\mathcal{V}$ and transmits the result,
\[ x_\ell = [t_\ell - d_\ell] \mod \Lambda. \] (10)
Relay $m$ observes $y_m$ and has access to every dither. It scales its observation by the minimum mean-squared error (MMSE) coefficient
\[ \alpha_m = \frac{P h_m^T a_m}{1 + P\|h_m\|^2} \] (11) and removes the dithers according to the desired coefficients $a_m t_\ell$. Afterwards, it quantizes the result onto the fine lattice and takes the modulus with respect to the coarse lattice,
\[ \tilde{v}^{(1)}_m = \frac{Q_{\Lambda_{FINE}}(a_m y_m + \sum_{\ell=1}^L a_m d_\ell)}{\mod \Lambda}. \] (12)
It can be shown that, with high probability, this is equal to
\[ v^{(1)}_m = \left[ \sum_{\ell=1}^L a_m t_\ell \right] \mod \Lambda \] (13) so long as
\[ R < \frac{1}{2} \log^+ \left( \frac{P}{a_m^2 + P\|a_m h_m - a_m\|^2} \right). \] (14)
Finally, the relay applies the inverse map to get its estimate $\hat{u}^{(1)}_m = \phi^{-1}(\tilde{v}^{(1)}_m)$. Assuming that $\hat{v}^{(1)}_m = v^{(1)}_m$, it can be shown that $\hat{u}^{(1)}_m = u^{(1)}_m$ (see [2, Lemma 6]).

V. SUCCESSIVE COMPUTE-AND-FORWARD

Successive interference cancellation is a powerful technique for decoding several messages at a single receiver. Assume that, given the channel observation $y_m$, a relay has correctly decoded $x_i$. It can now completely remove the effect of $x_i$ from its observation,
\[ y_m - h_m x_i = \sum_{\ell \neq i} h_m x_\ell + z_m. \] (15)
which reduces the interference and makes it easier to decode the next codeword.

As it turns out, we can employ a similar technique when decoding several linear combinations. Assuming the relay has decoded $\sum_{\ell \neq i} a_m x_\ell$, it can create a new effective channel
\[ y_m + \gamma_m \sum_{\ell=1}^L a_m x_\ell = \sum_{\ell=1}^L (h_m x_\ell + \gamma_m a_m) x_\ell + z_m. \] (16)

2These dithers can be replaced with deterministic sequences.
By adjusting the effective channel coefficients, we can make it easier for the relay to decode its second linear combination, and thus increase the computation rate region. Note that unlike successive interference cancellation, it is not always optimal to subtract the recovered linear combination. Below, we develop a successive computation scheme that follows the concept outlined above. We begin by showing that we can always recover the real sum of codewords if we have access to the modulo sum and the dithers.

**Remark 1:** In [2, Theorem 12], we described a limited version of successive computation. The key drawback is that this scheme only allows for integer-valued $\gamma_m$, owing to the fact that it works directly with the modulo sum of codewords.

As part of the compute-and-forward scheme, the relay recovers an estimate $\hat{v}_m^{(1)}$ of the modulo linear combination of codewords $v_m^{(1)}$ from (34). The lemma below shows that this modulo sum can be used to recover the real sum $\sum_\ell a_m \ell x_\ell$ that is needed for successive computation.

**Lemma 1:** The relay can make an estimate $\hat{s}_m$ of the real sum of codewords

$$
\hat{s}_m = \sum_{\ell=1}^L a_m \ell x_\ell
$$

with vanishing probability of error, $\lim_{n \to \infty} P(\hat{s}_m \neq s_m) = 0$, so long as $R < R(\{h_m, a_m\})$.

**Proof:** Since $R < R(\{h_m, a_m\})$, we can use Theorem 1 to make an estimate $\hat{v}_m^{(1)}$ that is equal to $v_m^{(1)}$ with high probability. For the remainder of the proof, we will assume that this estimate is correct. First, the relay dithers this linear combination and takes the modulus with respect to $\Lambda$. This gives it access to a modulo combination of the dithered lattice points,

$$
\left[ \hat{v}_m^{(1)} - \sum_{\ell=1}^L a_m \ell d_\ell \right] \mod \Lambda
$$

$$
= \left[ \sum_{\ell=1}^L a_m \ell t_\ell \right] \mod \Lambda - \sum_{\ell=1}^L a_m \ell d_\ell \mod \Lambda
$$

$$
= \sum_{\ell=1}^L a_m \ell (t_\ell - d_\ell) \mod \Lambda = \sum_{\ell=1}^L a_m \ell x_\ell \mod \Lambda.
$$

It then subtracts this quantity from $a_m y_m$,

$$
r_m = a_m y_m - \sum_{\ell=1}^L a_m \ell x_\ell \mod \Lambda
$$

$$
= \sum_{\ell=1}^L a_m h_m \ell x_\ell + a_m z_m - \sum_{\ell=1}^L a_m \ell x_\ell + Q_\Lambda \left( \sum_{\ell=1}^L a_m \ell x_\ell \right)
$$

$$
= Q_\Lambda \left( \sum_{\ell=1}^L a_m \ell x_\ell \right) + a_m z_m + \sum_{\ell=1}^L (a_m h_m \ell - a_m \ell) x_\ell ,
$$

to get a quantized version of the desired sum $s_m$ plus some effective noise with variance $a_m^2 + P \|a_m h_m - a_m\|^2$. To remove this noise, it applies the coarse lattice quantizer. This operation will be successful with high probability so long as the second moment of $\Lambda$ exceeds the effective noise variance, i.e., $P > a_m^2 + P \|a_m h_m - a_m\|^2$. Assuming that $R(\{h_m, a_m\}) > 0$, this condition holds and we have that

$$
Q_\Lambda(r_m) = Q_\Lambda \left( \sum_{\ell=1}^L a_m \ell x_\ell \mod \Lambda \right)
$$

with high probability. Finally, since the relay knows the quantized sum as well as its quantization error, it can infer the desired real sum. Assuming that (21) holds, we have that

$$
\hat{s}_m = Q_\Lambda(r_m) + \sum_{\ell=1}^L a_m \ell x_\ell = s_m .
$$

Using the real sum of the codewords, we can construct a successive compute-and-forward scheme. Define

$$
R_{SCF}(h, a, b) \triangleq \frac{1}{2} \log^2 \left( \frac{\|b\|^2 - \|a\|^2}{1 + P \left( \frac{\|a\|^2}{1 + \|b\|^2} \right)} \right)
$$

**Theorem 2:** For any set of channel vectors $h_m \in \mathbb{R}^L$ and coefficient vectors $a_m, b_m \in \mathbb{Z}^L$, the following computation rate region is achievable:

$$
R(\{h_m, a_m, b_m\}) = \max \{ R_{AB}, R_{BA} \}
$$

$$
R_{AB} = \min \left( R_{SCF}(h_m, a_m), R_{SCF}(h_m, a_m, b_m) \right)
$$

$$
R_{BA} = \min \left( R_{SCF}(h_m, b_m), R_{SCF}(h_m, b_m, a_m) \right).
$$

**Proof:** Fix an $\epsilon > 0$. The expressions $R_{AB}$ and $R_{BA}$ correspond to the two possible decoding orders. We will prove that $R_{AB}$ is achievable by first decoding the linear combination with coefficient vector $a_m$ and then that with $b_m$. The proof of $R_{BA}$ follows identically by exchanging the role of $a_m$ and $b_m$.

We employ the same encoding framework as in Theorem 1. Relay $m$ uses the same decoding framework to make an estimate $\hat{u}_m^{(1)}$ of $u_m^{(1)}$. For $n$ large enough, this estimate is incorrect with probability at most $\epsilon/3$ if

$$
R < \min_m R_{SCF}(h_m, a_m) .
$$

As a byproduct of successful decoding, the relay will obtain an estimate $\hat{u}_m^{(1)}$ of $u_m^{(1)}$. Using Lemma 1, it makes an estimate $\hat{s}_m$ of $s_m = \sum_\ell a_m \ell x_\ell$ that has probability of error at most $\epsilon/3$ for $n$ large enough. Below, we assume $\hat{s}_m = s_m$.

The relay removes the projection of $\hat{s}_m$ onto $y_m$ from $y_m$ to get

$$
r_m = y_m - \frac{a_m h_m}{\|a_m\|^2} \hat{s}_m
$$

$$
= \sum_{\ell=1}^L \left( h_m \ell - \frac{a_m h_m}{\|a_m\|^2} a_m \ell \right) x_\ell + z_m
$$

105
Define \( g_{m\ell} = h_{m\ell} - a_m^T h_{m\ell} a_m^T \) and \( g_m = [g_{m1} \cdots g_{mL}]^T \). It then forms a new effective channel observation equation
\[
\tilde{y}_m = \beta_m r_m + \mu_m s_m \quad (31)
\]
\[
= \sum_{\ell=1}^L (\beta_m g_{m\ell} + \mu_m a_m^T) x_\ell + \beta_m z_m . \quad (32)
\]
and proceeds to decode the linear combination with coefficient vector \( b_m \) as in Theorem 1. Specifically, it forms the estimate
\[
\hat{u}_m^{(2)} = \left[ Q_{\Lambda^m} \left( \tilde{y}_m + \sum_{\ell=1}^L b_m^\ell d_\ell \right) \right] \text{mod } \Lambda . \quad (33)
\]
It can be shown that, with probability of error at most \( \epsilon/3 \), this is equal to
\[
\hat{v}_m^{(2)} = \left[ \sum_{\ell=1}^L b_m^\ell t_\ell \right] \text{mod } \Lambda \quad (34)
\]
so long as
\[
R < \min_m \frac{1}{2} \log^+ \left( \frac{P}{\beta_m^2 + P^2 \|b_m^\ell a_m - b_m \|^2} \right) . \quad (35)
\]
Finally, the relay applies the inverse map to get its estimate \( \hat{u}_m^{(2)} \). It remains to solve for the \( \beta_m \) and \( \mu_m \) that minimize the effective noise variance
\[
f(\beta_m, \mu_m) = \beta_m^2 + P \|b_m^\ell a_m - b_m \|^2 . \quad (36)
\]
This is a convex function whose global minimum is attained at
\[
\beta_m^* = \frac{P a_m^T b_m}{1 + P^2 \|g_m a_m \|^2} \quad (37)
\]
\[
\mu_m^* = \frac{a_m^T b_m}{\|a_m \|^2} . \quad (38)
\]
Plugging this back in, we find that
\[
f(\beta_m^*, \mu_m^*) = P \|b_m \|^2 - \frac{P^2 (a_m^T b_m)^2}{\|a_m \|^2} - \frac{P^2 (g_m^T b_m)^2}{1 + P^2 \|g_m \|^2} . \quad (39)
\]
Substituting into (35), we get the desired condition
\[
R < \min_m R_{\text{CF}}(h_m, a_m, b_m) . \quad (39)
\]
By the union bound, the probability of error is at most \( \epsilon \).

**Example 1:** Consider a single relay with channel vector \( h_1 = [2 \ 1 \ 1]^T \) that wishes to decode the linear combinations with coefficient vectors \( a_1 = [1 \ 1 \ 1]^T \) and \( b_1 = [1 - 1 - 1]^T \) using Theorem 2. It is not possible to decode \( b_1 \) first as \( R_{\text{CF}}(h_1, b_1) = 0 \). Decoding \( a_1 \) first requires
\[
R < \frac{1}{2} \log^+ \left( \frac{1 + 6P}{3 + 2P} \right) . \quad (40)
\]
After recovering \( a_1 \), the relay can adjust the channel and decode \( b_1 \) so long as
\[
R < \frac{1}{2} \log^+ \left( \frac{9}{24 + P} \right) . \quad (41)
\]

The example above demonstrates that successive compute-and-forward can make it possible to recover linear combinations that are not available via a direct application of the original compute-and-forward framework. In other words, the relay can first target a linear combination that is “easy” to decode and then use it to create a better effective channel for decoding the second linear combination.

From another perspective, successive compute-and-forward can be used to enlarge the computation rate region for decoding a single linear combination with coefficient vector \( b_m \). The relay should order all viable coefficient vectors (i.e., those satisfying \( \|a_m \|^2 \leq 1 + P \|b_m \|^2 \)) by computation rate \( R_{\text{CF}}(h_m, a_m) \) and set aside those \( a_m \) with rates larger than \( R_{\text{CF}}(h_m, b_m) \). It can then calculate which pair \((a_m, b_m)\) offers the highest rate using Theorem 2. Finally, it applies successive compute-and-forward for this pair and keeps only the second equation. Example 1 demonstrates that this procedure does indeed enlarge the rate region.

**VI. GENERALIZATIONS AND EXTENSIONS**

Following the framework in [2], successive compute-and-forward can be generalized to include complex-valued channel models as well as unequal message rates. One can also envision extending this technique to the case where each relay may want more than two linear combinations. In this case, the linear combinations obtained thus far should be mixed together with the original channel observation to create a new effective channel for the next targeted linear combination.

**REFERENCES**


Tunable Sparse Network Coding

Soheil Feizi1, Daniel E. Lucani2, Muriel Médard1

1Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
2Instituto de Telecommunicações, DEEC, Faculdade de Engenharia da Universidade do Porto, Porto 4200-465, Portugal
Email: sfeizi@mit.edu; dlucani@fe.up.pt; medard@mit.edu

Abstract—A fundamental understanding of the relationship between delay performance and complexity in network coding is instrumental towards its application in practical systems. The main argument against delay-optimal random linear network coding (RLNC) is its decoding complexity, which is $O(n^3)$ for $n$ original packets. Fountain codes, such as LT and Raptor codes, reduce the decoding load on the receiver but at the cost of introducing additional, non-negligible delay. The source of this increased delay is the inherent sparsity of the code, which significantly reduces the impact of a new coded packet, i.e., the probability of a packet to be linearly independent from the knowledge at the receiver, as we approach the end of the transmission (when few degrees of freedom are missing). Thus, the additional overhead is mainly due to the transmission of the last data packets. Our key observation is that switching gears to denser codes as the transmission process progresses could considerably reduce the delay overhead while maintaining the complexity advantages of a sparse code. We propose tunable sparse network coding as a dynamic coding mechanism with a code structure with two regions: a sparse region, with various levels of sparsity, and a dense region, where packets are generated as per RLNC. We characterize the problem in multicast sessions on general networks and illustrate trade-offs in special cases of interest. We also present a novel mechanism to perform efficient Gaussian elimination for sparse matrices that guarantees linear decoding complexity with high probability.

I. INTRODUCTION

Network coding constitutes a disruptive concept for the operation of communication networks [1] that has evolved from an information theoretic result for achieving capacity in multicast to become an enabler of a wide variety of applications, including communications in ad-hoc networks, data gathering in sensor networks, and data storage. One of the key features of network coding is that it encourages the mixing (coding) of data packets at intermediate nodes, rather than limiting intermediate nodes to store and forward packets. Under this premise, it is no longer required for the system to keep track of which packets have been received; receivers need only accumulate enough coded packets in order to recover the information. Ref. [2] and [3] show that linear codes over a network are sufficient to achieve multicast capacity, while Ref. [4] proved that randomly generated linear codes achieve capacity with high probability.

One of the main arguments against the use of RLNC is that its decoding complexity is $O(n^3)$, where $n$ constitutes the number of original packets. Sparse end-to-end erasure correcting codes, such as LT [5] and Raptor codes [6], reduce the decoding load on the receiver but at the cost of introducing additional, non-negligible delay. More importantly, they lack RLNC’s capability to use re-encode at intermediate nodes. Sparse end-to-end erasure correcting codes tend to choose sparsity in a static manner. Although this sparsity helps to reduce the complexity, it also introduces additional delay overheads. In particular, this inherent (and static) sparsity reduces the impact of a coded packet as the transmission session progresses. In fact, the additional overhead of these codes is mainly due to the transmission of the latter packets.

We revisit the idea of using sparse coding to reduce the decoding complexity, but using a fundamentally different approach in which i) we tune the level of sparsity as the transmission process evolves, and ii) we transition to RLNC for the last transmissions. This tuning process can reduce the delay overhead by using denser codes towards the end of the transmission, while maintaining complexity advantages of a sparse code. We call this scheme Tunable Sparse Network Coding. Unlike RLNC, in tunable sparse network coding the coding coefficients are chosen sparsely, i.e., mostly zero. As the session progresses, these coefficients become denser, i.e., statistical properties of the coding coefficients vary in time.

There are three key ideas in tunable sparse network coding:

- **Sparse coding is more beneficial at the beginning of a transmission session.** Since destination nodes have only received a few degrees of freedom (independent linear combination of the original packets), any sparsely coded packet will constitute a new degree of freedom (or innovative packet) with high probability.

- **Dense coding is required towards the end of the sessions for fast completion** because dense coded packets are innovative with high probability, while sparse coded packets are innovative with low probability.

- **Sparse coding translates in reduced complexity.** A decoder can exploit an encoding matrix with a large number of zero coefficients for improved complexity.

This paper proposes a framework for studying our network coding mechanism over packet erasure networks to analyze the trade-off between the delay performance (i.e., mean number of transmissions) and the decoding complexity of multicast sessions. Finally, we introduce a simple, yet powerful algorithm to organize a sparse matrix in order to provide a linear decoding complexity with high probability.

II. MODEL AND PRELIMINARIES

Consider a general multicast network with one source and several receivers. We model the network as a directed graph, with independent edge erasure probabilities. This assumption may be relaxed in future studies, but it is a fair assumption in many networks. Suppose the source node has $n$ packets to transmit to a subset of the nodes in the network, i.e., a multicast session. Suppose $X$ is a $n \times 1$ vector of source packets whose $i^{th}$ component, $X_i$, represents the $i^{th}$ source packet in the finite field $\mathbb{F}_q$. Suppose the minimum min-cut between
each receiver and source is $m$. Therefore, the network can deliver $m$ packets at each transmission round. A transmission session is complete when all destination nodes receive $n$ source packets. Throughout this paper, we assume $m << n$. Hence, multiple transmission rounds of $m$ coded packets are required in a session. Each transmission round is indexed by $t$.

Let us leverage some of the ideas from the algebraic framework for linear network coding from Ref. [3] to model our problem. For the case of no erasures, the receiver $i$ gets $A_iX$, where $A_i$ is a $m \times n$ full-rank matrix (since the min-cut rate is $m$) after each round. $A_i$ is called the network coding matrix associated with the receiver $i$ and can be decomposed into two matrices $B$ and $G_i$ (i.e., $A_i = G_i \cdot B$). $B$ is a $m \times n$ precoding matrix which is applied at the source node. This matrix is the same for all receivers. $G_i$ is a $m \times n$ matrix representing the effect of performing network coding at intermediate nodes from the source node to the receiver $i$.

Now, consider the case of having packet erasures over the network links. Since some packets are lost as they traverse the network, the min-cut rate of this network is smaller than the error-free min-cut rate of the network (i.e., $m$). For receiver $i$, define a $m \times m$ diagonal matrix $P_i$, whose $j$th component on the diagonal (i.e., $P_{ij}$) is zero with probability $p_{ij}$, and one with probability $1 - p_{ij}$. Probabilities $p_{ij}$ depend on edge erasure probabilities. Hence, at the end of each transmission round, the destination node $i$, receives $P_iA_iX$. In other words, some of coded packets are erased over the network. Considering multiple transmission rounds, $Y(t) = P_tA_t(t)X$ represents the received coded packets at receiver $i$ in round $t$.

