International Zurich Seminar on Communications

Proceedings

International Zurich Seminar on Communications

February 26 – 28, 2014

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
Martin Bossert
Terence H. Chan
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Martin Hänggi
Franz Hlawatsch
Johannes Huber
Tobias Koch
Gerhard Kramer
Frank Kschischang
Hans-Andrea Loeliger
Thomas Mittelholzer
Stefan M. Moser
Nikolai Nefedov
Bernhard Plattner
Igal Sason
Jossy Sayir
Robert Schober
Giorgio Taricco
Emre Telatar
Emanuele Viterbo
Pascal Vontobel
Michèle Wigger
Armin Wittneben

Organizers of Invited Sessions

Jean-Claude Belfiore
Alex Grant
Deniz Gunduz
Olgica Milenkovic
Tsachy Weissman

Local Organization

Conference Secretaries

Rita Hildebrand
Silvia Tempel

Web and Publications

Michael Lerjen

Table of Contents

Keynote Talks

Session 1 Wed 10:00 – 12:00
Shannon Theory

Chaired by Shlomo Shamai

[*] Invited papers are marked by an asterisk.

Session 2 Wed 13:20 – 15:00

Advances in Shannon Theory

Invited session organizer: Tsachy Weissman

Session 3 Wed 15:30 – 16:50

Coding Theory

Invited session organizer: Jean-Claude Belfiore

Session 4 Wed 17:00 – 17:40

Secrecy

Chaired by Ashish Khisti

Session 5 Thu 10:00 – 12:00

Coding Theory

Chaired by Hans-Andrea Loeliger

Session 6 Thu 13:20 – 14:40

Information Theoretic Approaches to Database Management

Invited session organizer: Deniz Gunduz

Session 7 Thu 15:10 – 17:10
Wireless Communications

Chaired by Johannes Huber

Session 8 Fri 10:00 – 12:00
Relaying and Information

Chaired by Michael Gastpar

Session 9 Fri 13:20 – 15:00

Sparse Signal Processing and Coding

Invited session organizer: Olgica Milenkovic

*Model-based Sketching and Recovery with Expanders
V. Cevher

*Energy Allocation in Compressed Sensing of Non-uniformly Sparse Signals
W. Dai

*Noisy Boolean Compressed Sensing and Error-Correcting Codes
A. Mazumdar

*Communications over Sparse Channels: Fundamental Performance Limits and Practical System Design
P. Schniter

*Semi-Quantitative Group Testing
O. Milenkovic

Session 10 Fri 15:30 – 16:50

Information Rates and Coding for Networks

Invited session organizer: Alex Grant

*Twelve Short Schemes for Index Coding
Y.-H. Kim

*Constrained Entropy Maximisation 152
T. H. Chan and A. Grant

*On Capacity, Cooperation, and the Edge Removal Problem
M. Effros

*The Modulo-Lattice Output is a Sufficient Statistic
R. Zamir

A Rate-Splitting Approach to Fading Multiple-Access Channels with Imperfect Channel-State Information

Adriano Pastore*, Tobias Koch†, Javier R. Fonollosa*
* Universitat Politècnica de Catalunya, Signal Theory and Communications Department, 08034 Barcelona, Spain
Email: {adriano.pastore,javier.fonollosa}@upc.edu
† Universidad Carlos III, Signal Theory and Communications Department, 28911 Leganés, Spain
Email: koch@tsc.uc3m.es

***Abstract*—As shown by Médard, the capacity of fading channels with imperfect channel-state information (CSI) can be lower-bounded by assuming a Gaussian channel input and by treating the unknown portion of the channel multiplied by the channel input as independent worst-case (Gaussian) noise. Recently, we have demonstrated that this lower bound can be sharpened by a rate-splitting approach: by expressing the channel input as the sum of two independent Gaussian random variables (referred to as layers), say $X = X_1+X_2$, and by applying Médard's bounding technique to first lower-bound the capacity of the virtual channel from X_1 to the channel output Y (while treating X_2 as noise), and then lower-bound the capacity of the virtual channel from X_2 to Y (while assuming X_1 to be known), one obtains a lower bound that is strictly larger than Médard's bound. This rate-splitting approach is reminiscent of an approach used by Rimoldi and Urbanke to achieve points on the capacity region of the Gaussian multiple-access channel (MAC). Here we blend these two rate-splitting approaches to derive a novel inner bound on the capacity region of the memoryless fading MAC with imperfect CSI. Generalizing the above rate-splitting approach to more than two layers, we show that, irrespective of how we assign powers to each layer, the supremum of all rate-splitting bounds is approached as the number of layers tends to infinity, and we derive an integral expression for this supremum. We further derive an expression for the vertices of the best inner bound, maximized over the number of layers and over all power assignments.**

I. Introduction

Consider a discrete-time, memoryless, fading channel with imperfect channel-state information (CSI), whose time-k output ($k \in \mathbb{Z}$), conditioned on the channel input $X[k] = x \in \mathbb{C}$, is

$$Y[k] = (\hat{H}[k] + \tilde{H}[k])x + Z[k] \tag{1}$$

(with $\mathbb{C}$ and $\mathbb{Z}$ denoting the set of complex numbers and the set of integers, respectively). Here, the noise $\{Z[k]\}_{k\in\mathbb{Z}}$ is a sequence of independent and identically distributed (i.i.d.), zero-mean, circularly-symmetric, complex Gaussian random variables with variance σ^2. The fading processes $\{\hat{H}[k]\}_{k\in\mathbb{Z}}$ and $\{\tilde{H}[k]\}_{k\in\mathbb{Z}}$ are both sequences of i.i.d. complex random variables (of arbitrary distribution), the former with mean μ and variance $\hat{V}$ and the latter with mean zero and variance $\tilde{V}$. Assume that the processes $\{\hat{H}[k]\}_{k\in\mathbb{Z}}$, $\{\tilde{H}[k]\}_{k\in\mathbb{Z}}$, and $\{Z[k]\}_{k\in\mathbb{Z}}$ are independent of each other and of the input sequence $\{X[k]\}_{k\in\mathbb{Z}}$. Further assume that the receiver is cognizant of the realization of $\{\hat{H}[k]\}_{k\in\mathbb{Z}}$, but the transmitter is only cognizant of its distribution. Finally assume that both the transmitter and receiver are cognizant of the distributions of $\{\tilde{H}[k]\}_{k\in\mathbb{Z}}$ and $\{Z[k]\}_{k\in\mathbb{Z}}$ but not of their realizations.

The fading process $\{\hat{H}[k]\}_{k\in\mathbb{Z}}$ can be viewed as an estimate of the channel fading coefficient

$$H[k] \triangleq \hat{H}[k] + \tilde{H}[k], \quad k \in \mathbb{Z} \tag{2}$$

and $\{\tilde{H}[k]\}_{k\in\mathbb{Z}}$ can be viewed as the channel estimation error.

The capacity of the above channel (1) under the average-power constraint P is given by [1]

$$C(P) = \sup I(X;Y|\hat{H}) \tag{3}$$

where the supremum is over all distributions of X satisfying $\mathsf{E}[|X|^2] \le P$. Here and throughout the paper we omit the time indices k where they are immaterial. Since (3) is difficult evaluate, it is common to assess $C(P)$ using upper and lower bounds. A well-known lower bound is due to Médard [2]:

$$C(P) \ge \mathsf{E}\left[\log\left(1 + \frac{|\hat{H}|^2 P}{\tilde{V}P + \sigma^2}\right)\right]. \tag{4}$$

It is derived by assuming a Gaussian channel input X and by treating the term $\tilde{H}X + Z$ as independent worst-case (Gaussian) noise.

In [3], it was demonstrated that (4) can be sharpened by a rate-splitting and successive decoding approach: writing the input $X = X_1 + X_2$ as a sum of two independent Gaussian random variables (referred to as *layers*) of respective powers P_1 and P_2, using the chain rule

$$I(X;Y|\hat{H}) = I(X_1;Y|\hat{H}) + I(X_2;Y|\hat{H},X_1) \tag{5}$$

A. Pastore and J. R. Fonollosa have been supported by the Ministerio de Economía y Competitividad of Spain (TEC2010-19171 and CONSOLIDER-INGENIO 2010 CSD2008-00010 COMONSENS) and the Generalitat de Catalunya (2009SGR-1236). T. Koch has been supported by the European Community's Seventh's Framework Programme (FP7/2007-2013) under grant agreement No. 333680 and by the Ministerio de Economía y Competitividad of Spain (TEC2009-14504-C02-01, TEC2012-38800-C03-01, and CONSOLIDER-INGENIO 2010 CSD2008-00010 COMONSENS).

and applying Médard's bound on each term, we obtain a lower bound that is strictly larger than (4) except in the trivial cases where $P_1 = 0$, $P_2 = 0$, or $\mathsf{Pr}(\hat{H}\tilde{V} = 0) = 1$. This rate-splitting approach can be generalized to more than two layers. It was demonstrated that the supremum of all such rate-splitting bounds is approached as the number of layers tends to infinity and an integral expression of this supremum was presented [3, Theorem 4].