III. TUNABLE SPARSE NETWORK CODING

In this section, we introduce a framework for studying a time-varying coding scheme, with particular emphasis on the proposed case of tunable sparse network coding. A coded packet is uniformly coded when its coding coefficients are uniformly chosen from a finite field $\mathbb{F}_q$. On the other hand, a coded packet is $k$-sparse if it contains exactly $k$ non-zero coding coefficients. Non-zero coefficients are chosen uniformly from $\mathbb{F}_q - \{0\}$. Finally, a coded packet is $k$-sparse with high probability (w.h.p.) when, with probability $1 - k/n$, each coding coefficient is zero, otherwise, it is chosen uniformly from $\mathbb{F}_q - \{0\}$. For simplicity, consider that $q$ is sufficiently large, although our techniques extend naturally to small $q$.

Define $\rho(A_i(t))$, the density of a matrix $A_i(t)$, as the ratio of its non-zero components. A matrix is dense if this ratio is high, e.g., RLNC’s $A_i(t)$ is dense w.h.p. If the ratio is low, the matrix is sparse. In a tunable sparse network coding scheme, the density of $A_i(t)$ is tuned depending on the transmission round index $t$. The key idea is that, at the beginning of the transmission session, receivers have only a few coded packets making a sparsely coded packet innovative to the receivers w.h.p. However, towards the end of the transmission session, sparsely coded packets are innovative with low probability because destination nodes have received most of the required degrees of freedom (dof). Hence, $\rho(A_i(t))$ should be increased, i.e., densely coded packets should be transmitted.

Depending on the function $\rho(A_i(t))$ and its control method, we can classify various tunable sparse network coding schemes. If the density control of $A_i(t)$ is performed without having feedback from receivers, we say the coding scheme is open-loop. Otherwise, it is a closed-loop scheme. By having different time dependencies of elements of $A_i(t)$, a variety of coding schemes, such as window-based coding, sliding window-based codings, and generation-based encoding, can be characterized and studied with our framework.

This work focuses on an open loop approach with prior knowledge of the network’s erasure probabilities. In particular, we characterize the delay and complexity performance from the perspective of a receiver and for the special case where the error-free min-cut rate is one. However, the analysis extends seamlessly to (i) multiple receivers with an error-free min-cut rate $m$ greater than one, and (ii) close-loop schemes, because our analysis relies on characterizing receivers’ performance given the number and type of received coded packets.

A. From a Receiver’s Perspective

A tunable sparse network coding scheme with $k$-sparsely coded packets for $m = 1$ can be described as follows: the source node first transmits $n$ uncoded packets, where some of these packets are lost due to erasures. These uncoded packets are followed by sparse packets. We shall study two different approaches: i) transmission of $n_k$ $k$-sparse coded packets, and ii) transmission of $n_k$ $k$-sparse coded packets w.h.p., which we call deterministic density, and ii) transmission of $n_k$, $k$-sparse coded w.h.p., which we call randomized density. Finally, $n_{uni}$ uniformly coded packets are transmitted.

We analyze the effect of $n_k$ on the delay performance and the decoding complexity for the general case of several levels of sparsity. Without loss of generality, we restrict our delay performance achievability result in Theorem 3 to the case of one type of sparsely coded packets, i.e., only one $k \neq 1$ for which $n_k > 0$. This result extends seamlessly to the case of different levels of sparsely coded packets.

B. Delay Performance

This subsection characterizes the probability that a newly received, sparse coded packet is innovative under specific density distributions. This problem is key to modeling the general problem of a multicast session over a general network, where at each transmission time $t$ up to $m$ sparse packets will arrive to a receiver. In the general network case, the choice of precoding and the coding within the network as well as the link erasures will determine the sparsity of the received coded packets for that time $t$. Considering a scheme with multiple levels of sparsity, a transmission of $g(d_k|n_p)$ sparse packets for
a level \( k \) yields \( d_k \) dof, when \( n_p \) dof were previously received. The following lemma characterizes \( d_k(n_p) \) and \( r(n_p) \) for deterministic and randomized sparsity, respectively.

**Lemma 1.** For a receiver in a multicast session with \( n_p < n \) previously received degrees of freedom,

\[
g_d(d_k|n_p) = \sum_{i=n_p}^{n_p + d_k} \frac{1}{(\sum_{j=0}^{k-1} \frac{i-i}{n-j})}, \quad \text{if } k < (1-\alpha)n, \quad (1)\]

\[
gr(d_k|n_p) = \sum_{i=n_p}^{n_p + d_k} \frac{1}{(1 - k/n(1-i/n))}, \quad (2)\]

**Proof:** First we compute \( g_d(d_k|n_p) \) for the case of deterministic density and then point out the necessary changes for \( g_r(d_k|n_p) \). Suppose at time \( t \), we have \( d \) dof at the receiver. For the case of deterministic density, a \( k \)-sparse coded packet is not innovative with probability \( \left( \prod_{j=0}^{k-1} \frac{i-i}{n-j} \right) \), which represents the probability of all the \( k \) non-zero coefficients being chosen amongst packets that do not increase the knowledge space of the receiver. We model the reception of a new packet as the Markov chain shown in Figure 1. The expected number of required sparsely coded packets to have \( d_k \) new dof, conditioned on having \( n_p \) innovative packets at time \( t \) is Eq. (1).

For the case of randomized density, a \( k \)-sparsely coded packet is not innovative with probability \( (1 - k/n(1-i/n))^n \) and again model this as a Markov chain to show Eq. (2).

The following corollary provides approximate expressions for Eq. (1) and (2) for \( n \) is large enough and \( d_k << n - n_p \).

**Corollary 2.** For large \( n \) and \( d_k << n - n_p \). If \( (n_p/n)^k << 1 \) then, \( g_d(d_k|n_p) \approx d_k(1 + (1-n_p/n)^k) \). If \( (1 - k/n(1-n_p/n))^n << 1 \) then, \( g_r(d_k|n_p) \approx d_k(1 + (1 - k/n(1-n_p/n))^n) \).

**Proof:** Under conditions of Corollary 2, we have that

\[
g_d(d_k|n_p) \approx \sum_{i=n_p}^{n_p + d_k} \frac{n}{1 - x^k} \approx \sum_{i=n_p}^{n_p + d_k/n} \frac{n}{1 + x^k} \approx d_k(1 + (1-n_p/n)^k),\]

We use similar approximations for \( g_r(d_k|n_p) \).

Define an \((n_k, n_{uni})\) tunable sparse network coding scheme as a scheme that transmits \( n \) systematic packets, \( n_{uni} \) uniformly coded packets, and \( n_k \) \( k \)-sparse coded packets. This scheme is achievable if the receiver is expected to recover all \( n \) packets. The following result characterizes achievability for an erasure probability \( \alpha \) as seen by the receiver. For this result, define \( f(d_k) \) as the number of \( k \)-sparse coded packets to have \( d_k \) innovative ones in average. \( f(d_k) - d_k \) represents the performance loss due to using sparse network coding instead of RLNC with large \( q \).

**Theorem 3.** For a network with a single receiver and \( m = 1 \), a \((n_k, n_{uni})\) tunable sparse network coding scheme is achievable where \( n_k = \left\lfloor \frac{f(d_k)}{d_k} \right\rfloor \) and \( d_k + (1-\alpha)n_{uni} = \alpha n \). For a deterministic density of \( k < (1-\alpha)n \), \( f(d_k) = \sum \left( 1 - k/n(1-i/n) \right)) \).

For a randomized density, \( f(d_k) = g_r(d_k|1-\alpha)n \).

---

**Fig. 2.** Example of proposed decoding algorithm (Algorithm 1)

**Proof:** Since the expected value of \( n_p \) is \((1-\alpha)n \) we compute having \( f(d_k) = E[g_d(d_k|D_1)] \) and \( f(d_k) = E[g_r(d_k|D_1)] \) using Lemma 1. We conclude the proof following a typicality argument for large \( n \) since the distribution of \( n_p \) is given by a binomial distribution with success probability \( 1 - \alpha \).

C. Complexity Analysis

We present a simple algorithm that reorders the matrix to avoid operations in the forward pass w.h.p. for the processing of sparse packets. This has the crucial effect of avoiding additional density to be created in sparse packets, which keeps the number of operations in the backward pass for the sparse packets as \( O(d_k) \). Define \( P_1 \) as the \( i \)-th original packet, \( C_j(Z) \) as the set of coded packets of \( j \)-sparse packets drawn from \( Z \) a subset of the original packets, and \( S_1 \) as the set of packets in a 1-sparse combination. Define \( \text{sum}(F, Z) \) as the set of original packet of subset \( Z \) used only once in the set of coded packets \( F \). Finally, we define \( \Omega \) as the set of all \( n \) original packets.

Fig. 2 (a) illustrates the case of a transmission using tunable sparse network coding. For simplicity, we have assumed that received coded packets were either uncoded, i.e., from the packet region, or sparse, although for this example \( n \) is kept small for illustration purposes. Fig. 2 (a) shows that packets \( P_1 \) and \( P_2 \) were received uncoded and we use them to substract their contribution from coded packets with a non-zero coefficient for \( P_1 \) and \( P_2 \). i.e., all non-zero elements of the first two columns. The algorithm counts the number of times a packet corresponding to columns on the lower right-hand side, as defined by the dashed lines, is involved in a linear combination. Involvement of the packet in combinations below the dashed line are not counted. Rows and columns shall be reorganized so that packets with a single occurrence will be in the upper left corner, as defined by the dashed lines in Fig. 2 (b). At this point we repeat the process until no single-occurrence packets are identified or until all sparse coded packets have been processed. In Fig. 2 (b), it is clear that the latter is attained because packet \( P_3 \) is involved in only one linear combination below the dashed line. The algorithm is stated in the following.

**Algorithm 1**

STEP 0: \( R(0) = \{C_k(\Omega) : k > 1 \} \) \( C(0) = \{P_j : P_j \notin S_1 \} \)

STEP i: If \( \text{sum}(R(i), C(i)) \neq \emptyset \)

\( R(i+1) = R(i) - \{C_k(\text{sum}(R(i), C(i)) : k > 1 \}, \)

\( C(i+1) = C(i) - \text{sum}(R(i), C(i)) \)

\( i \leftarrow i + 1, \) Go to next step

Else Stop. Proceed with Standard Gaussian Elimination
Algorithm 1 is the original algorithm for sparse network coding. Theorem 4 states that the complexity of decoding sparse packets with a randomized density, given that \( n_p \) innovative packets have already been ordered as per Algorithm 1 is \( O(d_k) \).

**Theorem 4.** For tunable sparse network coding with randomized density with \( d_k \) \( k \)-sparsely random packets when \( n_p \) degrees of freedom have been processed before with Algorithm 1 and, the average number of operations to process the \( d_k \) sparse packets using Algorithm 1 with probability \( k = 1 \) is \( c(n_p, d_k) = k d_k \).

**Proof:** First, the probability of a packet being used only once in \( j \) independent linear \( k \)-sparsely coded packets is \( P_j = (1 - k/n)^j - 1 \). Second, we process on a row by row basis such that there is a decreasing number of linear combinations from the original \( d_k \). For the \( r \)-th processed row, the probability of having at least one packet being used only once is given by \( P(r) = 1 - (1 - (1 - k/n) d_k - 1 - r)^{n_p - 1} \). If all \( d_k \) packets can be processed exclusively with the reordering algorithm, only \( O(d_k) \) operations for reordering during the forward pass and \( O(d_k) \) operations for the backward elimination pass are required. The probability of this event is \( \prod_{l=0}^{d_k-1} P(l) \).

**IV. PERFORMANCE EVALUATION**

This section illustrates the trade-off between complexity and delay performance via numerical examples and show that the threshold to switch from sparse to dense codes is key in determining this trade-off. For simplicity, we assume that no packets are sent uncoded at the beginning of the transmission. We compare the following schemes:

- **3-sparse Network Coding:** This scheme considers the transmission of 3-sparse packets only. For determining the delay performance, we use Eq. (1) and use Algorithm 1 to determine complexity.
- **RLNC:** This constitutes a classical RLNC scheme with cubic complexity. In particular, we consider that \( GE(n) = n^3/3 + n^2/2 - 5n/6 \), which is a usual estimate of operations necessary to perform Gaussian elimination. For RLNC we assume that every incoming packet is innovative, which is a good approximation when using large field sizes.

**Tunable Sparse Network Coding Scheme:** This scheme relies on both a sparse and a dense region, with a tunable threshold to decide when to switch from one to the other. Packets generated in the sparse region are \( k \)-sparse, with \( k \) being incremented every time \( n/100 \) coded packets sent.

Fig. 3 shows the performance of the three schemes as more packets are coming into the system. The right and top-most point for each curve constitutes the final pair of values for delay and complexity. Fig. 3 shows our proposed scheme for different threshold levels, namely when 1, 2, 5, 10, 50 and 75% of the degrees of freedom are missing at the receiver. We observe that i) switching too fast to RLNC is counterproductive from a complexity perspective, and ii) gains in terms of complexity become important when the switch to RLNC happens for the very last packets (See case of 1%). Finally, the case of switching for the last 1% helps us confirm our initial intuition, namely that a sparse code’s delay performance is primarily degraded when few degrees of freedom are missing.

**V. CONCLUSIONS**

We revisit the problem of trading off delay performance and decoding complexity but providing a novel framework that uses both sparse and dense coded packets and allows coding at intermediate nodes for multicast sessions in general networks. More generally, we propose that the sparsity level should evolve as the transmission session progresses. To enable this coding flexibility, we proposed an efficient and intuitive algorithm for reordering rows and columns of a matrix in order to perform efficient Gaussian elimination in sparse matrices. This algorithm guarantees linear decoding complexity of sparse packets w. h. p.

Our approach can benefit considerably from the use of feedback in order to change the density of the code and studying efficient feedback mechanisms will be the focus of future work. Future work will also consider efficient and practical pre-coding mechanisms at the source for the case of single and multi source multicast to deal with specific network topologies and practical coding policies at intermediate nodes.

**VI. ACKNOWLEDGEMENTS**

This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under award No. FA9550-08-1-0159 and by the FCT project Pest-OE/EEI/LA0008/2011.

**REFERENCES**


The Flatness Factor in Lattice Network Coding: Design Criterion and Decoding Algorithm

Jean-Claude Belfiore
Department of Communications and Electronics
Telecom ParisTech, LTCI UMR 5141
46, rue Barrault, 75013 Paris, France
Email: belfiore@enst.fr

Cong Ling
Department of Electrical and Electronic Engineering
Imperial College London
London SW7 2AZ, United Kingdom
Email: cling@ieee.org

Abstract—In a recent work, Nazer and Gastpar proposed the compute-and-forward strategy as a physical-layer network coding scheme. They described a code structure based on nested lattices whose algebraic structure makes the scheme reliable and efficient. In a more recent paper, Niesen and Whiting revealed a fundamental limitation of the decoder used by Nazer and Gastpar. In this work, we consider maximum-likelihood decoding of compute-and-forward, aiming to overcome its limitation. By examining the decoding metric from the viewpoint of Gaussian measures over lattices, we present a design criterion of the lattice code based on the flatness factor, and propose a new decoding algorithm based on inhomogeneous Diophantine approximation.

I. INTRODUCTION

In [1], Zhang et al. introduced the physical-layer network coding concept (PNC) in order to turn the broadcast property of the wireless channel into a capacity boosting advantage. Instead of considering the interference as a nuisance, each relay converts an interfering signal into a combination of simultaneously transmitted codewords. PNC concept has received much interest in the last years because it provides means of embracing interference and improving network capacity.

In a recent work [2], Nazer and Gastpar proposed a new physical-layer network coding scheme. The proposed strategy, called compute-and-forward (CF), exploits interference to obtain higher end-to-end transmission rates between users in a network. The relays are required to decode noiseless linear combinations of the transmitted signals through the channel. The destination, given enough linear combinations, can solve the linear system for its desired messages. This strategy is based on the use of structured codes, particularly nested lattice codes to ensure that integer combinations of codewords are themselves codewords. The authors demonstrated its asymptotic gain using information-theoretic tools. Standard minimum Euclidean distance decoding was assumed.

The authors in [3] followed and generalized the framework of Nazer and Gastpar by introducing the Lattice Network Coding (LNC). They related the Nazer-Gastpar’s approach to the theorem of finitely generated modules over a principal ideal domain (PID). They gave sufficient conditions for lattice partitions to have a vector space structure which is a desirable property to make them well suited for PNC. Then, they generalized the lattice code construction and developed encoding and decoding methods. Standard minimum Euclidean distance decoding was still assumed.

In a more recent work [4], Niesen and Whiting revealed a fundamental limitation of standard minimum Euclidean distance decoding, namely, it achieves no more than 2 degrees of freedom. This is due to the quantization of channel coefficients, which leads to a tradeoff between channel coefficient approximation and noise amplification. This tradeoff was associated with a problem of Diophantine approximation in [4]. They further proposed a scheme achieving higher degrees of freedom based on real interference alignment. However, this scheme is not robust, among many of its shortcomings.

In [5], the first author essentially proposed an analysis of the compute-and-forward strategy when maximum-likelihood (ML) decoding is used. It should be emphasized that ML decoding does not subscribe to the limitation shown in [4]. A new quantity, the Flatness Factor, was introduced, which measures the quality of ML decoding. After changing the definition of the flatness factor in order to be consistent with [8], [9], this paper studies more deeply the properties of it and gives some new insights into lattice network coding.

II. COMPUTE AND FORWARD

In our model, we consider one relay receiving messages from \( k \) sources \( S_1, \ldots, S_k \) and transmitting a linear combination of these \( k \) messages. The relay observes a noisy linear combination of the transmitted signals through the channel. Received signal at the relay is expressed as,

\[
y = \sum_{j=1}^{k} h_j x_j + w
\]

where \( h_j \) is the channel coefficient between source \( S_j \) and the relay and \( x_j \) is the vector transmitted by source \( S_j \).

The relay searches to transmit a noiseless linear combination of the received signals with integer coefficients given by vector \( \alpha = [ a_1 \ a_2 \ \cdots \ a_k ]^T \). To achieve that goal, it decodes a noiseless linear combination of the transmitted vectors,

\[
x_R = \sum_{j=1}^{k} a_j x_j
\]
and retransmits it to the destination or another relay. We consider a complex-valued channel model with complex inputs and outputs. The channel coefficients $h_i$ are complex, circular, i.i.d. Gaussian, $h_i \sim \mathcal{N}(0,1)$. $z$ is Gaussian, zero mean, with variance $\sigma^2$ ($z \sim \mathcal{N}(0,\sigma)$). Let $h = [h_1 \ h_2 \ \cdots \ h_k]^T$ denote the vector of channel coefficients. Source vectors $x_j$ are carved from a lattice code. The sources have no channel side information (CSI). CSI is only available at the relay.

III. THE ML DECODER

A. Solving the System of Diophantine Equations

The relay aims to decode a linear system of equations of the transmitted messages and passes it towards the destination. After calculating the vector $a$ chosen following some given criterion, the relay recovers a linear combination of the transmitted signals. We rewrite the received (scaled) signal at the relay in the following form,

$$y = \lambda + \sum_{j=1}^{k} \xi_j x_j + w$$

where $\lambda$ is a point of a lattice $\Lambda$, $\xi_j = h_j - a_j$ and $w$ is the additive i.i.d. Gaussian noise. The recovered linear system of equations $\lambda = \sum_{j=1}^{k} a_j x_j$ is a linear system of Diophantine equations. This system admits the following solutions.