The above rate-splitting approach is reminiscent of a rate-splitting approach proposed by Rimoldi and Urbanke to achieve points on the capacity region of the Gaussian multiple-access channel [4]. For example, for the two-user Gaussian MAC, Rimoldi and Urbanke showed that any point in the capacity region can be achieved by splitting one user, say User 1, into two virtual users,[2] and by decoding first the codeword of the first virtual user while treating the codewords of the second virtual user and of User 2 as noise, by then decoding the codeword of User 2 upon subtracting the contribution of the first virtual user and treating the codeword of the second virtual user as noise, and by finally decoding the codeword of the second virtual user upon subtracting the contributions of the first virtual user and User 2.

In this paper, we blend the two rate-splitting approaches in [3] and [4] to derive a novel inner bound on the capacity region of the memoryless fading MAC with imperfect CSI. We show that, irrespective of how we assign powers to each layer, the supremum of all such rate-splitting bounds is approached as the number of layers tends to infinity, and we derive an integral expression for this supremum. We further derive an expression for the best inner bound, maximized over the number of layers and all power assignments.

II. Channel Model and Capacity Region

We consider the multiple-access generalization of (1): the time-k output $Y[k]$, conditioned on the channel inputs $X_1[k] = x_1 \in \mathbb{C}$ and $X_2[k] = x_2 \in \mathbb{C}$ corresponding to User 1 and User 2, respectively, is

$$Y[k] = (\hat{H}_1[k] + \tilde{H}_1[k])x_1 + (\hat{H}_2[k] + \tilde{H}_2[k])x_2 + Z[k] \quad (6)$$

where $\{Z[k]\}_{k\in\mathbb{Z}}$ is as in Section I, and where, for each user $i = 1, 2$, the fading processes $\{\hat{H}_i[k]\}_{k\in\mathbb{Z}}$ and $\{\tilde{H}_i[k]\}_{k\in\mathbb{Z}}$ are sequences of i.i.d. complex random variables, the former with mean μ_i and variance $\hat{V}_i$, and the latter with mean zero and variance $\tilde{V}_i$. We assume that the processes $\{\hat{H}_i[k]\}_{k\in\mathbb{Z}}$, $\{\tilde{H}_i[k]\}_{k\in\mathbb{Z}}$ $(i = 1, 2)$ and $\{Z[k]\}_{k\in\mathbb{Z}}$ are independent of each other and of the input sequences $\{X_i[k]\}_{k\in\mathbb{Z}}$, $i = 1, 2$. As in Section I, we assume that both transmitter and receiver are cognizant of the distributions of $\{\hat{H}_i[k]\}_{k\in\mathbb{Z}}$, $\{\tilde{H}_i[k]\}_{k\in\mathbb{Z}}$ $(i = 1, 2)$ and $\{Z[k]\}_{k\in\mathbb{Z}}$, and that the receiver is, in addition, cognizant of the realizations of $\{\hat{H}_i[k]\}_{k\in\mathbb{Z}}$, $i = 1, 2$.

The capacity region of the above channel (6) under the power constraints P_1 and P_2 is given by the closure of the convex hull of all rates (R_1, R_2) satisfying

$$R_1 \le I(X_1; Y|X_2, \hat{\mathbf{H}}) \triangleq I_{1|2} \quad (7a)$$

$$R_2 \le I(X_2; Y|X_1, \hat{\mathbf{H}}) \triangleq I_{2|1} \quad (7b)$$

$$R_1 + R_2 \le I(X_1, X_2; Y|\hat{\mathbf{H}}) \triangleq I_{\Sigma} \quad (7c)$$

for some product distributions of (X_1, X_2) satisfying $\mathsf{E}[|X_1|^2] \le P_1$ and $\mathsf{E}[|X_2|^2] \le P_2$ [5].

In [2, Equations (69)–(71)], an inner bound on the capacity region was derived by assuming zero-mean real Gaussian channel inputs and by lower-bounding the mutual informations $I_{1|2}$, $I_{2|1}$ and I_{Σ} using worst-case noise bounds like (4). In the following, we will derive an improved capacity inner bound (for complex signalling) by evaluating (7a)–(7c) for zero-mean, circularly-symmetric, complex Gaussian channel inputs of respective powers P_1 and P_2, and by using Médard's lower bound (4) together with the above presented rate-splitting approaches. Specifically, we follow the approach by Rimoldi and Urbanke [4] to characterize points (R_1, R_2) on the dominant face of (7a)–(7c), i.e., points satisfying

$$\begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = (1-\alpha)\begin{bmatrix} I_{\Sigma} - I_{2|1} \\ I_{2|1} \end{bmatrix} + \alpha \begin{bmatrix} I_{1|2} \\ I_{\Sigma} - I_{1|2} \end{bmatrix} \quad (8)$$

for some $0 \le \alpha \le 1$, by single-user constraints for each R_1 and R_2. We then follow the rate-splitting approach presented in [3] to derive evaluable lower bounds on these single-user constraints.