Suppose that the transmitted vectors $x_j \in \Lambda_j$ where $\Lambda_j$ are full rank lattices. Then, since

$$\lambda = \sum_{j=1}^{k} a_j x_j,$$

the lattice in which $\lambda$ lives is

$$\Lambda = a_1 \Lambda_1 + a_2 \Lambda_2 + \cdots + a_k \Lambda_k.$$

Our aim, here, is to express the messages $x_j$ as a function of $\lambda$, for a given $\lambda \in \Lambda$. But the system of Diophantine equations admits an infinite number of solutions. The Hermite Normal Form (HNF) allows to solve this system [7]. We show here how to proceed.

From Equation (4), we get $\lambda = \sum_{j=1}^{k} a_j M_j z_j$ where $z_j \in \mathbb{Z}[i]^n$ and $M_j$ is a generator matrix for the lattice $\Lambda_j$. Then, Equation (4) can be expressed as

$$\begin{bmatrix} a_1 M_1 \\ a_2 M_2 \\ \cdots \\ a_k M_k \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_k \end{bmatrix} = \lambda.$$ (5)

Define $M = \begin{bmatrix} a_1 M_1 & a_2 M_2 & \cdots & a_k M_k \end{bmatrix}$ and use the Hermite Normal Form of $M$ [7],

$$M \cdot U = \begin{bmatrix} 0 & B \end{bmatrix}.$$

where matrices $V_j$ are square $n \times n$ matrices. Then the solution to the system of Diophantine equations is,

$$x_j = s_j + v_j$$ (6)

where $v_j = M_j V_j B^{-1} \lambda$ and $s_j$ is any vector in the lattice $L_j \subset \Lambda_j$ of generator matrix $M_j U_j$.

B. ML Decoding Metric

The ML decoder maximizes $p(y/\lambda)$ over all possible values of $\lambda$. The conditional probability $p(y/\lambda)$ can be expressed as,

$$p(y/\lambda) = \sum_{\{x_1,\ldots,x_k\}} p(y/x_1,\ldots,x_k) p(x_1,\ldots,x_k)$$ (7)

where

$$p(y/x_1,\ldots,x_k) \propto e^{-\frac{\|y - \sum_{j=1}^{k} h_j x_j\|^2}{2\sigma^2}}$$ (8)

and $x_1, \ldots, x_k$ are (a priori) supposed equiprobable and given by (6). The decoding rule is now to find,

$$\hat{\lambda} = \arg \max_{\lambda \in \Lambda} \varrho(\lambda),$$ (9)

with

$$\varrho(\lambda) = \sum_{q \in \mathcal{L}} e^{-\frac{\|u(\lambda) - q\|^2}{2\sigma^2}}.$$ (10)

where $u(\lambda) = y - \left(\sum_{j=1}^{k} h_j M_j V_j B^{-1} \lambda\right)$ and $\mathcal{L}$ the $n$-dimensional $\mathbb{Z}[i]$-lattice with generator matrix $C = \sum_{j=1}^{k} h_j M_j U_j$. We observe that, for some values of the parameters, the function $\varrho(\lambda)$ can be almost constant or flat, which will yield a high error probability. Figure III-B illustrate the behavior of the likelihood function for difference noise variances. Clearly, we should avoid the second and third cases where it is almost flat. In the next section, we will use the flatness factor to assess the behavior of the likelihood function.

IV. THE FLATNESS FACTOR

A. Definition

For a given lattice $\mathcal{L}$ and a vector $y$, we consider the $n$-variable function

$$f_{\sigma,\mathcal{L}}(y) = \frac{1}{(2\pi\sigma)^{n/2}} \sum_{q \in \mathcal{L}} e^{-\frac{\|y - q\|^2}{2\sigma^2}},$$

which is a Guassian measure over a lattice. Obviously, $\varrho(\lambda) \propto f_{\sigma,\mathcal{L}}(u(\lambda))$. This function has been studied in [8]. We want here, to avoid that this function be flat for the lattice $\mathcal{L}$.

Definition 1. [Flatness factor] The flatness factor of the Gaussian measure is defined as the maximum ratio $f_{\sigma,\mathcal{L}}(y)$ can deviate from the uniform distribution:

$$\epsilon(\sigma) = \max_{y \in \mathbb{R}^n} \frac{|f_{\sigma,\mathcal{L}}(y) - E_y f_{\sigma,\mathcal{L}}(y)|}{E_y f_{\sigma,\mathcal{L}}(y)}.$$ (10)

Note that this definition is slightly different from that in [5], for the sake of consistency with [9] where the problem of physical-layer security is addressed (flatness is desired in there). The following proposition gives the expression of the
3.1

Figure 1. $f_{\sigma, \Lambda}(y)$ for $\Lambda=2 \mathbb{Z}^2$, $\sigma^2 = 0.3, 0.5$ and 0.8 respectively.

3.2

The flatness factor is closely related to the smoothing parameter studied in [8]. The advantage of the flatness factor in terms of the theta series (the proof is given in [9]).

**Proposition 2.** [Expression of $\epsilon_{\mathcal{L}}(\sigma)$]

$$
\epsilon_{\mathcal{L}}(\sigma) = \mu^n \Theta_{\mathcal{L}} \left( \frac{1}{2\pi \sigma^2} \right) = 1
$$

(11)

where $\mu = \frac{\sqrt{\chi(\mathcal{L})^2}}{2\pi \sigma^2}$ is the generalized signal-to-noise ratio, and $
\Theta_{\mathcal{L}}(z) = \sum_{\lambda \in \mathcal{L}} e^{-\pi \|z\|^2}$ is the theta series of $\mathcal{L}$.

The flatness factor is closely related to the smoothing parameter studied in [8]. The advantage of the flatness factor is two-fold:

- It gives a precise characterization $\epsilon_{\mathcal{L}}(\sigma)$ by the theta series, so it is more precise than the bounds available in [8].
- The smoothing parameter is mostly concerned with small values of $\epsilon_{\mathcal{L}}(\sigma)$, while the flatness factor can handle both small and large values. The latter is of interest in coding applications.

3.3

**B. Design Criterion**

We need that the flatness factor of $\mathcal{L}$ be as large as possible to ensure that the maximum likelihood metric will be as different as possible from the likelihood metrics of the other points of the lattice $\mathcal{L}$. Formally, we have the following design criterion of the lattice $\mathcal{L}$:

**Proposition 3.** [Goodness for network coding] In order to have good decoding performance in compute-and-forward, the theta series $\Theta_{\mathcal{L}}(z)$ should be large.

This requires from the lattice $\mathcal{L}$ to be as ‘bad’ as possible in the standard sense of coding (i.e, not dense at all). The generator matrix of $\mathcal{L}$ is $\sum_{j=1}^{k} a_j M_j U_j$. But, since $\sum_{j=1}^{k} a_j M_j U_j = 0$, from the HNF, we deduce that we have to align vector $a$ with vector $h$ as much as possible.

As an illustration, Figure 2 shows the values of some flatness factors as a function of the generalized signal to noise ratio $\mu$. For our purpose, the Leech lattice is worse than the $\mathbb{Z}^n$ lattice.

V. DECODING ALGORITHM

We are now interested in an algorithm which must maximize $\vartheta(\lambda)$ in an efficient way. To do this, we need the Fourier transform and Jacobi theta functions. In fact, we will not attempt to compute $\vartheta(\lambda)$ itself; instead, we will propose inhomogeneous Diophantine approximation as a solution to ML decoding.

**A. Fourier transform and Jacobi theta function**

Since $f_{\sigma, \mathcal{L}}(y)$ is obviously periodic on $\mathcal{L}$, its Fourier transform is defined on the dual lattice $\mathcal{L}^*$. The Fourier coefficients are given by

$$
f_{\sigma, \mathcal{L}}(q^*) = \frac{1}{V(\mathcal{L})} e^{-2\pi^2 \sigma^2 \|q^*\|^2}, \quad q^* \in \mathcal{L}^*.
$$

(12)

The Fourier expansion of $f_{\sigma, \mathcal{L}}(y)$ is then

$$
f_{\sigma, \mathcal{L}}(y) = \frac{1}{V(\mathcal{L})} \sum_{q^* \in \mathcal{L}^*} e^{-2\pi^2 \sigma^2 \|q^*\|^2} e^{2\pi i (q^* \cdot y)}. \quad (13)
$$

Now we express the Gaussian measure as the Jacobi theta function (sometime referred to as the Riemann theta function in the multi-dimensional case), via the Fourier transform (13). Let $\mathbb{H}_n$ be set of symmetric square matrices whose imaginary part is positive definite. For an $n \times n$ matrix $T \in \mathbb{H}_n$ and an $n$-dimensional complex vector $\mathbf{z}$, the multi-dimensional theta function is defined as

$$
\theta(\mathbf{z}; T) = \sum_{q \in \mathbb{Z}^n} e^{\pi i q^T T q + 2\pi i q^T \mathbf{z}}.
$$

(14)

Comparing with (13), we recognize that

$$
f_{\sigma, \mathcal{L}}(y) = \frac{1}{V(\mathcal{L})} \theta(\mathbf{z}; T)
$$

(15)

Figure 2. Flatness factor of the Leech lattice vs the cubic lattice
for \( z = C^{-1} y \) and \( T = 2\pi \sigma^2 (CTC)^{-1} \). Known results of the Jacobi theta function may be applied to analyze the Gaussian measure over lattices.

**B. The 1-D Real Case**

Here,

\[
\vartheta(\lambda) = \sum_{k \in \mathbb{Z}} e^{-\frac{(y-z\alpha+\beta k)^2}{2\sigma^2}},
\]

for some given values of \( \alpha \) and \( \beta \). So, we have to study the function \( f_{\sigma,\zeta}(y) \) since, obviously, we have

\[
\vartheta(\lambda) \propto f_{\sigma,\zeta}(y - \alpha \lambda).
\]

We will show now that ML decoding is equivalent to Diophantine approximation

\[
\arg \max_{\lambda \in \Lambda} \vartheta(\lambda) = \arg \min_{\lambda, k \in \mathbb{Z}} |y - \alpha \lambda - \beta k|.
\]  

(16)

To do this, we specialize (15) to the 1-D case:

\[
f_{\sigma,\zeta}(y) = \frac{1}{\sqrt{2\pi \sigma}} \sum_{k \in \mathbb{Z}} e^{-\frac{(y-z\zeta)^2}{2\sigma^2}} = \sum_{k \in \mathbb{Z}} e^{-2\sigma^2 \pi^2 k^2} e^{2\pi i k \zeta}.
\]  

(17)

In the 1-D case, the Jacobi theta function is a function defined for two complex variables \( z \) and \( \tau \):

\[
\theta(z; \tau) = \sum_{k = -\infty}^{\infty} e^{\pi i k^2 \tau + 2\pi i k z} = 1 + 2\sum_{k=0}^{\infty} e^{\pi i k \tau} \cos(2\pi k z)
\]  

(18)

where \( z \) can be any complex number and \( \tau \) is confined to the upper half-plane. Comparing (17) and (18), we recognize that

\[
f_{\sigma,\zeta}(y) = \theta(z; \tau), \quad \text{for} \quad z = y, \tau = 2\pi i \sigma^2.
\]  

(19)

By using the Jacobi theta function, we have,

**Proposition 4.** \( f_{\sigma,\zeta}(y) \) is monotonically decreasing for \( y \in (0, 1/2) \).

**Proof:** Let \( r = e^{\pi i \tau} \) and \( z = y \). We fix \( r \) so that \( \theta(y; r) = \theta(y/r) \) is a function of \( y \) for \( r = e^{-2\pi^2 \sigma^2} \). Using the triple-product property of the Jacobi theta function [10]

\[
\theta(y; \tau) = \prod_{m=1}^{\infty} \left( 1 - r^{2m} \right) \left( 1 + e^{2\pi y} r^{2m-1} \right) \left( 1 + e^{-2\pi y} r^{2m-1} \right),
\]

(20)

we obtain an alternative expression of the Gaussian measure

\[
f_{\sigma,\zeta}(y) = \prod_{m=1}^{\infty} \left( 1 - r^{2m} \right) \left( 1 + 2 \cos(2\pi y) r^{2m-1} + r^{4m-2} \right).
\]  

(21)

Obviously, \( f_{\sigma,\zeta}(y) \) is periodic in \( y \) with period \( 1 \), and its monotonicity is the same as that of \( \cos(2\pi y) \).

**Corollary 5.** The Diophantine-approximation decoder (16), intuitively conjectured in [5], is in fact optimum in 1-D.

**Proof:** Follows directly from proposition 4. More precisely, \( f_{\sigma,\zeta}(y) \) is monotonically decreasing as \( y \) gets away from an integer \( a \), until it reaches the minimum \( a \pm 0.5 \).

**C. The Multidimensional Case**

The multidimensional case requires to be able to derive product formulas from Jacobi theta series and show that the function \( f_{\sigma,\zeta}(y) \) has a regular behavior (decreases when the distance between \( y \) and a lattice point increases) inside a Voronoi cell, or inside a certain region centered at a lattice point. If it is true (and we conjecture it is), then we have

**Conjecture 6.** Maximizing the ML metric is equivalent or close to solving a multidimensional inhomogeneous Diophantine approximation problem. More precisely,

\[
\arg \max_{\lambda \in \Lambda} \vartheta(\lambda) \simeq \arg \min_{\lambda \in \Lambda, q \in \mathcal{L}} \| u(\lambda) - q \|^2.
\]  

(22)

Even if we are unable to prove the conjecture for now, there are evidences supporting a decoder like (22). Since \( f_{\sigma,\zeta}(y) \) is the Fourier transform of a positive measure, we have

**Proposition 7.** \( f_{\sigma,\zeta}(y) \) is a positive-definite function. In particular, \( f_{\sigma,\zeta}(y) \leq f_{\sigma,\zeta}(0) \).

Moreover, it is easy to see

**Proposition 8.** The first derivative \( f'_{\sigma,\zeta}(y) = 0 \), while the second derivative \( f''_{\sigma,\zeta}(y) < 0 \), when \( y \in \mathcal{L} \).

Therefore, \( f_{\sigma,\zeta}(y) \) is bell-shaped in the vicinity of a lattice point, which is a local (and also global) maximum. The Diophantine-approximation decoder is a good approximation with respect to this monotonic decoding metric (at least locally).

We close this investigation with a remark on the different roles played by Diophantine approximation in this paper and in [2]. Diophantine approximation was a detrimental effect in [2], while it plays a constructive role in this paper. We expect it will achieve a better performance in compute-and-forward.

**REFERENCES**


Analysis of the Stochastic Channel Model by Saleh & Valenzuela via the Theory of Point Processes

Morten Lomholt Jakobsen, Troels Pedersen and Bernard Henri Fleury {mlj,troels,bfl}@es.aau.dk
Section Navigation and Communications, Dept. of Electronic Systems, Aalborg University
Fredrik Bajers Vej 7B, DK-9220 Aalborg East, Denmark

Abstract—In this paper we revisit the classical channel model by Saleh & Valenzuela via the theory of spatial point processes. By reformulating this model as a particular point process and by repeated application of Campbell’s Theorem we provide concise and elegant access to its overall structure and underlying features, like the intensity function of the component delays and the delay-power intensity. The flexibility and clarity of the mathematical instruments utilized to obtain these results lead us to conjecture that the theory of spatial point processes provides a unifying mathematical framework to define, analyze, and compare most channel models already suggested in literature and that the powerful tools of this framework have not been fully exploited in this context yet.

I. INTRODUCTION

Literature regarding channel models for (indoor) radio propagation dates back earlier than 1960, and most commonly the wireless multipath channel is characterized via its (time and space varying) impulse response [1]. Two classic and seminal contributions within channel modeling are those by Turin et al. [2] and Saleh & Valenzuela [3]. To some extent the (indoor) model by Saleh & Valenzuela can be seen as a generalization of the (urban) model by Turin. Specifically, the generalization aimed at mimicking cluster alike behavior since this effect was reported to have been observed experimentally.

Ever since the model by Saleh & Valenzuela (for short the S-V model) was proposed in 1987, many refined or marginally extended variants have appeared, see e.g. [4] and [5]. Unfortunately, these channel models have not been developed within any unifying mathematical framework. Instead their treatment is of rather ad-hoc nature and, as a result, their inherent features remain essentially veiled and any two different models are not easily comparable.

Recently the authors of [6] and [7] reformulated and outlined the S-V model in terms of marked point processes. The S-V model has also been revisited in [8] by use of shot-noise tools and point process theory. Among other things the analysis in [7] and [8] show that the overall intensity of the relative delays of multipath components grows linearly with the propagation delay. Unfortunately, the mathematical tools used in [7] to extract the features of the model are not directly associated with the general theory of point processes. On the other hand, the tools used in [8] are rather advanced and the derivations less transparent. Accordingly, the potential theoretical benefits arising through these point process reformulations are not immediately evident.

In this paper we showcase how the general theory of spatial point processes provides an insightful view upon the inherent structure and features of the classical S-V model. Like [7] and [8] we revisit the model and reformulate it as a particular point process. Aligned with [7] we show that the component delays consist of the union of a Poisson point process and a Cox point process and we derive the associated intensity function as an immediate consequence of Campbell’s Theorem. The derivation in [7] is similar but with no reference to Campbell’s Theorem. Furthermore, and in contrast to the involved proofs relying on shot-noise tools in [8], we obtain the delay-power intensity in a simple and direct way by invoking once more Campbell’s Theorem. These results demonstrate the potential of this well-known theorem from the theory of spatial point processes in the context of stochastic channel modeling. In view of this, our conclusion is that the theory of spatial point processes and its powerful tools have not been fully exploited yet to analyze the properties of most proposed stochastic channel models. This theory appears to provide the necessary unifying framework for which these models can be contrasted within.

II. POINT PROCESS FRAMEWORK

We assume familiarity with the basics of the theory of spatial point processes (see [9, Sec. 1.3, Chap. 2] and [10, Sec. 1.5, 6.2] for highly recommendable introductions). Concepts from abstract measure theory will be kept at a minimum.

A. Locally finiteness and simplicity

Denote by $Y$ a locally finite and simple point process defined on a $d$-dimensional space $S \subseteq \mathbb{R}^d$. For intuitive, practical and mathematical reasons, these two properties are convenient to impose since several technical aspects can then be disregarded. A point process is locally finite if the number of points falling in every bounded Borel set $B \subseteq S$ is almost surely finite. A point process is simple if, almost surely, no two points of the process coincide. Accordingly, any realization of the point process $Y$ can be identified as a countable set of points $\{y_1, y_2, \ldots, y_i, \ldots\}$, $y_i \in S$, where the index $i$ of $y_i$ serves solely as a dummy label. Thus, the index is used only to distinguish points and to indicate countability. It does not indicate any ordering of the points.