To illustrate this approach, let us split User 1 into two virtual users, i.e., let $X_1 = X_{11} + X_{12}$, where X_{11} and X_{12} are independent, zero-mean, circularly symmetric, complex Gaussian random variables of respective powers $(1-\beta)P_1$ and βP_1. By performing successive decoding of X_{11}, X_2 and X_{12} (in this order), we can achieve the rates

$$R_{11} = I(X_{11}; Y|\hat{\mathbf{H}}) \quad (9a)$$

$$R_{12} = I(X_{12}; Y|\hat{\mathbf{H}}, X_{11}, X_2) \quad (9b)$$

$$R_2 = I(X_2; Y|\hat{\mathbf{H}}, X_{11}) \quad (9c)$$

giving rise to the single-user constraints

$$R_1 \le I(X_{11}; Y|\hat{\mathbf{H}}) + I(X_{12}; Y|\hat{\mathbf{H}}, X_{11}, X_2) \quad (10a)$$

$$R_2 \le I(X_2; Y|\hat{\mathbf{H}}, X_{11}). \quad (10b)$$

The mutual informations on the right-hand side (RHS) of (10a)–(10b) can then be lower-bounded following the rate-splitting approach presented in [3]. In this example, we first decode all layers of X_{11}, then all layers of X_2, and finally all layers of X_{12}. By introducing more than two virtual users, we can construct different decoding orders that potentially give rise to sharper inner bounds.

III. Power Allocations and Inner Bounds

The most general rate-splitting scheme on the two-user MAC can be represented as follows: the transmit signals of User $i = 1, 2$ are written as sums of independent, zero-mean, circularly-symmetric, complex Gaussian random variables $X_{i,\ell}$, $\ell = 1, \dots, L$ with respective powers $P_{i,\ell} \ge 0$

[2] The virtual users correspond to the layers in [3].

summing up to P_i, i.e.,

$$X_i = \sum_{\ell \in L} X_{i,\ell} \quad \text{and} \quad P_i = \sum_{\ell \in L} P_{i,\ell}. \tag{11}$$

The signals are decoded in an alternating decoding order

$$X_{1,1}, X_{2,1}, X_{1,2}, X_{2,2}, \ldots, X_{1,L}, X_{2,L}. \tag{12}$$

This yields the rate pair

$$J_1 = \textstyle\sum_{\ell=1}^{L} I\big(X_{1,\ell}; Y \mid \mathbf{X}_1^{\ell-1}, \mathbf{X}_2^{\ell-1}, \hat{\mathbf{H}}\big) \tag{13a}$$

$$J_2 = \textstyle\sum_{\ell=1}^{L} I\big(X_{2,\ell}; Y \mid \mathbf{X}_1^{\ell}, \mathbf{X}_2^{\ell-1}, \hat{\mathbf{H}}\big) \tag{13b}$$

where $\mathbf{X}_i^j$ stands for the collection $X_{i,1}, \ldots, X_{i,j}$. Note that this decoding order incurs no loss in generality, since setting a power $P_{i,\ell}$ to zero effectively suppresses the decoding step. With the decoding order held fixed, any rate-splitting scheme is fully described by the power allocations $\{P_{i,\ell}\}$. However, we shall find it convenient to define power allocations via so-called *layering functions*.

Definition 1. *A continuous surjective non-decreasing function $K_i \colon [0;1] \to [0;1]$ is called a layering function for user i. The set of layering functions is denoted as $\mathcal{K}$.*

We shall define a rate-splitting scheme by the pair of layering functions $\mathbf{K} = (K_1, K_2) \in \mathcal{K}^2$ and the number of layers L. The corresponding power allocations can then be obtained by

$$P_{i,\ell} = P_i\Big(K_i\big(\tfrac{\ell}{L}\big) - K_i\big(\tfrac{\ell-1}{L}\big)\Big). \tag{14}$$

Note that $\mathbf{K}$ does not depend on L.

A. Infinite-layer rate region

Upon applying Médard's bound on each summand on the RHS of (13a)–(13b), a given rate-splitting scheme $(\mathbf{K}, L)$ yields an achievable-rate pair $\underline{\mathbf{J}}(\mathbf{K}, L) \triangleq (\underline{J}_1(\mathbf{K}, L), \underline{J}_2(\mathbf{K}, L))$. The following theorem shows that, for any $\mathbf{K}$, the supremum over all rate pairs is approached as L tends to infinity.

Theorem 1. *For every pair of layering functions $\mathbf{K}$, the supremum of $\underline{J}_i(\mathbf{K}, L)$ over the number of layers is given by the Lebesgue-Stieltjes integral*

$$\sup_{L \in \mathbb{N}} \underline{J}_i(\mathbf{K}, L) = \lim_{L \to \infty} \underline{J}_i(\mathbf{K}, L) = \int_0^1 f_i(\zeta)\, \mathrm{d}K_i(\zeta) \tag{15}$$

with

$$f_i(\zeta) = \mathsf{E}\left[\frac{|\hat{H}_i|^2 P_i}{\sigma^2 + \sum_{j=1}^{2}[\tilde{V}_j P_j K_j(\zeta) \Xi_j + (|\hat{H}_j|^2 + \tilde{V}_j) P_j \bar{K}_j(\zeta)]}\right]$$

where Ξ_1 and Ξ_2 are two independent unit-mean exponentially distributed random variables, and $\bar{K}_i(\zeta) \triangleq 1 - K_i(\zeta)$. We shall denote this infinite-layering limit (15) *as $\underline{J}_i^{\infty}(\mathbf{K})$.*

Proof outline: The proof is a generalization of the proof of [3, Theorem 4] and hinges on similar ideas. The main difference is that, in the single-user setting in [3], the achievable rate converges to an expression that does not depend on the layering function. This allows for a simplified analysis where L-variate power allocations are approximated by N-variate (for N sufficiently large) equi-power allocations

$$P_1 = \ldots = P_N = \frac{P}{N}. \tag{16}$$

In contrast, for the fading MAC, the infinite-layering limit (15) depends on the pair $\mathbf{K}$ of layering functions, so a refined analysis is required. ■

Note that [3, Theorem 4] follows from Theorem 1 by setting $P_2 = 0$ and by the change of variable $\xi = K_1(\zeta)$.

B. Vertices of the rate region

By a change of variable applied to the integral on the RHS of (15), it can be shown that $\underline{J}_i^{\infty}(\mathbf{K})$ can be written as

$$\underline{J}_i^{\infty}(K_1, K_2) = \underline{J}_i^{\infty}(\widetilde{K}_1, \widetilde{K}_2). \tag{17}$$

where

$$\widetilde{K}_1(\zeta) \triangleq \zeta + \Lambda(\zeta), \quad \zeta \in [0;1] \tag{18a}$$

$$\widetilde{K}_2(\zeta) \triangleq \zeta - \Lambda(\zeta), \quad \zeta \in [0;1] \tag{18b}$$

for some function $\Lambda \colon [0;1] \to \big[-\frac{1}{2}; \frac{1}{2}\big]$ satisfying

$$\Lambda(0) = \Lambda(1) = 0 \tag{19a}$$

and

$$\sup_{0 \le \zeta_1 < \zeta_2 \le 1} \frac{|\Lambda(\zeta_2) - \Lambda(\zeta_1)|}{\zeta_2 - \zeta_1} \le 1. \tag{19b}$$

This allows us to write $\underline{J}_i^{\infty}(\mathbf{K})$ as a functional of one function (Λ) instead of two (K_1 and K_2), i.e., $\underline{J}_i^{\infty}(\mathbf{K}) = \underline{J}_i^{\infty}(\Lambda)$.

Definition 2. *A function $\Lambda \colon [0;1] \to \big[-\frac{1}{2}; \frac{1}{2}\big]$ with border values $\Lambda(0) = \Lambda(1) = 0$ satisfying the Lipschitz condition*

$$\sup_{0 \le \zeta_1 < \zeta_2 \le 1} \frac{|\Lambda(\zeta_2) - \Lambda(\zeta_1)|}{\zeta_2 - \zeta_1} \le 1 \tag{20}$$

is called a relative layering function. *The set of relative layering functions is denoted as $\mathcal{L}$.*

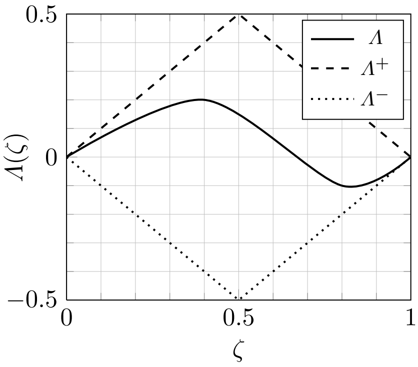


Fig. 1. Example of a relative layering function.

Figure 1 shows three examples of relative layering functions. The relative layering function Λ has the following interpretation: having decoded a proportion $\zeta \in [0;1]$ of the overall signal power $P_1 + P_2$, the value $2\Lambda(\zeta)$ quantifies the power by which the first user precedes (or lags behind, if $\Lambda(\zeta)$ is negative) the second user.