B. The intensity function and Campbell’s Theorem

Consider the counting function defined, using a generic indicator function $\mathbb{I}[\cdot] \in \{0, 1\}$, as

$$N_Y(B) := \sum_{y \in V} \mathbb{I}[y \in B],$$

which equals the random number of points from $Y$ falling in the set $B$. For any fixed and bounded $B$, the count $N_Y(B)$ is
a non-negative integer-valued random variable. The expected value of the counting function $\mu_\nu(B) := \mathbb{E}[N_\nu(B)]$ defines a measure on $S$, the so-called intensity measure of $Y$. If the intensity measure can be expressed as

$$\mu_\nu(B) = \int_B \varrho_\nu(y)dy, \quad B \subseteq S,$$

for a locally integrable function $\varrho_\nu : S \to [0, \infty)$, then $\varrho_\nu$ is called the intensity function of $Y$. The case when the intensity function exists is by far the most important for applications [11]. The importance of the intensity function is evident from the following result, often referred to as Campbell’s Theorem.

**Campbell’s Theorem.** Let $Y$ be a point process on $S \subseteq \mathbb{R}^d$ with intensity function $\varrho_\nu$. Then for a real or complex-valued measurable function $h : S \to \mathbb{R}$ (or $\mathbb{C}$), the random variable

$$\sum_{y \in Y} h(y)$$

has expected value

$$\mathbb{E}\left[ \sum_{y \in Y} h(y) \right] = \int_S h(y)\varrho_\nu(y)dy,$$

provided that the integral on the right exists.

Proofs with varying degrees of detail can be found in [9, Sec. 3.2], [11, Prop. 4.1] and [12, Thm. 2.2]. Often, the theorem is stated only for non-negative functions $h$, since the equality in (1) is then unconditionally true, i.e. the integral is always well-defined but possibly divergent. When $h$ is real-valued some care must be taken since the integral at the right hand side of (1) has no meaning if the positive and the negative part of $h$ are not integrable. Similar care must be taken for complex $h$.

**C. Poisson and Cox point processes**

We now define two classes of point processes which are particularly important for our treatment in the forthcoming section, namely Poisson point processes and Cox point processes. These definitions can be found in many text books covering the theory of spatial point processes. Our treatment is directly inspired by [11] and the interested reader may consult [10]–[12] for further details.

**Definition.** A point process $Y$ on $S \subseteq \mathbb{R}^d$ is called a Poisson point process with intensity function $\varrho_\nu$ if:

(i) For any $B \subseteq S$ with $\mu_\nu(B) = \int_B \varrho_\nu(s)ds < \infty$ the count $N_\nu(B)$ is Poisson distributed with mean $\mu_\nu(B)$.

(ii) Given that $N_\nu(B) = n \in \mathbb{N}$ where $0 < \mu_\nu(B) < \infty$, the distribution of $Y \cap B$ is the same as that of $n$ points drawn i.i.d. according to $\varrho_\nu$, where

$$f_n(s) := \frac{\varrho_\nu(s)I[s \in B]}{\mu_\nu(B)}.$$

We write $Y \sim \text{PoissonPP}(S, \varrho_\nu)$.

**Definition.** Let $Z(s), s \in S$, be a non-negative random field such that, almost surely, every realization of $Z$ is a locally integrable function on $S$. If a point process $Y$, conditioned on $Z$, is a Poisson point process with intensity function $Z$, then $Y$ is called a Cox point process driven by $Z$.

Cox point processes (also often referred to as doubly stochastic Poisson point processes [10]) are flexible models for clustered point patterns. Specifically, the two-level construction most commonly entails the Cox class to exhibit so-called over-dispersion compared to the Poisson class [11, Sec. 5.2].

**III. The Model by Saleh & Valenzuela**

In this section we analyze the impulse response of the classical S-V model within the framework of spatial point processes. The main purpose of this effort is to straightforwardly derive the features of this model through a flexible and powerful theory. Several relevant aspects of the model are revealed through this reformulation, e.g. its overall delay intensity, a concise and clear derivation of the average power gain and, a simple derivation of the delay-power intensity as well.

**A. Classical formulation**

Saleh & Valenzuela define the channel impulse response with cluster and within-cluster delays as [3, Eq. (25)]

$$h(t) = \sum_{\ell=0}^{\infty} \sum_{k=0}^{\infty} \beta_{k,\ell} \exp(j\theta_{k,\ell}) \delta(t - (T_{\ell} + \tau_{k,\ell})), \quad \ell \neq \infty,$$

where $\delta$ is the Dirac delta and $j$ is the imaginary unit. The index $\ell$ indicates a certain cluster and $k$ is the within-cluster index. By definition in [3], $T_0 = 0$ and $\tau_{0,\ell} = 0$ for each $\ell \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$. Beside these fixed delay components, a sequence of Poisson point processes are suggested such that:

- $\{T_{\ell}\}_{\ell \in \mathbb{N}} \sim \text{PoissonPP}(\mathbb{R}_+, \Lambda)$
- $\{\tau_{k,\ell}\}_{k \in \mathbb{N}} \sim \text{PoissonPP}(\mathbb{R}_+, \lambda)$ for each $\ell \in \mathbb{N}_0$.

Here, $\Lambda, \lambda > 0$ being two parameters. Moreover, conditional second-order moments are modeled such that [3, Eq. (26)]

$$\mathbb{E}[\beta_{k,\ell}^2 | T_{\ell}, \tau_{k,\ell}] = Q \exp(-T_{\ell}/\Gamma) \exp(-\tau_{k,\ell}/\gamma),$$

with $\Gamma, \gamma > 0$ and $Q$ being the average power gain of the first component within the first cluster (i.e. corresponding to the fixed delay $T_0$). Conditioned on all $T_{\ell}$'s and all $\tau_{k,\ell}$'s, the $\beta_{k,\ell}$'s are assumed to be mutually independent random variables. Specifically, each power gain $\beta_{k,\ell}$, conditioned on $T_{\ell}$ and $\tau_{k,\ell}$, should follow an exponential distribution with mean parameter decaying as described by (3). Fig. 1 illustrates the Poisson point processes involved in the S-V model.

Finally, it was mentioned in [3] that practically the doubly-infinite sum in (2) should "stop" whenever each of the exponentially decaying terms in (3) had become small enough. Through the insight gained via the forthcoming reformulation of this classical channel model we are able to motivate a less heuristic "stopping criterion".

**B. Point process formulation**

Naturally, we select the space $S = \mathbb{R}_+$ and let $T_0 = 0$ as above. In addition, we introduce the point processes:

- $C := \{T_{\ell}\}_{\ell \in \mathbb{N}}$ (all cluster delays except $T_0$)
- $W_{\ell} := \{T_{\ell} + \tau_{k,\ell}\}_{k \in \mathbb{N}}$ (delays within the $\ell$'th cluster)
- $W := \bigcup_{\ell=0}^{\infty} W_{\ell}$ (all within-cluster delays)
- $Y := C \cup W$ (all propagation delays except $T_0$).
Notice that $C$ is the Poisson point process specified at first in the previous paragraph. Its intensity function has a simple form, namely $\varrho_C(t) = \Lambda$ for all $t \in S$. By conditioning, we immediately identify a sequence of Poisson point processes\footnote{\[\text{Note that in (6) we abuse notation since the collections $W_c$ are not explicitly defined. We only defined these as $W_s$ via the counting index $s$.}\]}

$$W_i|T_i \sim \text{PoissonPP} \left( R_i, \lambda [t > T_i] \right), \quad \ell \in \mathbb{N}_0,$$

and since the Poisson class is stable with respect to countable superpositions \cite[Prop. 3.6]{11, chapter 5.2.3}, we see that

$$W|[C \sim \text{PoissonPP} \left( R_s, \bar{\varrho}_W \right),$$

with the staircase-alike intensity function

$$\bar{\varrho}_W(t) = \lambda + \lambda \sum_{c \in C} \mathbb{1}[t > c], \quad t \in S. \quad (4)$$

Accordingly, we identify that the point process $W$, without conditioning on $C$, is a Cox point process driven by a stochastic process $Z$ having the same functional form as $\bar{\varrho}_W$ in (4) but with $C$ being random. The intensity function of the Cox point process $W$ is $\varrho_v(t) = \mathbb{E}[Z(t)]$ \cite[Sec. 5.2.3]{11}, and by direct application of Campbell’s Theorem we get

$$\varrho_v(t) = \lambda + \lambda \mathbb{E} \left[ \sum_{c \in C} \mathbb{1}[t > c] \right] = \lambda + \lambda \Lambda t, \quad t \in S.$$

Since $Y = C \cup W$ is an union of almost surely disjoint point processes, its associated intensity function reads \cite[Sec. 6.2.3]{10}

$$\varrho_Y(t) = \varrho_v(t) + \varrho_C(t) = \lambda + \lambda + \lambda \Lambda t, \quad t \in S.$$

It is interesting to notice that the entire set of propagation delays (excluding the first component $T_0$) is the union of a Poisson point process and a Cox point process. Of course, the realization of $W$ depends upon the realization of $C$, i.e., these two point processes are not independent. In [7] this interpretation was inherently adopted, without being explicitly mentioned. Another interesting yet observed fact is that the intensity function $\varrho_Y$ rises linearly with propagation delay, see Fig. 2. The jump of height $\Lambda + \lambda$ at $T_0 = 0$ in the graph of $\varrho_Y$ appears due to the cluster delays and the delays within the very first cluster. The term $\lambda \Lambda t$ result from the fact that, on average, a total of $\Lambda t$ additional clusters emerge during the interval $[0, t]$, with each and every one of them spawning further delay components at rate $\lambda$.
as was also reported in a footnote in [3]. Yet, the original sequence of arguments used to obtain this result may appear less instructive, see [3, Eq. (27), (31)] for comparison. Notice that, depending on how we choose to write out the product in (7), we end up with different interpretations of individual average power contributions.

D. Delay-power intensity

Motivated by the definition of $G$ in (5) together with the relationship in (7), we consider

$$p(t) := \sum_{t=0}^{\infty} \sum_{k=0}^{\infty} \beta_{k,t} \delta(t - (T_k + \tau_{k,t})).$$

We wish to calculate how the average power gains are distributed across delay. From (7) we already know the mean total power gain, yet we seek to obtain further insight. The above definition of $p(t)$ is motivated by the fact that $\int_0^\infty p(t) dt = G$, and since $\mathbb{E}[G]$ is finite, the non-negative random variable $G$ is itself finite almost surely. Accordingly, we define

$$P(t) := \mathbb{E}[p(t)], \ t \in S,$$

and we refer to this function as the delay-power intensity. By similar manipulations as in the previous paragraph (conditioning, Campbell’s Theorem, etc.) we find

$$\frac{P(t)}{Q} = \delta(t) + \left\{ \begin{array}{ll}
    k_1 \exp \left( -\frac{t}{\gamma} \right) + k_2 \exp \left( -\frac{t}{\Gamma} \right), & \Gamma, \gamma \neq \gamma \\
    \varphi_v(t) \exp \left( -\frac{t}{\Gamma} \right), & \Gamma = \gamma
    \end{array} \right.$$

where we have conveniently introduced the two constants

$$k_1 := \Lambda \left( 1 + \lambda \frac{\gamma}{\Gamma} \right) \quad \text{and} \quad k_2 := \lambda \left( 1 - \Lambda \frac{\gamma}{\Gamma} \right).$$

The same expression for $P(t)$ is obtained in [8, Chap. 2.3] using rather involved shot-noise tools with weighty notational overhead. Notice the particular relationship

$$\mathbb{E}[G] = \mathbb{E} \left[ \int_0^\infty p(t) dt \right] = \int_0^\infty \mathbb{E}[p(t)] dt = \int_0^\infty P(t) dt.$$

The delay-power intensity of the S-V model is depicted in Fig. 3. Notice that $P(t)$ is not exponentially decaying, not even when $\Gamma = \gamma$ since $\varphi_v$ rises linearly (compare with the dotted line in Fig. 3).

Finally, as mentioned in the beginning of this section, we are now able to make a simple "stopping criterion" suitable, e.g. for simulation purposes. Specifically, select a delay threshold $t_{\text{max}}(\alpha)$ such that

$$\int_0^{t_{\text{max}}(\alpha)} P(t) dt = \alpha \mathbb{E}[G],$$

for a relevant choice of $\alpha \in (0, 1)$, e.g. $\alpha = 0.99$.

IV. CONCLUSION

In this contribution we have revisited the radio channel model by Saleh & Valenzuela (the S-V model) within the framework of spatial point processes. We have shown that the component delays in the S-V model emerge from the union of a Poisson point process and a Cox point process. Furthermore, we have demonstrated that the intensity function of the component delays and the delay-power intensity can be derived in a straightforward and rigorous manner as an immediate consequence of Campbell’s Theorem.

The above results indicate that the theory of spatial point processes yields a natural, unifying theoretical framework for dealing with stochastic channel models. This applies in particular to most channel models already suggested in literature, including the models by Turin et al. [2], Spencer et al. [4], and Chong et al. [5]. Our results also reveal that the powerful tools available in this framework, like Campbell’s Theorem, have not been exploited to their full extent in this context yet. Overall the considered application to the S-V model and to some extent the work in [6]–[8] show that the resulting mathematical treatments inherit clarity and conciseness, in addition to rigorousness, in contrast to the traditionally used ad-hoc and heuristic arguments.

ACKNOWLEDGMENT

This work has been funded by the project ICT-248894 Wireless Hybrid Enhanced Mobile Radio Estimators 2 (WHERE2), the 4GMCT cooperative research project co-financed by Intel Mobile Communications, Agilent Technologies, AAU and the Danish National Advanced Technology Foundation, and by the Danish Agency for Science, Technology and Innovation.

REFERENCES

Low-Complexity Soft-Output Detection for Impulse-Radio Ultra-Wideband Systems Via Combining Multiple Observations

Andreas Schenk$^a$ and Robert F.H. Fischer$^b$

$^a$Lehrstuhl für Informationsübertragung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany,
$^b$Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany,
Email: andreas.schenk@int.de, robert.fischer@uni-ulm.de

Abstract—Avoiding explicit channel estimation, autocorrelation-based receivers enable low-complexity detection of impulse-radio ultra-wideband (IR-UWB) signals. In this paper, we present power-efficient IR-UWB detection schemes, which process the receive symbol stream partitioned into overlapping blocks. Properly combining the individual observations gained from each blockwise detection, a final (soft)-decision with improved reliability is obtained. By means of numerical simulations we show that the presented approach enables a reduction of the hardware complexity of the autocorrelation receiver at no degradation in performance and only slightly increased computational complexity, and is perfectly suited especially for coded transmission.

I. INTRODUCTION

Noncoherent receivers are widely regarded as the key to low-complexity detection in impulse-radio ultra-wideband (IR-UWB) systems, which are especially well-suited for short-range low-data-rate communications, such as, e.g., in wireless-sensor networks [1]. The inherent loss in performance of conventional noncoherent autocorrelation-based differential detection (DD), as compared to coherent detection based on explicit channel estimation, can be alleviated by advanced autocorrelation-based detection schemes, which process blocks of receive symbols jointly, i.e., perform multiple-symbol differential detection (MSDD) [2]–[4] or decision-feedback differential detection (DF-DD) [5]. However, significant challenges for implementation of the autocorrelation receiver (ACR) front-end are induced, as delaying the (analog) receive signal over several symbol durations is required to compute the correlation coefficients required for these detection schemes.

In this paper, we propose to obtain multiple observations of the same symbol by processing overlapping blocks of symbols. The resulting information gained from these individual observations is then combined to obtain a final decision with improved reliability. Depending on the choice of the blockwise detection scheme (we consider MSDD and DF-DD) and the method, how the individual observations are combined, the presented scheme is able to deliver hard and soft decisions, and is thus suited for coded and uncoded transmission. The benefit of the presented scheme is to reduce the hardware implementation complexity of the ACR front-end at no degradation in performance.

The proposed scheme can be seen as a development of subset MSDD introduced for noncoherent detection of differential phase-shift keying (DPSK, or more general differential space-time modulation) [6], [7], where similarly multiple observations are generated with an overlapping block-structure. However, as opposed to our approach, in [6], [7] most of the observations are simply treated as unreliable, thus discarded, and only a single observation is processed further.

After defining the IR-UWB system model and briefly reviewing autocorrelation-based detection in Sec.II, multiple-observations combining based on blockwise detection is introduced in Sec. III. The effectiveness of the approach is validated in Sec. IV by means of numerical simulations of uncoded and coded transmission; final remarks conclude the paper.

II. AUTOCORRELATION-BASED DETECTION OF IR-UWB

A. IR-UWB System Model

We consider differentially encoded binary pulse-amplitude-modulated IR-UWB transmission. Assuming no inter-symbol-interference (i.e., the symbol duration $T$ is chosen sufficiently large, such that each pulse has decayed before the next pulse is received) and a time-invariant multipath channel, the sampled receive signal in the $k$th symbol interval may compactly be written as an \((N_s = f_s T)\)-dimensional vector

$$\vec{r}_k = b_k \vec{p} + \vec{n}_k$$

(1)

where $f_s$ is the sampling frequency, $b_k$ are differentially encoded information symbols $a_k \in \{\pm 1\}$ (i.e., $b_k = b_{k-1} a_k$ with $b_0 = 1$), and $\vec{p}$ is the (sampled) receive pulse shape, resulting from the convolution of the transmit pulse, multipath channel and receive filter impulse response. Neglecting the reference symbol, after suitable normalization, the energy per symbol is given as $E_s = 1$. $\vec{n}_k$ collects the samples of the zero-mean additive Gaussian noise, each with variance $\sigma_n^2 = f_s N_0/2$, where $N_0/2$ is the two-sided power-spectral density of the underlying additive white Gaussian noise.

B. Autocorrelation Receiver Front-End

Avoiding explicit channel estimation, autocorrelation-based detection computes the correlation of the current symbol with up to $L$ preceding symbols, cf., e.g., [2]–[5], [8]. Significant gains are achieved by adapting the correlation interval $T_i$ to the channel characteristics at hand [1], i.e., considering only...
the relevant part $r_k$ of $\tilde{r}_k$ for autocorrelation by selecting $N = T_s f_c$ successive elements (typically the first) out of $N_0$. This effectively reduces the time-bandwidth product to $N = f_c T_s$. Thus, the output of the $L$-branch ACR for the $k$th symbol is given as, for $l = 1, \ldots, L$,

$$z_{k-l,k} = r_{k-l}^T r_k = b_{k-l}^T b_k + \eta_{k-l,l}. \tag{2}$$

Assuming that the correlation interval captures the entire pulse energy (i.e., due to normalization $p^T p = 1$, $z_{k-l,k}$ is composed of the phase transition from $b_{k-l}$ to $b_k$, and “information x noise” and “noise x noise” terms, summarized in the equivalent noise term $\eta_{k-l,l} = b_{k-l}^T p n_k + b_k n_{k-l}^T p + n_{k-l}^T n_k$. It has been shown that already for moderate time-bandwidth products $N$ this equivalent noise is well approximated as uncorrelated zero-mean Gaussian random variables with variance $\sigma_n^2 = 2\sigma_s^2 + N(\sigma_n^2)^2$ [8]. For clarity, we note that, given the assumptions above, this Gaussian-approximated discrete-time equivalent model is independent of the actual pulse shape, i.e., the channel realization [9]. In this paper, all simulations of the IR-UWB system use this equivalent model.