Writing the infinite-layering limit as a function of a relative layering function allows us to establish the following monotonicity result, which will be used later to determine the vertices of the best achievable-rate region $\underline{\mathcal{J}}$, maximized over all number of layers and over all possible layering functions.

Theorem 2. *Let the relative layering functions Λ and $\tilde{\Lambda}$ satisfy*

$$\Lambda(\zeta) \leq \tilde{\Lambda}(\zeta), \quad 0 \leq \zeta \leq 1 \tag{21}$$

with the inequality being strict for at least one $0 \leq \zeta \leq 1$. Then

$$\begin{array}{rcl} \underline{J}_2^\infty(\Lambda) & \leq & \underline{J}_2^\infty(\tilde{\Lambda}) \\ \underline{J}_1^\infty(\Lambda) & \geq & \underline{J}_1^\infty(\tilde{\Lambda}). \end{array} \tag{22}$$

Proof outline: Using a convexity argument, it can be shown that there exists a partial ordering for the layering functions according to which $K_1(\zeta) \leq \tilde{K}_1(\zeta)$ for all $0 \leq \zeta \leq 1$ implies $\underline{J}_1(K_1, K_2) \geq \underline{J}_1(\tilde{K}_1, K_2)$ and $\underline{J}_2(K_1, K_2) \leq \underline{J}_2(\tilde{K}_1, K_2)$. By an appropriate transformation (using variable substitutions in the Lebesgue-Stieltjes integral), the property is carried over to $\underline{J}_i^\infty(\Lambda)$, $i = 1, 2$, yielding (22). ■

Theorem 2 suggests that successive decoding penalizes users decoded first, while it benefits users decoded last.

A direct implication of Theorem 2 is that the vertices of the rate region $\underline{\mathcal{J}}$ are obtained for the extremal functions

$$\Lambda^+(\zeta) \triangleq \begin{cases} \zeta & \text{for } 0 \leq \zeta \leq \frac{1}{2} \\ 1 - \zeta & \text{for } \frac{1}{2} \leq \zeta \leq 1 \end{cases} \tag{23}$$

and $\Lambda^-(\zeta) = -\Lambda^+(\zeta)$, $0 \leq \zeta \leq 1$.

Corollary 1. *The relative layering functions Λ^+ and Λ^- satisfy*

$$\underline{J}_2(\Lambda^+) = \sup_{\Lambda \in \mathcal{L}} \underline{J}_2^\infty(\Lambda) \tag{24a}$$

$$\underline{J}_1(\Lambda^-) = \sup_{\Lambda \in \mathcal{L}} \underline{J}_1^\infty(\Lambda). \tag{24b}$$

While Theorem 2 provides an easy handle on the vertices of $\underline{\mathcal{J}}$, it is difficult to investigate the set of points in $\underline{\mathcal{J}}$ of maximal sum rate that are not vertices. To better understand the behavior of these points, we define four families of relative layering functions $\Lambda_{k,\alpha}$, $k = 1, \ldots, 4$ parametrized by a scalar α where $\Lambda_{k,\alpha}$ is continuous in α and the extremal functions Λ^+ and Λ^- are contained in each family. By varying α, we can move from one vertex point to the other. It is unknown whether any of these functions achieves the maximal sum rate for α's for which $\Lambda_{k,\alpha}$ is neither Λ^+ nor Λ^-.

We define $\Lambda_{k,\alpha}$ as integrals $\Lambda_{k,\alpha}(\zeta) = \int_0^\zeta \Lambda'_{k,\alpha}(z)\,\mathrm{d}z$ over their respective derivatives:

$$\Lambda'_{1,\alpha}(z) = \alpha\Big(\mathbb{I}_{[0;1/2[}(z) - \mathbb{I}_{[1/2;1]}(z)\Big),\ \alpha \in [-1;1]$$

$$\Lambda'_{2,\alpha}(z) = \mathbb{I}_{[0;\alpha[\cup[\alpha+(1/2);1]}(z) - \mathbb{I}_{[\alpha;\alpha+(1/2)[}(z),\ \alpha \in [0;\tfrac{1}{2}]$$

$$\Lambda'_{3,\alpha}(z) = \mathrm{sgn}(\alpha)\Big(\mathbb{I}_{[0;|\alpha|[}(z) - \mathbb{I}_{[1-|\alpha|;1]}(z)\Big),\ \alpha \in [-\tfrac{1}{2};\tfrac{1}{2}]$$

$$\Lambda'_{4,\alpha}(z) = \mathbb{I}_{[0;\alpha[\cup[1/2;1-\alpha[}(z) - \mathbb{I}_{[\alpha;1/2[\cup[1-\alpha;1]}(z),\ \alpha \in [0;\tfrac{1}{2}].$$

Here, $\mathrm{sgn}(\cdot)$ denotes the sign function and $\mathbb{I}_{\mathcal{A}}$ denotes the indicator function of the set $\mathcal{A}$.

Figure 2 shows the sum rates achieved by $\Lambda_{k,\alpha}$, $k = 1, \ldots, 4$ for a symmetric fading MAC with parameters $P_1 = P_2 = 10$ and $\sigma^2 = 1$, plotted against R_1. The channel components $\hat{H}_i$, $i = 1, 2$ are both zero-mean, circularly-symmetric, complex Gaussian random variables with variance $\frac{1}{2}$. Moreover, we choose $\tilde{V}_1 = \tilde{V}_2 = \frac{1}{2}$. Observe that the sum rate critically depends on the chosen rate-splitting scheme. Further observe that $\Lambda_{2,\alpha}$ does not achieve its largest sum rate at a vertex point. Consequently, there exist rate pairs that cannot be achieved by time sharing between the vertices.

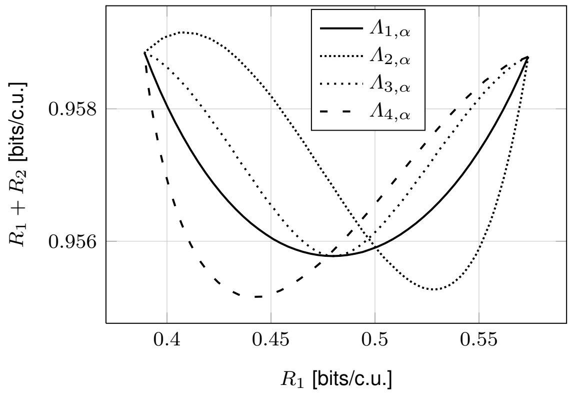


Fig. 2. Comparison of different relative layering functions.

IV. Conclusion

We have blended the rate-splitting approaches by [3] and [4] in order to derive a novel inner bound on the capacity region of the fading MAC with imperfect receiver CSI. We have shown that, for every pair of layering functions $\mathbf{K}$, the supremum of this inner bound is approached as the number of layers tends to infinity, and have derived an integral expression for this supremum. In addition, we have determined the vertices of the best inner bound, maximized over the number of layers and all layering functions. Our analysis has revealed that, in contrast to the setting with *perfect* receiver CSI, certain rate pairs cannot be achieved by time sharing between the vertices.

References

[1] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, pp. 2619–2692, 1998.

[2] M. Médard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

[3] A. Pastore, T. Koch, and J. Fonollosa, "A rate-splitting approach to fading channels with imperfect channel-state information," *IEEE Transactions on Information Theory*, 2013, submitted. [Online]. Available: http://arxiv.org/abs/1301.6120

[4] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[5] S. Shamai and A. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels—Part I," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1877–1894, Nov. 1997.

Expurgated Random-Coding Ensembles: Exponents, Refinements and Connections

Jonathan Scarlett[1], Li Peng[1], Neri Merhav[2], Alfonso Martinez[3] and Albert Guillén i Fàbregas[134]
[1]University of Cambridge, [2]Technion, I.I.T., [3]Universitat Pompeu Fabra, [4]ICREA
e-mail: {jms265,lp327}@cam.ac.uk, merhav@ee.technion.ac.il, {guillen,alfonso.martinez}@ieee.org

***Abstract*—This paper studies expurgated random-coding bounds and exponents for channels with maximum-metric decoding. A simple non-asymptotic bound is shown to attain an exponent which coincides with that of Csiszár and Körner for discrete memoryless channels, while remaining valid for continuous alphabets. Using an alternative approach based on statistical-mechanical methods, an exponent for more general channels and decoding metrics is given.**

I. Introduction

Achievable performance bounds for channel coding are typically obtained by analyzing the average error probability of an ensemble of codebooks with independently generated codewords. At low rates, the error probability of the best code in the ensemble can be significantly smaller than the average. In such cases, better performance bounds are obtained by considering an ensemble in which a subset of the randomly generated codewords are expurgated from the codebook.

The main approaches to obtaining expurgated bounds and exponents are those of Gallager [1, Sec. 5.7] and Csiszár-Körner-Marton [2, Ex. 10.18] [3]. Gallager's approach is based on simple inequalities such as Markov's inequality, and has the advantage of being simple and applicable to channels with continuous alphabets. On the other hand, the techniques of [2], [3] are based on the method of types, and are applicable to channels with input constraints. While the exponents of [1]–[3] all coincide after optimizing the input distribution, the exponents of [2], [3] can be higher than that of [1] for a given input distribution [4].