The correlation coefficients serve as input for various detection schemes [1]–[5], [8], the most simple being symbolwise simulations of the IR-UWB system use this equivalent model. In this paper, all that, given the assumptions above, this Gaussian-approximated log-likelihood ratios (LLRs), soft output for the $k$th (binary) information symbol is given as

$$\text{LLR}_k = \frac{1}{\sigma_n^2 (L + 1)} \left[ a_k^{\text{MSDD}} (\Lambda^\text{MSDD} - \Lambda^\text{MSDD}) \right] \tag{5}$$

where the decision metric is defined as

$$\Lambda(\tilde{a}) = \sum_{k=1}^{L} \left( \sum_{l=0}^{k} \left( z_{l,k} - b_l \tilde{b}_k z_{l,k} \right) \right) \tag{4}$$

and $\tilde{b}_k$, $0 \leq k \leq L$, and $z_{l,k}$, correspond to the differentially encoded hypothesis $\tilde{a}$, and the correlation coefficients (cf., (2)) of the considered block, respectively.

Soft decisions, i.e., reliability information on the decision of each symbol, is obtained by extending MSDD to so-called soft-output MSDD (SO-MSDD) [4]. In terms of (max-log approximated) log-likelihood ratios (LLRs), soft output for the $k$th (binary) information symbol is given as

$$\text{LLR}_k = \frac{1}{\sigma_n^2 (L + 1)} \left[ a_k^{\text{MSDD}} (\Lambda^\text{MSDD} - \Lambda^\text{MSDD}) \right] \tag{5}$$

where $\Lambda^\text{MSDD} = \Lambda(a^{\text{MSDD}})$ denotes the metric associated with the MSDD estimate (3), and

$$\Lambda_k^{\text{MSDD}} = \min_{\tilde{a}_k \in \{\pm 1 \}} \Lambda(\tilde{a}) \tag{6}$$

the decision metric of the corresponding counter hypothesis, i.e., the minimum metric with the restriction $\tilde{a}_k = -a_k^{\text{MSDD}}$.

Utilizing the triangular structure of the decision metric, MSDD can be viewed as a tree-search-problem, which in turn can efficiently be solved employing the single-tree-search soft-output sphere decoder (SO-SD) [10], [4]. To further reduce the computational complexity of the SO-SD in particular at low SNR, LLR clipping during the SO-SD search process (clipping level LLR$_{\text{max}}$) [10], [4] is applied in addition. In this case, we employ an optimized decision order within the search process. This sorting of the SD input is based on the same optimization criterion as of DD (cf. Sec. III-A2 and [5]), and requires some modifications to the original SO-SD algorithm [4]. Minimum throughput of the MSDD device, despite the varying complexity of the SD, can effectively be guaranteed by introducing a cumulative maximum complexity limit (e.g., per codeword) similar to [10]. Due to space limitations, these techniques are not described in greater detail.

2) Decision-Feedback Differential Detection: A closely-related detection scheme is blockwise DD [5], which decides the symbols within each block in a successive manner taking into account the feedback from already decided symbols within the block. In [5] it has been shown that DD achieves close-to-optimum performance at very low, and in particular constant computational complexity, if the decision order is optimized, such that in each step the most reliable symbol is decided next (taking into account the actual correlation coefficients and the previous decisions, see [5] for details).

B. Obtaining Multiple Observations

In the related context of noncoherent detection of DPSK (and more general unitary differential space-time modulation), in [6], [7] it has been proposed to employ MSDD in combination with an overlapping block-structure. Since multiple blocks thus contain the same symbol, processing of each block delivers (possibly different) beliefs on the same symbol, i.e., multiple observations are available. Observing that the reliability of the estimates degrades at the block edges for
combining selection
received symbols
final (soft) decisions
reliability
symbol index $k$
observations block-based
$s = 1s \ldots$ blocksize $L = 3$, 5 and 10. We consider the two detection
schemes combining multiple soft decisions obtained from SO-
traditional single observation per symbol for
observations for each symbol. Thus, we obtain on average
note that, in the context of MSDD of IR-UWB, significantly
multiple observations to deliver a final (soft) decision for
side of Fig. 1. There are different options how to combine
final decision. The basic concept is illustrated on the right hand
pose to suitably combine all observations obtained from pro-
edges are treated as unreliable and are simply discarded.
Fig. 1. Illustration of subset MSDD (left) and multiple-observations com-
posing (right). Blocksize $L = 3$, maximum overlap (mutual shift $s = 1$).

DPSK [6], [7], only the middle part of each block is processed
further, so-called subset MSDD, as illustrated in the left part of
Fig. 1. Already available observations of symbols at the block
edges are treated as unreliable and are simply discarded.

Avoiding to discard already available observations, we propo-
tose to suitably combine all observations obtained from pro-
cessing of each block, resulting in a (possibly more reliable)
final decision. The basic concept is illustrated on the right hand
side of Fig. 1. There are different options how to combine
multiple observations to deliver a final (soft) decision for the
respective symbol, as discussed in Sec. III-C. In addition
note that, in the context of MSDD of IR-UWB, significantly
different reliabilities within each block cannot be observed.
and, as we employ an optimized decision order within each
block (cf., [5] and Sec. III-A), the positions of the most reliable
symbols are no longer limited to the block center.

Of course, the generation of multiple observations, i.e.,
processing of overlapping blocks, leads to an increased computa-
tional complexity and processing delay; however, the analog
ACR front-end remains unchanged. As we aim for a flexible
tradeoff between performance and computational complexity,
we vary the mutual shift $s = 1, \ldots, L$ of adjacent blocks (i.e.,
overlap of $L - s$ symbols) to adjust the number of multiple
observations for each symbol. Thus, we obtain on average $L/s$
observations per symbol (for finite symbol streams this value
can be ensured by reducing the blocksize at the beginning
and at the end), with maximum overlap for $s = 1$, and the
traditional single observation per symbol for $s = L$. Note that
blocksize and overlap are defined based on the information
symbols; the respective blocks of $L + 1$ differentially encoded
symbols overlap by at least one symbol.

C. Combining Multiple Observations

Depending on the employed blockwise detection scheme,
different options can be considered how multiple observations
are combined to a final estimate of the information symbol.

1) Combining Multiple Soft Decisions: If soft decisions are
available for each individual observation, as obtained from SO-
MSDD, we may consider these observations as independent
(though independence cannot be assured), and obtain the
resulting final estimate according to the principle of parallel
information combining [11]. In terms of LLRs, the resulting
LLR for each symbol is then directly given as the sum of the
corresponding individual LLRs.

In the case of coded transmission, the resulting LLR can
directly be passed to the subsequent (soft-input) channel
decoder. Since LLR clipping affects each SO-SD run, the
maximum absolute input value is $L \cdot \text{LLR}_{\text{max}}$. Final hard-
decisions can be obtained by evaluating only the sign of the
sum of the LLRs, corresponding to combining the individual
hard decisions according to their reliability.

2) Combining Multiple Hard Decisions: Based on multiple
hard decisions, here obtained from DF-DD (of course, hard-
output MSDD is also possible, but not considered for brevity),
a final hard decision is directly given following the majority
of the individual decisions (i.e., evaluating the sign of the
sum of the individual estimates), which effectively treats each
observation with equal reliability (in case of a draw, we
choose $\pm 1$ at random). Clearly, this strategy leads to improved
performance only if more than two observations are available
per symbol, i.e., only if $s < L/2$.

However, the most interesting option is to combine multiple
hard decisions of the same symbol to form a single soft
decision. Exemplarily, consider the case that three (possibly
different) hard decisions are available. If these coincide, the
reliability of this symbol is three times as large as the
reliability of only a single observation, whereas if only two
coincide and one differs, the reliability is reduced to the case
of a single observation. This method can be implemented by
using the sum of the individual hard-decisions as (quantized
and scaled) “soft-output”. Depending on the employed channel
decoder, scaling of these soft-output values may be neglected
if the decoding metric is scale-invariant (e.g., in a convolu-
tional coded system employing the Viterbi algorithm for soft-
decision decoding). However, e.g., in an LDPC-coded system,
the soft output should be scaled to match the decoder input
range.

IV. PERFORMANCE AND COMPLEXITY

Modeling a realistic IR-UWB scenario [1], for numerical
simulations we set the time-bandwidth product to $N = 400$,
corresponding, e.g., to an ACR correlation interval of $T_i = 33$
ns and a Nyquist frequency of 6 GHz ($f_s = 12$ GHz).
Typically, the symbol duration is chosen larger (e.g., $T = 75$
ns) to preclude inter-symbol interference.

We evaluate the effectiveness of the presented approach
based on the key figures performance (i.e., the required energy
per bit $E_b$ vs. spectral noise density $N_0$) and computational
complexity at a desired bit error rate (BER) operating point.
As a measure for the computational complexity of MSDD
we use the average number of visited nodes per symbol of the
SO-SD tree-search process. Due to LLR clipping and the
optimized decision order, its variance is relatively low and
can effectively be upper-bounded by introducing a cumulative
maximum complexity limit. On the contrary, DF-DD has
constant complexity, which, using its relation to SD-based tree
search [5], corresponds to a single node per symbol.

First, we consider uncoded transmission ($E_b = E_o$), i.e.,
multiple observations are combined to yield a final hard
decision. Fig. 2 depicts the tradeoff performance vs. (average)
complexity at an operating point of $\text{BER} = 10^{-3}$, obtained
by adjusting the mutual shift of the blocks ($s = 1, \ldots, L$) for
a blocksize $L = 3, 5$ and 10. We consider the two detection
schemes combining multiple soft decisions obtained from SO-
MSDD (sign of sum of LLRs), and combining multiple hard decisions obtained from DF-DD (sign of sum of estimates). For SO-MSDD, the LLR clipping level is set to $LLR_{\text{max}} = 15$. For both cases the upper-left point corresponds to the maximum block-overlap (maximum number of $L$ observations, $s = 1$), and the lower-right point to traditional detection with only a single observation per symbol (no overlap, $s = L$).

The performance of SO-MSDD improves with increasing number of combined individual soft-decisions, with gains of up to 0.8 dB compared to traditional MSDD; the increase in complexity is proportional to the average number of observations per symbol, $L/s$. For comparison, tradeoff-curves employing the SO-SD without LLR clipping are also given, which prove that LLR clipping is crucial to achieve low complexity at only slight performance degradation [4], [10].

Combining multiple hard decisions (DF-DD) improvements are only achieved for $s < L/2$, as otherwise only up to two observations are available. However, as opposed to the varying complexity of SD-based MSDD, DF-DD achieves even larger gains at lower and, in particular, fixed complexity.

Finally, the performance-complexity tradeoff is evaluated for coded transmission at $BER = 10^{-6}$ in Fig. 3. We consider a bit-interleaved coded modulation scheme employing maximum-free-distance convolutional codes (non-recursive, non-systematic, zero-padded encoder [12]) with constraint length $\nu = 4$, a random block-interleaver (size 1500 symbols), and, following the reasoning in [9], a code-rate of $R_c = 2/3$ ($E_b = E_a R_c$). As we aim for soft-decision decoding using the Viterbi algorithm, the final soft decision is obtained by combining either multiple soft decisions (sum of LLRs) in case of SO-MSDD as blockwise detection scheme, or multiple hard decisions (sum of estimates) in case of DF-DD, cf. Sec. III-C.

It can be observed that for all blocksizes each additional observation leads to notable gains. Combining of multiple observations is especially effective, if multiple hard decisions are combined to a single soft decision (DF-DD); the difference between the traditional single observation and $L$ observations, e.g., amounts to almost 2 dB for $L = 10$. Consequently, this scheme enables to effectively tradeoff low-cost computational complexity against costly implementation complexity of the ACR front-end at no degradation in performance by processing shorter, but overlapping blocks (thus, lower maximum delay required in the ACR) and combining the multiple observations.

V. CONCLUSIONS

In this paper, we have proposed a power-efficient, yet low-complexity detection scheme for autocorrelation-based IR-UWB receivers, which is based on combining multiple observations for the same symbol obtained from blockwise detection using an overlapping block structure. Our approach is able to deliver hard and soft decisions with improved reliability at only marginal increase in computational complexity, by, simplified speaking, repeating very-low-complexity steps multiple times and combining the individual observations. It is perfectly suited especially for coded transmission and enables to reduce the blocksize, and thus also the implementation complexity of the analog autocorrelation receiver front-end, at no degradation in performance.

REFERENCES


Amplify-and-Forward Versus Decode-and-Forward Relaying: Which is Better?

Georgy Levin and Sergey Loyka

Abstract—Performance of multi-hop MIMO relay channels under the amplify-and-forward and decode-and-forward protocols are compared via the capacity and SNR gains. In an $N$-hop channel of linear topology with multi-antenna source and destination and single-antenna relay nodes, the capacity gain of the DF relaying over the AF one does not exceed $\log_2 N$ bits/Hz and its SNR gain does not exceed $N$, for any channel realization. This conclusion also applies to selection relaying, to the outage probability/capacity and the ergodic capacity in an arbitrary block-fading channel, and can be further extended to hybrid relaying. The conditions under which the DF and AF relaying have nearly identical performance are identified.

I. INTRODUCTION

Due to its numerous advantages (e.g. improved coverage, throughput, system capacity, power/battery life etc.), multi-hop relaying has recently attracted significant attention in both academia [1] and industry. Most common relaying strategies are decode-and-forward (DF) and amplify-and-forward (AF). While a DF relay decodes, re-modulates and retransmits the received signal, an AF one simply amplifies and retransmits the signal without decoding. Compared to an AF relay, the complexity of a DF one is significantly higher due to its full processing capability. The DF protocol also requires a sophisticated media access control layer, which is unnecessary in the AF protocol. Overall, a DF relay is nearly as complex as a base station. Does the performance improvement of DF relaying outweigh its complexity burden? To answer this question (in part), the present paper will quantify this performance improvement.

There seems to be no consensus in the literature as to the question: What relaying protocol is better, AF or DF? While some studies find the DF protocol to be superior, others find the other way around. Indeed, it has been shown in [2] based on numerical simulations, that the AF multi-hop relaying outperforms the DF one under uncoded BPSK modulation in terms of outage probability and bit error rate (BER), which was explained by the error propagation effect in the DF relaying outweighing the noise amplification of the AF relaying. An analysis of a maximum likelihood demodulation presented in [3] for coherent cooperative diversity in uncoded BPSK systems shows that the DF relaying with more than one relay loses about half of the diversity of the AF relaying.

These studies, however, are limited to uncoded systems. Most modern communications systems rely on powerful channel codes and the uncoded results do not extrapolate directly to those systems. Indeed, it was shown in [4] that single-antenna multi-hop Rayleigh-fading relay channels under the DF protocol achieve higher ergodic (mean) capacity than under the AF one. A similar conclusion was obtained in [5] for the ergodic capacity of MIMO multi-hop relay systems. The study in [6] shows that the outage probability in multi-hop relay channels is higher under the AF protocol, which automatically transforms into smaller outage capacity. Therefore, one has to conclude that while the AF protocol is better for uncoded systems (where the error propagation effect outweighs the noise amplification), the opposite is true for systems using powerful capacity-approaching codes. Finally, the popular diversity-multiplexing framework (DMT, see e.g. [7]) has been also applied to relay channels [8][9]. We caution the reader that this framework, due to its asymptotic nature ($SNR \rightarrow \infty$), is known to provide misleading conclusions in many scenarios (including relay ones) as far as the finite-SNR performance is concerned [6][10].

In this paper, we consider coded systems via their capacity analysis (i.e. the fundamental limit to error-free data transmission) and quantify the performance superiority of the DF relaying over the AF one, for a fixed channel (or, alternatively, for a given realization of an arbitrary-fading channel), via two performance metrics: capacity and SNR gains. The main result is that, in an $N$-hop channel with multi-antenna source and destination and single-antenna relay nodes, the capacity gain of the DF relaying does not exceed $\log_2 N$ bits/Hz and its SNR gain does not exceed $N$. This conclusion also applies to selection relaying, to the outage probability/capacity and the ergodic capacity in an arbitrary block-fading channel, and can be further extended to hybrid relaying. The conditions under which the DF and AF relaying have nearly identical performance and, thus, the DF relaying is not worth the effort due to its higher complexity, are identified.

II. RELAY CHANNEL MODEL AND CAPACITY

Let us consider a multi-hop relay channel when the source (transmitter) and the destination (receiver) are equipped with multiple antennas, and the full-duplex relay nodes have a
single antenna, as shown in Fig. 1. In this paper, we assume that the channel is of linear topology, i.e., only adjacent relays can “see” each other [8] and there is no direct (source-destination) link (this is the case when relay nodes are needed most), and that the full CSI is available at the source and the destination 1.

A. Amplify-and-Forward Protocol

The signal received by i-th relay is $y_i = h_i x_{i-1} + \xi_i$, $i \geq 1$, where $x_{i-1}$ is the signal transmitted by (i-1)-th relay, $h_i$ is the i-th hop channel (between (i-1) and i-th relays), and $\xi_i$ is i-th relay AWGN noise, and we assume the baseband, discrete-time, frequency-flat channel model. The signal transmitted by i-th relay in the AF mode is $x_i = \sqrt{K_i} y_i$, where $K_i$ is its power gain. Thus, the input-output relationship of the whole N-hop AF relay channel is

$$y = \sqrt{K_{N-1}} h_N \cdot h_1^* x + \sqrt{K_{N-2}} h_{N-1} \cdot h_2 \xi_2 + \ldots$$

$$+ \sqrt{K_1} h_N \xi_{N-1} + \xi_N$$

(1)

where $x$ and $y$ are the transmitted (source) and received (destination) signal vectors, $h_1$, $h_N$ are the 1-st and last hop channels (bold and regular characters denote vectors and scalars); note that 1-st and N-th hops are vector channels, all others are scalars, $\xi_i \sim \mathcal{CN}(0, \sigma_i^2)$, $i = 1 \ldots N - 1$, $\xi_N \sim \mathcal{CN}(0, \sigma_N^2)$. $I$ is identity matrix, and $\cdot^*$ denotes Hermitian conjugation. Without loss of generality, the relay power gains $K_i$ are further absorbed into $h_2 \ldots h_N$ via the substitution $h_i \rightarrow \sqrt{K_{i-1}}h_i$.

To find the capacity of the channel in (1), consider the sufficient statistics for $y$, which is (see e.g. [6][7]),

$$z = |h_N|^{-1} h_N^* y = |h_N| \cdot h_1^* x + |h_N| \cdot h_2 \xi_2 + \ldots$$

$$+ |h_N| \xi_{N-1} + |h_N|^{-1} h_N^* \xi_N$$

(2)

where $|h|^2 = h^* h$. The 1st term represents the signal received at the destination, and the other terms represent the relay noise propagated to the destination and the destination own noise. Note that sufficient statistics preserves the mutual information and capacity, as well as error rate under maximum-likelihood decoding, so that (2) is a scalar channel equivalent to (1). From this, the SNR at the destination can be expressed as

$$g_N \ldots g_2 h_1^* R_s h_1 / \sigma_s^2 \leq \sigma_s^2 \sum_{k=1}^{N} g_k / \sigma_k^2 = \gamma_{AF}$$

(3)

where $R_s = E[xx^\dagger]$ is the covariance matrix of the source transmitted signal, $\sigma_s^2 = tr R_s = E[xx^\dagger]$ is the total power, $g_k = |h_k|^2$ is the power gain of k-th hop channel, and $\sigma_k^2 = \sigma_k^2 + \sum_{k=1}^{N-k} \sigma_k^2 \prod_{i=k+1}^{N} g_k$ is the total noise power at the destination. The inequality in (3) follows from $h_1^* R_s h_1 \leq |h_1|^2 \sigma_s^2$, and the equality is achieved when $R_s = \sigma_s^2 h_1 h_1^*/|h_1|^2$, i.e., beamforming from the source to the 1-st relay $2$, $x = s \cdot h_1/|h_1|$, where $s$ is the scalar transmitted symbol of the total power $\sigma_s^2$. From (2), achieving the upper bound in (3) also requires maximum ratio combining (MRC, also known as spatial matched filtering) at the destination.

The AF relay channel capacity, i.e., the ultimate bound to error-free data transmission under the AF protocol, can now be compactly expressed from the upper bound in (3) as

$$C_{AF} = \log(1 + \gamma_{AF}) [\text{bits/Hz}]$$

(4)

where $\gamma_{AF} = (\sum_{i=1}^{N} \gamma_i^{-1})^{-1}$ is the equivalent SNR, i.e., SNR in the AWGN channel that has the same capacity as the N-hop AF relay channel above, and $\gamma_i = g_1 \ldots g_i \sigma_i^2 / \sigma_s^2$ is the SNR at i-th relay if all previous relays were noiseless ($i = N$ corresponds to the destination), which is also the SNR at the destination when all the relays but i-th, and also the destination are noiseless. The capacity in (4) is achieved by the beamforming at the source (towards 1-st relay), the MRC at the destination and using a temporal capacity-achieving code on the equivalent scalar channel. Since $\gamma_{AF}$ is an increasing function of $\gamma_i$, each AF relay has to use the maximum available power/gain.

B. Decode-and-Forward Protocol

The system model in this case is $y_1 = h_1 x + \xi_1, y_2 = h_2 x + \xi_2, \ldots, y = h_N x + \xi_N$.

(5)

where $x$ and $y$, $i = 1 \ldots N - 1$, are the transmit and received signals at the relays. The DF relay channel capacity is limited by the weakest link,

$$C_{DF} = \min_{i=1 \ldots N} \{C_i\} = \log(1 + \gamma_{DF})$$

(6)

where $C_i = \log(1 + \gamma_i)$ is the capacity of i-th hop, $\gamma_i = g_1 \ldots g_i \sigma_i^2 / \sigma_s^2$ is the SNR at i-th relay, and $\gamma_{DF} = \min_{i=1 \ldots N} \{\gamma_i\}$ is the effective SNR of an AWGN channel of the same capacity as the DF relay channel. This is also the relay channel capacity for the linear topology in Fig. 1. To make a fair comparison, we have set the transmitted signal power at each relay node to be equal to that in the AF protocol, i.e., $g_1 \ldots g_i \sigma_i^2$ for i-th relay 3. Note that the capacity in (6) is achieved by beamforming at the source and MRC at the destination, in combination with a temporal capacity-achieving code for each hop (unlike the AF protocol, where a single temporal code is required for the whole channel).

III. COMPARISON OF AF AND DF PROTOCOLS

We are now in a position to compare the capacities in (4) and (6), which represent the ultimate rate bounds for coded systems under these protocols.

2When no source CSI is available, the best transmission strategy is isotropic signaling, $R_s = \sigma_s^2 I/n_s$ (see e.g. [6][7]), where $n_s$ is the number of source antennas, which achieves $\gamma_{AF}/n_s$ at the destination, so that all the results below will hold with appropriately scaled SNR.

3The signal power constraint is the standard one for the RF/microwave amplifier design [11]. Some authors set the same total (signal + noise) power. Our analysis can be extended to that scenario as well. In particular, our lower bounds will hold and also the upper bounds in the high SNR regime.
Proposition 1: The capacity gain $\Delta C = C_{DF} - C_{AF}$ of the DF relaying over the AF one in a N-hop relay channel in Fig. 1 is bounded as follows:

$$0 \leq \Delta C \leq \log N$$

for any channel realizations $h_1, h_2, \ldots, h_N$. When $C_{DF} < \infty$, the lower bound is achieved if and only if the destination and all the relays but one are noiseless. The upper bound is achieved if and only if all hops are equally strong and the SNR is sufficiently high, $\gamma_1 = \gamma_2 = \ldots = \gamma_N \rightarrow \infty$.

Proof: Since $\Delta C = \log(1 + \gamma_{DF})(1 + \gamma_{AF})^{-1}$, the lower bound follows from,

$$\gamma_{AF} = \left(\gamma_1^{-1} + \ldots + \gamma_N^{-1}\right)^{-1} \leq \min(\gamma_1, \ldots, \gamma_N) = \gamma_{DF}$$

where the equality is achieved if and only if $\gamma_k < \infty$ for some $k$ and $\infty$ otherwise, provided that $\gamma_{DF} < \infty$. To prove the upper bound, let $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_N$ be the ordered SNRs. Note that $\gamma_{DF} \geq \gamma_{AF}/N$ and $\gamma_{DF} = \gamma_{AF}$, so that

$$\Delta C \leq \log N + \log(1 + \gamma_1)(N + \gamma_1)^{-1} \leq \log N$$

The 1st inequality becomes equality when $\gamma_1 = \gamma_2 = \ldots = \gamma_N$ and the 2nd one – when $\gamma_1 \rightarrow \infty$.

It follows from Proposition 1 that $C_{DF} \geq C_{AF}$, i.e. the AF relaying is sub-optimal in general (for arbitrary channel realization). In practical terms, $C_{DF} \approx C_{AF}$, i.e. no significant gain is provided by the DF relaying over AF one, when one relay (or the destination) is much weaker than the others, $\gamma_1 \ll \gamma_2$. Under such condition, the AF relaying is a preferable solution since DF one incurs a significant complexity penalty in the relay design and implementation, and its capacity $C_{DF} = \log(1 + \gamma_{DF}) \approx C_{AF}$ is almost the same as that of the AF one (dominated by the weakest hop). The gain is maximum when all hops are equally strong and, for a two-hop channel, it never exceeds 1 bit/s/Hz.

The advantage of the DF relaying over the AF one can also be cast in terms of an SNR gain. Define the SNR gain $G$ of the DF relaying from the following:

$$C_{AF}(G \gamma_0) = C_{DF}(\gamma_0)$$

where $\gamma_0 = \sigma_s^2/\sigma_k^2$ is the source SNR. It shows how much more SNR (or, equivalently, the source power) is required for the AF relaying to achieve the same capacity as the DF relaying. This gain can be characterized as follows.

Proposition 2: The SNR gain of DF relaying over AF one for the N-hop relay channel in Fig. 1 is bounded as follows:

$$1 \leq G \leq N$$

and the lower bound is achieved when the destination and all the relays but one are noiseless. The upper bound is achieved when all hops are equally strong, $\gamma_1 = \gamma_2 = \ldots = \gamma_N$.

Proof: Follows along the same line as that of Proposition 1 using $G = \gamma_{DF}/\gamma_{AF}$.

We note that an equivalent form of (9) is

$$C_{AF}(\gamma_0) \leq C_{DF}(\gamma_0) \leq C_{AF}(N\gamma_0)$$

and the condition for achieving the upper bound in (9) can be re-written as follows,

$$\sigma_N^2 = g_N \sigma_k^2 \gamma_{AF}^2 = \ldots = g_N g_{N-1} \ldots g_2 \sigma_1^2$$

i.e. the destination own noise power equals to the noise power of any relay propagated to the destination. This balance of all noise powers is required to exploit the full advantage of the DF relaying. When all noise powers are the same, i.e. $\sigma_N^2 = \ldots = \sigma_1^2$, this reduces to $g_N = \ldots = g_2 = 1$, which is equivalent to $|h_i|^2 K_{i-1} = 1$, i.e. each relay compensates for the path loss of the hop following it.

An approximate (practical) condition for achieving the lower bound is $\gamma_{AF} \ll \gamma_{DF}$. For a two-hop channel, this reduces to $\sigma_2^2 \gg g_1 \sigma_1^2$ or $\sigma_2^2 \ll g_2 \sigma_2^2$, i.e. either the destination is significantly more noisy than the relay (as seen at the destination) or the other way around, which is illustrated in Fig. 2. This is the case when the DF relaying is not worth the complexity effort. In any case, the DF relaying outperforms the AF relaying by not more than 1 bit/s/Hz or 3 dB in a two-hop channel.

To capture the conditions when the upper and lower bounds are achieved in Propositions 1 and 2, we introduce the following definition.

Definition 1: The relay channel in Fig. 1 is balanced when $\gamma_1 = \gamma_2 = \ldots = \gamma_N$ (the upper bounds are achieved in (9), and in (7) when the SNR is sufficiently high). The relay channel is misbalanced when $\gamma_1/\gamma_2 \rightarrow 0$ (the lower bounds are achieved in (7) and (9)).

While Propositions 1 and 2 compare the DF and AF relaying for any given channel realization (or a fixed channel), the corresponding comparison can also be made for randomly-block-fading channels in terms of their two main performance metrics, outage probability and outage capacity [7]. In particular, the SNR gain bounds in (9) also apply to the latter metrics, and also to the ergodic capacity. The outage probability is

4e.g. when there is a line-of-sight (LOS) link between the base station and the relay, but no LOS between the relay and a mobile user.
defined as [7],

\[ P_{\text{out}}(\gamma_0) = \Pr \{ C(\gamma_0) < R \} \quad (12) \]

i.e. the probability that the channel is not able to support the target rate \( R \), where \( C(\gamma_0) \) is the instantaneous channel capacity. We further comment that the channel outage probability is also the best achievable codeword error probability. It follows from (10) that, for any fading distribution,

\[ P_{\text{out}}^{AF}(N\gamma_0) \leq P_{\text{out}}^{DF}(\gamma_0) \leq P_{\text{out}}^{AF}(\gamma_0) \quad (13) \]

From this, the DF relaying has at most \( N \)-fold SNR gain over the AF relaying in terms of the outage probability, i.e. as in (9), and the DF and AF relaying have the same diversity gain (under arbitrary fading distribution). The outage capacity is the largest rate such that the outage probability does not exceed a given threshold \( \varepsilon \) [7],

\[ C_\varepsilon(\gamma_0) = \max \{ R : P_{\text{out}}(\gamma_0, R) \leq \varepsilon \} \quad (14) \]

where \( P_{\text{out}}(\gamma_0, R) \) denotes the outage probability as a function of the SNR \( \gamma_0 \) and the target rate \( R \). Using this definition and (13), it follows that

\[ C_\varepsilon^{AF}(\gamma_0) \leq C_\varepsilon^{DF}(\gamma_0) \leq C_\varepsilon^{AF}(N\gamma_0) \quad (15) \]

so that the SNR gain of the DF relaying in terms of the outage capacity does not exceed \( N \) either. From (10), the same inequality also holds in terms of the ergodic (mean) capacity \( \overline{C}(\gamma_0) \): \( \overline{C}_{AF}(\gamma_0) \leq \overline{C}_{DF}(\gamma_0) \leq \overline{C}_{AF}(N\gamma_0) \).

Note that the results above hold for arbitrary fading distribution. They can also be extended to more complicated relay channel topologies under selection relaying, which is done in the next section.

IV. SELECTION RELAYING

Let us now consider a two-hop relay channel under selection relaying, where only the best relay node (out of \( N \)) is used at any time. It is motivated by its low complexity and also by the fact that little interference is created to other users since only one relay is transmitting [9]. Below, we compare the DF and AF protocols for this selection relaying scheme.

**Proposition 3:** The capacity gain \( \Delta C = C_{DF} - C_{AF} \) and the SNR gain \( G = \gamma_{DF}/\gamma_{AF} \) of the DF relaying over the AF relaying in a two-hop selection relaying channel are bounded as follows:

\[ 0 \leq \Delta C \leq \log N, \quad 1 \leq G \leq 2 \quad (16) \]

for any channel realization. The lower bounds are achieved when either the destination or the best DF relay is noiseless, i.e. the channel is misbalanced. The upper bounds are achieved when the best AF and DF relays are the same, their two hops are equally strong, i.e. the channel is balanced, \( \gamma_1 = \gamma_2 \) and, for the capacity gain, the SNR is sufficiently high, \( \gamma_1 = \gamma_2 \rightarrow \infty \).

Note that the bounds in (16) are independent of the total number of nodes \( N \) (out of which the best one is selected) and are the same as in (7) and (9) for \( N = 2 \), i.e. selection relaying in a two-hop channel does not improve the maximum possible gain of the DF over AF relaying.

We note that these selection relaying results can also be extended to the \( N \)-hop selection relaying, as in Fig. 3, using the same reasoning as above.

**Proposition 4:** The capacity gain \( \Delta C \) and the SNR gain \( G \) of the DF relaying over the AF relaying in a \( N \)-hop selection relaying channel in Fig. 3 are bounded as follows:

\[ 0 \leq \Delta C \leq \log N, \quad 1 \leq G \leq N \quad (17) \]

The lower bounds are achieved when the destination and all the relays but one of the best DF relaying path are noiseless (misbalanced channel). The upper bounds are achieved when the best DF and AF relaying paths are the same and all the hops of the best path are equally strong (balanced channel) and, for the capacity gain, the SNR is sufficiently high, \( \gamma_1 = \gamma_2 = \ldots = \gamma_N \rightarrow \infty \).

Finally, we point out that similar results also hold for a hybrid relaying scheme, where some of the relay nodes are DF ones and the rest are AF, and that all the bounds also extend to the outage probability/capacity and the ergodic capacity under arbitrary fading distribution, as discussed in Section III.

REFERENCES


Complexity analysis for ML-based sphere decoder achieving a vanishing performance-gap to brute force ML decoding

Arun Singh and Petros Elia
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: {singhak, elia}@eurecom.fr

Joakim Jaldén
ACCESS Linnaeus Center, Signal Processing Lab
School of Electrical Engineering, KTH, Stockholm, Sweden
Email: jalden@kth.se

Abstract—This work identifies the computational reserves required for the maximum likelihood (ML)-based sphere decoding solutions that achieve, in the high-rate and high-SNR limit, a vanishing gap to the error-performance of the optimal brute force ML decoder. These error performance and complexity guarantees hold for most multiple-input multiple-output scenarios, all reasonable fading statistics, all channel dimensions and all full-rate lattice codes. The analysis also identifies a rate-reliability-complexity tradeoff establishing concise expressions for the optimal diversity gain achievable in the presence of any run-time constraint imposed due to the unavailability of enough computational resources required to achieve a vanishing gap.

I. INTRODUCTION

For multiple-input multiple-output (MIMO) systems, error probability and encoding-decoding complexity are widely considered to be two limiting and interrelated bottlenecks (cf. [1], [2]). Specifically, if a small gap to the brute-force maximum likelihood (ML) performance is acceptable then different branch-and-bound algorithms such as the sphere decoder (SD) [2]–[4] have been known to provide a complexity-performance tradeoff. Albeit suboptimal in terms of the error-performance, these SD based solutions might result in a preferable rate-reliability-complexity tradeoff, a metric that is pertinent for the practical implementations.

A. System model

We consider general \( m \times n \) point-to-point MIMO channel representation

\[
y = \sqrt{\rho} R x + w
\]  

where \( x \in \mathbb{R}^m \), \( y \in \mathbb{R}^n \) and \( w \in \mathbb{R}^n \) respectively denote the transmitted codewords, the received signal vectors, and the additive white Gaussian noise with unit variance, where \( \rho \) denotes the signal to noise ratio (SNR), and where the fading matrix \( R \in \mathbb{R}^{m \times n} \) is assumed to be random, with elements drawn from arbitrary statistical distributions.

The sphere decoding solutions require that the underlying code be linear, an assumption that we adopt here and consider encoding and decoding schemes relating to real lattices (cf. [5]). Specifically for \( r \geq 0 \), a (sequence of) full-rate linear (lattice) code(s) \( X_r \) is given by \( X_r = \Lambda \cap R \) where the shaping region \( R \) is a compact convex subset of \( \mathbb{R}^n \) that is independent of \( \rho \), where \( \Lambda = \rho^{-\frac{1}{2}} \Lambda \) and \( \Lambda = \{ G s \mid s \in \mathbb{Z}^n \} \), where \( \mathbb{Z}^n \) denotes the \( \kappa = \min \{ m, n \} \) dimensional integer lattice, and where generator matrix \( G \in \mathbb{R}^{m \times n} \) is full rank and independent of \( \rho \). After vectorization the codewords take the form

\[
x = \rho^{-\frac{1}{2}} G s, \quad s \in \mathbb{Z}^n \cap \rho^{-\frac{1}{2}} R,
\]  

where \( \rho \in \mathbb{R}_+ \) is a natural bijection of the shaping region \( R \) that preserves the code, and contains the all zero vector 0. For simplicity we consider \( R = [-1, 1]^n \) to be a hypercube in \( \mathbb{R}^n \), although this could be relaxed. Combining (1) and (2) yields the equivalent system model

\[
y = M s + w, \quad \text{(3a)}
\]

where \( M = \rho^{-\frac{1}{2}} R G \in \mathbb{R}^{n \times n} \). \( \text{(3b)} \)

B. Sphere Decoder

Let \( QR = M \) be the thin QR factorization of the code-channel matrix \( M \) and \( r \triangleq Q^H y \), then (3a) yields \( r = R s + Q^H w \) and the ML decoder for this system takes the form

\[
\hat{s}_{ML} = \arg \min_{\hat{s} \in \mathbb{Z}^n} \| r - R \hat{s} \|^2. \quad \text{(4)}
\]

We use SD to implement the decoder in (4), which identifies as candidates the vectors \( \hat{s} \in \mathbb{Z}^n \) that for some search radius \( \xi > 0 \) satisfy \( \| r - R \hat{s} \|^2 \leq \xi^2 \). The algorithm specifically uses the upper-triangular nature of \( R \) to recursively identify partial symbol vectors \( \hat{s}_k, k = 1, \ldots, \kappa \), for which

\[
\| r_k - R_k \hat{s}_k \| \leq \xi^2, \quad \text{(5)}
\]

where \( \hat{s}_k \) and \( r_k \) respectively denote the last \( k \) components of \( \hat{s} \) and \( r \), and where \( R_k \) denotes the \( k \times k \) lower-right submatrix of \( R \).

We note that the error performance and the total number of visited nodes is a function of the search radius \( \xi \). We use fixed search radius \( \xi = \sqrt{\frac{\log \rho}{\kappa}} \) for some \( z > \kappa d(r) \) such that

\[
P \left( \| Q^H w \|^2 > \xi^2 \right) \leq P \left( \| w \|^2 > \frac{\xi^2}{\kappa} \right) \leq \rho^{-d(r)}, \quad \text{(6)}
\]

which implies a vanishing probability of excluding the transmitted information vector from the search. We use \( \leq \) to denote the exponential equality, i.e., we write \( f(\rho) = \rho^{B} \) to denote \( \lim_{\rho \to \infty} \frac{\log f(\rho)}{\log \rho} = B \), and \( \leq, \geq \) are defined similarly.
C. Rate-reliability-complexity tradeoff in outage-limited MIMO communications

In the high SNR regime, a given encoder $\mathcal{X}$ and decoder $\mathcal{D}$ are said to achieve a multiplexing gain $r$ and diversity gain $d_D(r)$ if (cf. [1])

$$\lim_{\rho \to \infty} \frac{R(\rho)}{\log \rho} = r, \quad \text{and} \quad \lim_{\rho \to \infty} \frac{\log P_e}{\log \rho} = d_D(r) \quad (7)$$

where $P_e$ denotes the probability of codeword error with a ML-based sphere decoder $\mathcal{D}$ employing time-out policies.

We characterize complexity in terms of the complexity exponent (cf. [4], [6]). Let $N_{\text{max}}$ denote the amount of computational reserves, in floating point operations (flops) per $T$ channel uses, that the transceiver is endowed with, in the sense that after $N_{\text{max}}$ flops, the transceiver must simply terminate, potentially prematurely and before completion of its task. The complexity exponent then takes the form

$$c(r) := \lim_{\rho \to \infty} \frac{N_{\text{max}}}{\log \rho}. \quad (8)$$

We note that the complexity exponent is intimately intertwined with the achievable error performance and that any attempt to reduce $c(r)$ may be at the expense of a substantial degradation in error-performance.

For ML-based SD a vanishing performance gap to ML can, in the high SNR regime, be quantified as

$$g(c) \triangleq \lim_{\rho \to \infty} \frac{P_e}{P(\hat{s}_{\text{ML}} \neq s)} = 1 \quad (9)$$

where $P(\hat{s}_{\text{ML}} \neq s) \leq \rho^{-d(c)}$ describes the error probability of the brute force ML decoder, and where $c$ is the complexity exponent that describes the computational resources required to achieve this performance gap. Generally a smaller performance gap requires a larger complexity exponent.

At this point a natural question to ask would be - how large computational reserves are required to achieve a vanishing gap to the brute force ML performance. While this question was first addressed and partially answered in [4] for the specific settings of i.i.d. Rayleigh fading quasi-static channels with specific channel dimensions, specific codes, and specific permutation orderings, we here provide answers for the most general MIMO settings, i.e., for all reasonable fading statistics, all channel dimensions, all MIMO scenarios and all full-rate lattice codes.

II. COMPLEXITY OF ML-BASED SPHERE DECODING

The total number of visited nodes is commonly taken as a measure of the sphere decoder complexity1 which is given by

$$N_{\text{SD}} = \sum_{k=1}^{\kappa} N_k, \quad (10)$$

where $N_k$ denotes the number of visited nodes at layer $k$ that corresponds to the $k$th component of the transmitted symbol vector $s$ and is given by $N_k \triangleq |N_k|$ where $N_k \triangleq \{\hat{s}_k \in \mathbb{S}_c^n \mid \|r_k - R_k\hat{s}_k\| \leq \xi^2\}$.

At this point we want to clarify that the analysis presented here is specific to sphere decoding, and that it does not account for any other ML based solutions that could, under some (arguably rare) circumstances, be more efficient. A classical example of such rare circumstances would be a MIMO scenario, or equivalently a set of fade statistics, that always generate diagonal channel matrices.

We are interested in the ML-based SD complexity required to achieve a vanishing gap to brute force ML performance. We recall that a ML-based SD with run-time constraints, in addition to making the ML errors ($\hat{s}_{\text{ML}} \neq s$, also makes errors when the run-time limit of $\rho^x$ flops for $x > c(r)$ becomes active, as well as when the fixed search radius $\xi$ causes $N_k = 0$. Consequently the corresponding performance gap to the brute force ML decoder, takes the form (cf. (9))

$$g(x) = \lim_{\rho \to \infty} \frac{P(\{\hat{s}_{\text{ML}} \neq s\} \cup \{N_{\text{SD}} \geq \rho^x\} \cup \{N_k = 0\})}{P(\hat{s}_{\text{ML}} \neq s)} = 0. \quad (11)$$

Now going back to (8), and having in mind appropriate timeout policies that guarantee a vanishing gap, the complexity exponent $c(r)$ can be bounded as $\zeta(r) \leq c(r) \leq \tau(r)$, where

$$\zeta(r) \triangleq \inf \{x \mid - \lim_{\rho \to \infty} \log P(\{N_{\text{SD}} \geq \rho^x\}) > d(r)\},$$

$$\tau(r) \triangleq \sup \{x \mid - \lim_{\rho \to \infty} \log P(\{N_{\text{SD}} \geq \rho^x\}) < d(r)\}, \quad (11a)$$

respectively denote sufficient and necessary conditions that guarantee a vanishing gap to ML performance.

Though our complexity results are applicable for all channel dimensions, we here assume2 $n \geq m$ and define $\mu_i \triangleq -\log_2 \min \{g_{\text{ML}}(\mathbf{H}^\dagger \mathbf{w}) \mid \mathbf{w} \in \mathbb{R}^m\}$, $i = 1, \ldots, m$. The upper bound follows from [4, Theorem 2] and is given by $\tau(r) \leq \tilde{c}(r)$ where

$$\tilde{c}(r) \triangleq \max_{\mu_1, \ldots, \mu_m} \sum_{i=1}^{\kappa} \min_{\mu} \left\{ \frac{rT}{\mu} - \frac{1}{2} (1 - \mu), \frac{rT}{\mu} \right\}^+ \quad (12a)$$

s.t. $I(\mu) \leq d(r), \quad (12b)$

$$\mu_1 \geq \cdots \geq \mu_m \geq 0, \quad (12c)$$

where $\mu \triangleq (\mu_1, \ldots, \mu_m)$ satisfies the large deviation principle with rate function $I(\mu)$. Equivalently for $\mu^* = (\mu_1^*, \ldots, \mu_m^*)$ being one of the maximizing vectors such that $I(\mu^*) = d(r)$,

2For $n < m$, we can consider a new $\tilde{G} = \tilde{U}^\dagger G \in \mathbb{R}^{n \times n}$ which is full-rank and square matrix, and where $\tilde{U} \in \mathbb{R}^{m \times n}$ has orthogonal columns and a new $\tilde{H} = \tilde{H} \tilde{U} \in \mathbb{R}^{n \times n}$. As no explicit assumption is made regarding the fading distribution of $\tilde{H}$, results will directly apply for $\tilde{H}$ and $\tilde{G}$. 

1It is easy to show that in the scale of interest the SD complexity exponent $c(r)$ would not change if instead of considering the total number of visited nodes, we considered the total number of flops spent by the decoder.
we have that $\hat{c}(r) = \sum_{i=1}^{n} \min \left( \frac{rT_i}{\kappa} - \frac{1}{2} (1 - \mu_i^*), \frac{rT_i}{\kappa} \right)^+$. Furthermore, given the monotonicity of the rate function $I(\mu)$, and the fact that the objective function in (12) does not increase in $\mu_i$ beyond $\mu_i = 1$, we may also assume without loss of generality that $\mu_i^* \leq 1$ for $i = 1, \ldots, \kappa$. It then follows that

$$\tau(r) \leq \hat{c}(r) = \sum_{i=1}^{n} \left( \frac{rT_i}{\kappa} - \frac{1}{2} (1 - \mu_i^*) \right)^+.$$ (13)

A. Universal Lower Bound on Complexity

In this section we establish that $\hat{c}(r) = \hat{c}(r)$, i.e., the sphere decoder visits a total number of nodes that is close to $\rho^{\hat{c}(r)}$ with a probability that is large compared to the probability of decoding error $P(\hat{S}_{ML} \neq \hat{s}) \leq \rho^{-\hat{c}(r)}$.

We let $q \in [1, \kappa]$ be the largest integer for which $\frac{rT_i}{\kappa} - \frac{1}{2} (1 - \mu_i^*) > 0$, in which case (13) takes the form

$$\hat{c}(r) = \sum_{i=1}^{q} \left( \frac{rT_i}{\kappa} - \frac{1}{2} (1 - \mu_i^*) \right).$$ (14)

We quickly note that without loss of generality we can assume that $q \geq 1$ as otherwise $\tau(r) = \hat{c}(r) = 0$. Consequently it is the case that $\mu_i^* > 0$ for $i = 1, \ldots, q$.

We proceed to define three events $\Omega_1$, $\Omega_2$ and $\Omega_3$ which we will prove to be jointly sufficient so that the total number of nodes visited by the sphere decoder, employing a channel dependent fixed decoding order, is close to $\rho^{\hat{c}(r)}$. These events are given by

$$\Omega_1 \triangleq \{ \mu_i^* - 2\delta < \mu_i < \mu_i^* - \delta, j = 1, \ldots, q \}$$

$$0 < \mu_i < \delta, i = q + 1, \ldots, \kappa \},$$

for a given small $\delta > 0$,

$$\Omega_2 \triangleq \left\{ \|w\|^2 \leq \frac{\epsilon^2}{\kappa} \right\},$$

$$\Omega_3 \triangleq \left\{ \|s\| < \frac{1}{2} \rho^{\epsilon/\kappa} \right\}.$$ (17)

Note also that by choosing $\delta$ sufficiently small, and using the fact that $\mu_i^* > 0$ for $i = 1, \ldots, q$, we may without loss of generality assume that $\Omega_1$ implies that $\mu_i > 0$ for all $i = 1, \ldots, \kappa$.

Following the footsteps of [4, Lemma 2] it can be shown that in the presence of events $\Omega_1$, $\Omega_2$ and $\Omega_3$ we can remove the ML-based SD boundary constraints $\mathbb{S}_p^\kappa$ (cf.(4)). This removal allows us to lower bound the number of nodes visited at layer $k$ as (cf. [4, Lemma 1])

$$N_k \geq \sum_{i=1}^{n} \left[ \frac{2\kappa}{\sqrt{k} \sigma_i(R_k)} - \sqrt{k} \right]^+.$$ (18)

In the following, and up until (28), we will work toward upper bounding $\sigma_i(R_k)$ for the case of $q \in [1, \kappa - 1]$, the case of $q = \kappa$ is treated separately later on. Towards this we first consider a Greedy QR decomposition. The diagonal elements of $\tilde{R}$ satisfy $\tilde{r}_{11} \geq \cdots \geq \tilde{r}_{pp}$. Let $M_{ij} \in \mathbb{R}^{n \times p}$ contains the first $p$ columns of $MI$. It then follows that

$$M_{ij} = \tilde{R}_i\tilde{R}_j = \tilde{Q}\tilde{R}_i\tilde{R}_j.$$ (19)

where, $\tilde{R}_i$ and $\tilde{R}_j$ denote the sub matrices consisting of the first $p$ columns of $\tilde{R}_i$ and $\tilde{R}_j$ respectively. Now let $R_p$ be the $p \times p$ upper triangular matrix consisting of the first $p$ rows of $R_p$, then we get that $\sigma_i(M_{ij}^M_{ij}) = \sigma_i(R_i^R R_j^R)$ for $i = 1, \ldots, p$. For $R_i^R R_j^R$, having diagonal entries $\tilde{r}_{ii}^2 \geq \cdots \geq \tilde{r}_{pp}^2$ and singular values $\sigma_i(M_{ij}^M_{ij}) \leq \cdots \leq \sigma_p(M_{ij}^M_{ij})$, we have that for $k = 1, \ldots, p$ then (cf. [8, Theorem 2.3])

$$\prod_{i=1}^{k} \tilde{r}_{ii}^2 \leq \prod_{i=1}^{k} \sigma_{p-i+1}(M_{ij}^M_{ij}).$$ (20)

From [7, Lemma 4.3] regarding the Greedy QR decomposition, we have that $\tilde{r}_{kk}^2 \geq \frac{\sigma_{k+1}(M_{ij}^M_{ij})}{\kappa - k + 1}$. Hence, we have that $k = 1, \ldots, p$

$$\sigma_{p-k+1}(M_{ij}^M_{ij}) \geq \sigma_{k+1}(M_{ij}^M_{ij}) \prod_{i=1}^{k} \frac{1}{\kappa - i + 1}.$$ (22)

We then have $\sigma_{p-k+1}(M_{ij}^M_{ij}) \leq \sigma_{k+1}(M_{ij}^M_{ij})$ for $k = 1, \ldots, p$ (cf. [9, Theorem 4.3.15]), and it follows that for $k = 1, \ldots, p$

$$\sigma_{p-k+1}(M_{ij}^M_{ij}) = \sigma_{k+1}(M_{ij}^M_{ij}).$$ (23)

Recalling that $\sigma_1(M_{ij}^M_{ij}) \leq \cdots \leq \sigma_\kappa(M_{ij}^M_{ij})$, we have that for $k = 1, \ldots, p$

$$\sigma_{p-k+1}(M_{ij}^M_{ij}) = \sigma_{k+1}(M_{ij}^M_{ij})$$ (24)

The above inequality allows us to apply Lemma 3 from [4], which in turn gives that

$$\sigma_i(R_k) \leq \frac{\sigma_i(M_{ij}^M_{ij}) + 1}{\sigma_{p-k+1}(M_{ij}^M_{ij})} + \sigma_i(M) \leq \frac{\sigma_i(M_{ij}^M_{ij}) + 1}{\sigma_{q+1}(M)} + \sigma_i(M),$$ (25)

for $i = 1, \ldots, q$, where exponential equality follows from (23). From (3b) for $i = 1, \ldots, \kappa$, we have that

$$\sigma_i(M) \leq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2} (1 - \mu_i).$$ (26)

Furthermore (15) gives that

$$\sigma_i(M) \leq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2} (1 - \mu_i) \quad \text{for} \quad i = 1, \ldots, q$$ (27a)

$$\sigma_\kappa(M) \leq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2} (1 - \mu_{\kappa}) \leq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2},$$ (27b)

$$\sigma_{q+1}(M) \geq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2} (1 - \delta).$$ (27c)

Substituting (27) in (25) gives that

$$\sigma_i(R_k) \leq \rho^{-\frac{\kappa}{\epsilon}} + \frac{1}{2} \delta + \frac{1}{2} (1 - \mu_i), \quad i = 1, \ldots, q.$$ (28)
Consequently, going back to (18), we have that

\[ \left[ \frac{2\bar{\rho}}{\sqrt{k} \sigma_i(R_k)} - \sqrt{k} \right]^{\gamma} \geq \rho \left( \frac{4\delta - \bar{\rho}^2(1-\mu)}{2\bar{\rho}} \right). \tag{29} \]

As a result, for \( k = q \) with \( q \in [1, K - 1] \) we have that

\[ N_q \geq \rho \sum_{i=1}^{K} \left( \frac{4\delta - \bar{\rho}^2(1-\mu)}{2\bar{\rho}} \right) = \rho \left( \bar{\rho}(r-\kappa - \delta) \right), \tag{30} \]

where the last equality follows from (14). For the case of \( q = \kappa \), from (18) and (27a) we have that

\[ N_q \geq \rho \sum_{i=1}^{K} \left( \frac{4\delta - \bar{\rho}^2(1-\mu)}{2\bar{\rho}} \right) = \rho \left( \bar{\rho}(r-\kappa - \delta) \right). \tag{31} \]

Consequently for \( q \in [1, \kappa] \) we have that \( N_{SD} \geq \rho \bar{\rho}(r-\kappa - \delta) \) for small \( \delta > 0 \), where \( K \in \{ \frac{4\delta}{\kappa}, \kappa \} \).

We note that (15)-(17) jointly imply that \( N_{SD} \geq \rho \bar{\rho}(r-\kappa - \delta) \).

For some \( \delta \in K \) and \( \delta > \delta_1 > 0 \), it follows that

\[ P \left( N_{SD} \geq \rho \bar{\rho}(r-\kappa - \delta) \right) \geq P \left( \Omega_1 \cap \Omega_2 \cap \Omega_3 \right) \neq P \left( \Omega_1 \right) \tag{32} \]

where exponential equality follows from the independence of the events \( \Omega_1, \Omega_2 \) and \( \Omega_3 \) and from the fact that \( P \left( \Omega_2 \right) \approx \rho \delta \) (cf.(6)) and \( P \left( \Omega_3 \right) \approx \rho \bar{\rho} \). With \( \Omega_1 \) being an open set, we have that

\[ \lim_{\rho \to \infty} \frac{P \left( \Omega_1 \right)}{\log \rho} \leq \inf_{\mu \in \Omega_1} I(\mu) = I(\tilde{\mu}) < I(\mu^*) = d(r) \tag{33} \]

where \( \tilde{\mu} = \{ \mu_1, \cdots, \mu_n \} \approx \{ 26, 0, \cdots, 0 \} \), the last inequality follows from the monotonicity of the rate function \( I(\mu) \) and where the last equality follows from the fact that, by definition, \( I(\mu^*) = d(r) \).

Consequently (32) and (33) along with the definition of the lower bound in (11b) imply that \( \bar{c}(r) = \tilde{c}(r) \), for arbitrarily small \( \delta > 0 \). The following lemma directly holds corresponding to a vanishing performance gap.

**Lemma 1:** Irrespective of channel fading statistics and of the full-rate code applied, for every realization of channel \( M \) there exists a channel dependent column permutation matrix \( \Pi \) such that the ML-based sphere decoder with decoding order \( \Pi \) has the complexity exponent \( c(r) = \tilde{c}(r) \).

To show the dependence of \( \Pi \) on \( M \), we henceforth use \( \Pi_{M} \) instead of \( \Pi \). Under the assumption that all column permutation matrices ‘appear’ with non-zero probability, then for every column permutation matrix \( \Pi_{k} \in \mathbb{R}^{n \times \infty} \) we have that \( P \left( \Pi_{M} = \Pi_{k} \right) \approx \rho \delta \), where probability is taken over random \( M \). Then the following theorem is a consequence of Lemma 1.

**Theorem 1:** For any full-rate code and fading distribution such that \( P \left( \Pi_{M} = \Pi_{k} \right) > \epsilon \) \( \forall \ \Pi_{k} \), for some \( \epsilon > 0 \), the complexity exponent of the ML-based sphere decoder with any fixed decoding order is given by \( c(r) = \tilde{c}(r) \).

### III. Rate-Reliability-Complexity Tradeoff

In this section we present a rate-reliability-complexity tradeoff for ML-based SD, which identifies the optimal diversity gain achievable in the presence of any run-time constraint imposed due to the unavailability of enough computational resources required to achieve a vanishing gap. The proofs are omitted from this writeup due to the lack of space.

**Theorem 2:** For any full-rate code with ML-based diversity gain \( d(r) \) and any fading distribution such that \( P \left( \Pi_{M} = \Pi_{k} \right) > \epsilon \) \( \forall \ \Pi_{k} \), for some \( \epsilon > 0 \), the achievable diversity performance \( d_{SD}(r) \) for ML-based SD with any fixed decoding order and a run-time constraint \( \rho^{2\bar{\rho}}(r) \) flops, is uniquely described by

\[ d_{SD}(r) = \min \{ d(r), d_{SD}(r, x) \} \ \forall \ \epsilon, \tag{34} \]

where \( d_{SD}(r, x) \) \( \approx \) \( \max_{\epsilon} \in \epsilon \), \( d_{SD}(r, x) \approx \epsilon \), and where \( d_{SD}(r, x) + \epsilon \approx \inf I(\mu) \)

\[ \left\{ \sum_{i=1}^{n} \left( \frac{r_i}{r} - \frac{1}{2} \right) \geq \epsilon, \right. \tag{35} \]

\[ \left. \min_{\epsilon} \geq \mu_1 \geq \cdots \geq \mu_\kappa \geq 0. \right. \]

Example: For a \( 2 \times 2 \) i.i.d. Rayleigh channel with the \( 2 \times 2 \) Perfect code [5] and a ML-SD with run-time constraint of \( \rho^{2} \) flops, the achievable diversity gain is depicted in Fig.1.

![Fig. 1. Achievable diversity gain for 2 x 2 Perfect codes.](image)

### IV. Conclusions

The presented performance guarantees hold for the most general MIMO settings, i.e., for all reasonable fading statistics, all channel dimensions, all MIMO scenarios and all full-rate lattice codes. Such guarantees may be utilized for the practical implementation of telecommunication systems.

### REFERENCES


Abstract—An efficient sub-optimum symbol-by-symbol soft in/soft-out a posteriori probability (APP) decoding algorithm for linear binary block codes is presented. The three-step decoding procedure consists of: 1) list decoding of two supercodes, 2) calculating the intersection of these two lists, and 3) symbol-by-symbol APP decoding for the intersection. The presented results indicate that the proposed decoding procedure obtains good estimates of the ideal soft-output values and reduces the decoding complexity significantly. It may be an attractive alternative to conventional implementations of APP decoding.

I. INTRODUCTION

Most commonly, symbol-by-symbol a posteriori probability (APP) decoding of block codes is implemented by representing the code as a graph, the so-called trellis, and applying the well-known BCJR algorithm [1]. Alternative implementations of soft in/soft-out decoding are based on the soft-output Viterbi algorithm [2] or sub-optimum versions of the BCJR algorithm [3].

In this paper we consider a different approach which is based on the concept of supercode decoding introduced by Barg et al. [4]. The algorithm from [4] actually combines several previously known decoding strategies for discrete memoryless channels with the new idea of supercode decoding. A generalization of all employed decoding steps to channels with continuous output alphabets appears to be difficult. Therefore, we concentrate on the problem of supercode decoding. Consequently the intermediate decoding steps differ from that described in [4]. In particular we present efficient solutions to the problems: 1) list decoding of supercodes and 2) calculating the intersection \( L = \cap_i L_i \). The list \( L \) is represented as a trellis and the final decoding step 3) can be performed using, for example, Viterbi decoding in this trellis. Such a trellis based supercode decoding was devised and investigated for the binary symmetrical channel (BSC) in [5].

In this work we generalize trellis based supercode decoding to symbol-by-symbol soft in/soft-out a posteriori probability (APP) decoding for linear binary block codes, i.e., in the last decoding step we decode the trellis representing the intersection with the BCJR algorithm. Due to the trellis based approach we are not limited to the Hamming metric. Hence, the presented approach is suitable for soft in/soft-out decoding. It is inherently sub-optimum, because the APP decoding does not consider the complete code. Nevertheless, the presented simulation results indicate that the proposed decoding procedure obtains good estimates of the ideal soft-output values (similar to the well known Max-Log-MAP approximation [3]) and reduces the decoding complexity significantly.

In Section II, we briefly discuss the basic concept of supercode decoding. Sections III and IV describe a novel implementation of the trellis based supercode decoding-algorithm presented in [5]. When implementing an algorithm there is always a trade off between memory usage and speed. The new implementation reduces the runtime complexity compared to the approach presented in [5]. Finally, some simulation results are presented in Section V.

II. BASIC CONCEPT

A supercode \( C_i \) of the binary block code \( C \) is a code containing all codewords of \( C \). For a linear code \( C \) with parity-check matrix \( H \), we can construct two supercodes \( C_1 \) and \( C_2 \) such that \( C = C_1 \cap C_2 \). Let \( H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \) be the parity-check matrix of the code \( C \), this means that \( H_1 \) and \( H_2 \) are two sub-matrices of \( H \). Then the sub-matrices \( H_1 \) and \( H_2 \) define the supercodes \( C_1 \) and \( C_2 \), respectively.

Example 1: Consider for example the code \( C = \{(0000), (1110), (1011), (0101)\} \) with parity-check matrix 

\[
H = \begin{pmatrix}
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 1
\end{pmatrix}
\]

We obtain

\[
H_1 = \begin{pmatrix} 1 & 1 & 0 & 1 \end{pmatrix}
\]

\[
\downarrow
\]

\[
C_1 = \{(0000), (1100), (1110), (0010), (1011), (1001), (1111), (0101)\}
\]

and

\[
H_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \end{pmatrix}
\]

\[
\downarrow
\]

\[
C_2 = \{(0000), (1000), (0110), (1110), (1011), (1101), (0011), (0101)\}
\]
where the underlined vectors are the codewords of the code \( C \).

Supercode decoding is performed in three steps:

1) List decoding of each supercode \( C_t \) such that the transmitted codeword is in the list \( L_t \), provided that the weight of the error pattern is less or equal \( \rho \).

2) Calculating the intersection \( L = \cap L_t \). These are the codewords of the code \( C \) that are in the ball around the received sequence.

3) Applying symbol-by-symbol soft in/soft-out a posteriori probability decoding to the list \( L \).

The idea here is that list decoding of the supercodes is less complex than a decoding approach which is based on the code itself. The supercodes are represented by their minimal trellis [6]. The first decoding step is performed by a marking algorithm which labels all state transitions in a supercode trellis that corresponds to a codeword in the list \( L_t \).

The list \( L \) is also represented by a trellis. This trellis is constructed using the merging algorithm. The merging algorithm calculates the intersection \( L = \cap L_t \) by merging the labeled state transitions from both supercode trellises.

Before we describe the algorithm’s implementation we introduce the data structures which the trellis is composed of.

III. TRELLIS REPRESENTATION

A trellis \( T = (S, W) \) is a directed graph with a set of states \( S = \{ \sigma \} \) and a set of branches \( W = \{ w \} \) which connects the nodes. The set \( S \) can be split into \( n + 1 \) subsets \( S = S_0 \cup S_1 \cup \cdots \cup S_n \), where \( n \) denotes the length of the code. Such a subset is also called a level \( S_t \) of the trellis. Each branch \( w \) has a label \( b \in \{ 0, 1 \} \) which represents the code symbol \( v_t(w_t) \).

To represent the trellis, three data structures are used. One for the node \( \sigma \), a list of nodes for a level \( S_t \), and a list of the list of nodes for the complete trellis \( S \). The branches \( W \) are references which are stored in the nodes.

A node can have up to two predecessors \( N^{\rho}_b(\sigma) \) and successors \( N^\omega_b(\sigma) \), \( b \in \{ 0, 1 \} \). The references to the predecessors and successors are assigned to zero if there is no corresponding node. Each incoming branch \( w^\sigma_b \) which connects \( \sigma \) with \( N^\rho_b(\sigma) \) and outgoing branch \( w^\omega_b \) which connects \( \sigma \) and \( N^\omega_b(\sigma) \) has a metric. We differentiate between two metric values. The forward metric \( \Lambda^F(\sigma) \), which is the metric between \( \sigma \) and \( N^\rho_b(\sigma) \), and the backward metric \( \Lambda^B(\sigma) \), which is the metric between \( \sigma \) and \( N^\omega_b(\sigma) \). Furthermore a node stores the markings \( M_0(\sigma) \) for the outgoing branches \( w^\omega_b \). Hence, a node of the super code trellises can be described as a 10-tupel.

\[
\sigma = (\Lambda^B_0(\sigma), \Lambda^B_1(\sigma), \Lambda^B_2(\sigma), \Lambda^F_0(\sigma), \Lambda^F_1(\sigma), N^\rho_b(\sigma), N^\omega_b(\sigma), M_0(\sigma), M_1(\sigma))
\]

These nodes are used to construct the trellises of the supercodes. For the trellis representing the list \( L \) we have to extend the node with four attributes. Two references \( N^{(1)}(\sigma) \) and \( N^{(2)}(\sigma) \) which point to the corresponding nodes in the supercode trellises. Additionally the level \( L_t(\sigma) \) of the node and a visited flag \( V(\sigma) \) were added to implement the breadth first search efficiently. Thus, a node of the trellis can be described as a 14-tupel.

\[
\sigma = (\Lambda^B_0(\sigma), \Lambda^B_1(\sigma), \Lambda^B_2(\sigma), N^\rho_b(\sigma), N^\omega_b(\sigma), N^{(1)}(\sigma), N^{(2)}(\sigma), M_0(\sigma), M_1(\sigma), V(\sigma), L(\sigma))
\]

The implementation uses three trellises for decoding. The two supercode trellises \( T^{(1)} \) and \( T^{(2)} \) for the supercodes \( C_1 \) and \( C_2 \), and the trellis \( T \) for the code \( C \). Each node in \( T \) has a reference to a node in \( T^{(1)} \) and \( T^{(2)} \), such that

\[
\sigma = \begin{pmatrix} t^{(1)} \sigma \\ t^{(2)} \sigma \end{pmatrix}
\]

IV. SUPERCODE DECODING

The first step of supercode decoding is list decoding of each supercode \( C_t \) such that the transmitted codeword is in the list \( L_t \) provided that the weight of the error pattern is less or equal \( \rho \). With trellis based supercode decoding the list decoding is performed in the trellis of the supercodes and each list is represented by a marked trellis. Hence, the first decoding step is called the marking algorithm.

a) The marking algorithm: The marking algorithm is applied to the two supercode trellises.

Step 1: Set \( \Lambda^F_0(\sigma_0) = \Lambda^F_1(\sigma_0) = 0 \) and \( M_0(\sigma_0) = M_1(\sigma_0) = 0 \).

Step 2: Calculate the forward metric for each node \( \sigma_t \), \( t = 1, \ldots, n \) when \( N^\rho_b(\sigma_t) \neq 0 \), \( b = 0, 1 \):

\[
\Lambda^F(\sigma_t) = \begin{cases} 
\min(\Lambda^F_0(\sigma_t), \Lambda^F_1(\sigma_t)), & \text{if } N^\rho_b(\sigma_t) \neq 0 \land N^\rho_b(\sigma_t) \neq 0 \\
\Lambda^F_0(\sigma_t), & \text{if } N^\rho_b(\sigma_t) \neq 0 \land N^\rho_b(\sigma_t) = 0 \\
\Lambda^F_1(\sigma_t), & \text{if } N^\rho_b(\sigma_t) = 0 \land N^\rho_b(\sigma_t) \neq 0 \\
\end{cases}
\]

Reset all markings: \( M_0(\sigma_t) = 0 \).

Step 2.1: If \( \Lambda^F(\sigma_n) \leq \rho \), stop decoding.

Step 3: Set \( \Lambda^B_0(\sigma_0) = \Lambda^B_1(\sigma_0) = 0 \).

Step 4: Calculate the backward metric for each node \( \sigma_t \), \( t = n - 1, \ldots, 0 \) when \( N^\omega_b(\sigma_t) \neq 0 \), \( b = 0, 1 \):

\[
\Lambda^B(\sigma_t) = \begin{cases} 
\min(\Lambda^B_0(\sigma_t), \Lambda^B_1(\sigma_t)), & \text{if } N^\omega_b(\sigma_t) \neq 0 \land N^\omega_b(\sigma_t) \neq 0 \\
\Lambda^B_0(\sigma_t), & \text{if } N^\omega_b(\sigma_t) \neq 0 \land N^\omega_b(\sigma_t) = 0 \\
\Lambda^B_1(\sigma_t), & \text{if } N^\omega_b(\sigma_t) = 0 \land N^\omega_b(\sigma_t) \neq 0 \\
\end{cases}
\]

Step 3.1: If \( \Lambda^B(\sigma_0) \leq \rho \), stop decoding.
Set markings for $b = 0, 1$:
$M_0(\sigma_i) = M(N^{(1)}_0(\sigma_i)) \land ((\Lambda^F(\sigma) + \Lambda^B(\sigma)) \leq \rho)$
with $M(\sigma) = M_0(\sigma) \lor M_1(\sigma)$

Considering that each node of both supercode trellises has to be visited twice the runtime complexity of the marking algorithm is given by the size of the supercode trellises.

The second step of supercode decoding is the calculation of the intersection $L = \cap_i L_i$. Again, the intersection $L = \cap_i L_i$ is represented by a trellis. The intersection is calculated with the merging algorithm.

b) The merging algorithm (Breath first Viterbi):
Before we introduce the merging algorithm we have to define a queue. This queue is needed for the breath first search implementation. A queue is a list of $m$ nodes $Q = \{\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+m-1}\}$ which can shrink and grow dynamically. The queue has two operations. One to add a node at the end $Q = enqueue(Q, \sigma)$ and a second to remove the first node $\sigma = dequeue(Q) = \sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+m-1}, \sigma_{i+m}$.

Step 1: Set $\Lambda^F(\sigma_0) = \Lambda^F(\sigma_0) = 0$ and $Q = enqueue(Q, \sigma_0)$

Step 2: Breath first search
while ($\sigma = dequeue(Q)$)
{
    if ($L(\sigma) > 0$)
    {
        $\Lambda^F(\sigma) = \Lambda^F(N^F_0(\sigma)) + \lambda(\omega_{L(\sigma)} - 1)$
        $\Lambda^F(\sigma) = \Lambda^F(N^F_1(\sigma)) + \lambda(\omega_{L(\sigma)} - 1)$
    }

    $M_0(\sigma) = M_0(N^{(1)}(\sigma)) \land M_0(N^{(2)}(\sigma))$
    $\neg \epsilon(\sigma, \sigma_0) \land \neg V(N^F_0(\sigma))$
    $V(N^F_0(\sigma)) = 1$
    $Q = enqueue(Q, N^F_0(\sigma))$
}

$M_1(\sigma) = M_1(N^{(1)}(\sigma)) \land M_1(N^{(2)}(\sigma))$
$\neg \epsilon(\sigma, \sigma_0) \land \neg V(N^F_1(\sigma))$
$V(N^F_1(\sigma)) = 1$
$Q = enqueue(Q, N^F_1(\sigma))$

Step 3: Find best path from $\sigma_{n-1}$ to $\sigma_0$ considering only the marked nodes.

Step 4: Reset markings and visited flags. (Second breath first search run, with inverted visited flag)

This algorithm avoids iterating over the supercode trellis nodes. In order to reduce the runtime complexity we introduce a pairing of supercode trellis nodes which can be done once before the actual decoding. The simplest way to do this is to create the trellis of the code $C$. To achieve that a node of the trellis is implicitly marked we add two references to the corresponding nodes in the supercode trellises. As a result, a node in the trellis is marked if both corresponding nodes in the supercode trellises are marked. To calculate the metrics of the marked nodes we cannot iterate over all trellis nodes to check if a node is marked. This would mean that the merging algorithm has the same runtime complexity as Viterbi decoding. Instead of iterating over all nodes we use a breath first search [7] to visit only the marked nodes. The breath first search was chosen because it traverses the trellis level by level. Consequently, the metrics can be calculated as in the marking algorithm.

V. SIMULATION RESULTS

In this chapter we present some simulation results for binary Reed-Muller code. We compare the trellis based supercode decoder with the Viterbi and BCJR algorithms.

The simulation results for APP decoding are presented in Fig. 1. The left diagram shows the bit error rates for the $RM(3,6)$ code (with code length $n = 64$ and dimension $k = 42$). With $\rho = 5$ the APP decoding based on supercode decoding achieves almost BCJR-performance. With $\rho = 4$ the decoding performance is close to that of the Max-Log-MAP algorithm. In the right diagram we present the quality of the soft output values as the error compared to the bit error rate $P_b$. We estimate the bit error rate from the soft output values as

$$P_b = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{1 + e^{L_s}}$$

where $L_s$ is a soft output value [8], [9]. The right diagram in Fig. 1 presents the relative error

$$|\tilde{P}_b - P_b|$$

It can be seen that with $\rho = 5$ the quality of the soft output is comparable to the Max-Log-MAP algorithm.

Table I shows the number of nodes visited by the supercode decoding algorithm for different boundaries. Additionally the number of nodes visited by the Viterbi algorithm is presented. As said before, the complexity of the marking algorithm is independent of the input sequence and the bounded distance. Hence we can divide the complete decoding complexity in a static part, caused by the marking, and a dynamic part, caused by the merging. For $RM(2,5)$ (with code length $n = 32$ and dimension $k = 16$) and $\rho = 4$ the number of visited nodes by the Viterbi and supercode decoding algorithm are almost similar. But for $RM(3,6)$ and $\rho = 4$ the Viterbi algorithm has to visit eight times more nodes than the supercode decoding algorithm.

VI. CONCLUSIONS

In this work we have presented an efficient sub-optimum symbol-by-symbol soft in/out a posteriori probability (APP) decoding algorithm for linear binary block codes. The new method is based on trellis based supercode decoding
which was introduced in [5]. The new implementation presented in this paper reduces the runtime complexity compared to the approach presented in [5].

The presented simulation results indicate that the proposed APP decoding procedure obtains good estimates of the ideal soft-output values and reduces the decoding complexity significantly compared to decoding in the complete syndrome trellis. It is therefore an attractive alternative to conventional implementations of APP decoding for linear binary block codes.

LDPC codes or turbo codes that are based on binary block codes usually employ high rate Hamming codes [10], [11], [12] that allow conventional low complexity APP decoding. The new algorithm may enable other code constructions that are based on block codes with higher error correcting capabilities.

REFERENCES

Author Index

Abbe, E. ..........44, 51
Altu˘g, Y. ..........15
Amir, E. ..........8
Belfiore, J.-C. ..111
Blasco-Serrano, R. .55
Bossert, M. .......6, 63
Boston, N. ..........83
Bross, S. I. ........24
Bustin, R. ........20
Conti, D. ..........83
Cuff, P ..........11
Dimakis, A. G. .......49
D¨orpinghaus, M. ...28
Elia, P ..........127
Erez, U ..........99
Etezadi, F ..........95
Feizi, S. ........107
Fischer, R. F. H. ...119
Fleury, B. H. .......115
Freudenberg, J. ...131
Gastpar, M. .......51
Geiger, B. ........32
Goela, N. ........51
Grant, A ........67
Huber, M. ..........87
Jakobsen, M. L. ...115
Jaldén, J ..........127
Kaiser, U ..........131
Karzand, M ........59
Khisti, A ..........95
Kim, Y.-H ..........5
Koch, T ..........36
Kontoyiannis, I .......48
Levin, G ..........123
Lapidoth, A ..........36
Li, X ..........75
Loyka, S ..........123
Lucani, D. E ........107
Macris, N ..........79
Mathar, R ..........28
Matsumoto, R ........75
Médard, M .......107
Nazer, B ..........103
Orlitsky, A ..........5
Pedersen, T ........115
Permuter, H ..........24
Rooij, S. de .......40
Rüngeler, M ..........28
Saber Tehrani, A .......49
Schenk, A ..........119
Shah, D ..........119
Shamai, S ..........20
Singh, A ..........127
Skoglund, M ..........55
Soljanin, E ..........6
Steinberg, Y ..........8
Subramanian, R ........71
Telatar, E ..........5
Thobaben, R ..........55
Trott, M ..........95
Tuninetti, D ..........12
Urbanke, R ..........5
Venkat, K ..........43
Viterbo, E ..........6
Vuffray, M ..........79
Wachter, A ..........63
Wagner, A. B ..........15
Wang, G ..........67
Weissman, T ..........43
Wigger, M ..........19
Wornell, G. W ..........99
Zamir, R ..........91
Zeh, A ..........63