In this paper, we provide techniques that attain the best of each of the above approaches. Our main contributions are as follows:

1) We give the precise connection between the exponents of [1]–[3] using Lagrange duality [5], as well as generalizing the exponents of [1], [2] to the setting of mismatched decoding [3], [6].
2) We show that variations of Gallager's techniques can be used to obtain a simple non-asymptotic bound which recovers the exponent of [2], [3], as well as a generalization to the case of continuous alphabets.
3) We present an alternative analysis technique based on statistical-mechanical methods (e.g. see [7], [8]), and use it to derive an achievable exponent for general channels and metrics (e.g. channels with memory).

Due to space constraints, full proofs of the main results are omitted; details can be found in [4].

A. Notation

We use bold symbols for vectors (e.g. $\boldsymbol{x}$), and denote the corresponding i-th entry using a subscript (e.g. x_i). The set of all empirical distributions (i.e. types [2, Ch. 2]) on a given alphabet, say $\mathcal{X}$, is denoted by $\mathcal{P}_n(\mathcal{X})$. For a given type $Q \in \mathcal{P}_n(\mathcal{X})$, the type class $T^n(Q)$ is defined to be the set of all sequences in $\mathcal{X}^n$ with type Q. For two positive sequences f_n and g_n, we write $f_n \doteq g_n$ if $\lim_{n\to\infty} \frac{1}{n}\log\frac{f_n}{g_n} = 0$, and we write $f_n \,\dot{\le}\, g_n$ if $\limsup_{n\to\infty} \frac{1}{n}\log\frac{f_n}{g_n} \le 0$, and analogously for $\dot{\ge}$. All logarithms have base e, and all rates are in units of nats. We define $[c]^+ = \max\{0, c\}$, and denote the indicator function by $\mathbb{1}\{\cdot\}$.

B. System Setup

We consider block coding over a memoryless channel $W^n(\boldsymbol{y}|\boldsymbol{x}) \triangleq \prod_{i=1}^n W(y_i|x_i)$ with alphabets $\mathcal{X}$ and $\mathcal{Y}$. The encoder takes as input a message m uniformly distributed on the set $\{1,\dots,M\}$, and transmits the corresponding codeword $\boldsymbol{x}^{(m)}$ of length n. Given $\boldsymbol{y}$, the decoder forms the estimate

$$\hat{m} = \operatorname*{arg\,max}_{j\in\{1,\dots,M\}} q^n(\boldsymbol{x}^{(j)}, \boldsymbol{y}), \tag{1}$$

where $q^n(\boldsymbol{x},\boldsymbol{y}) \triangleq \prod_{i=1}^n q(x_i, y_i)$ for some non-negative function $q(x,y)$. When $q(x,y) = W(y|x)$, (1) is the optimal maximum-likelihood (ML) decoding rule. For other decoding metrics, this setting is that of *mismatched decoding* [3], [6], which is relevant when ML decoding is not feasible.

Except where stated otherwise, we assume that the codewords are unconstrained. However, in some cases we will consider input constraints of the form

$$\frac{1}{n}\sum_{i=1}^n c(x_i) \le \Gamma, \tag{2}$$

where $c(\cdot)$ is referred to as a cost function, and Γ is a constant.

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme (PEOPLE-2011-CIG) under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03. The work of N. Merhav was partially supported by the Israeli Science Foundation (ISF), grant no. 412/12.

C. Expurgated Exponents and Duality

We will primarily be interested in the following exponent, which was derived in [3] for the case of finite alphabets:[1]

$$E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) \triangleq \min_{\substack{P_{X\overline{X}Y}\in\mathcal{T}^{\mathrm{cc}}(Q)\\ I_P(X;\overline{X})\le R}} D(P_{X\overline{X}Y}\|Q\times Q\times W) - R, \quad (3)$$

where

$$\mathcal{T}^{\mathrm{cc}}(Q) \triangleq \Big\{P_{X\overline{X}Y} \,:\, P_X = Q, P_{\overline{X}} = Q, \\ \mathbb{E}_P[\log q(\overline{X},Y)] \ge \mathbb{E}_P[\log q(X,Y)]\Big\} \quad (4)$$

and Q is an arbitrary input distribution. The following theorem links this exponent with those given in [1], [2].

Theorem 1. *For any input distribution Q and rate R, we have*

$$\begin{aligned} E_{\mathrm{ex}}^{\mathrm{cc}}(Q,R) &= \sup_{s\ge0} \min_{\substack{P_{X\overline{X}}\,:\,P_X=Q,P_{\overline{X}}=Q,\\ I_P(X;\overline{X})\le R}} \\ &\qquad \mathbb{E}_P[d_s(X,\overline{X})] + I_P(X;\overline{X}) - R \quad (5)\\ &= \sup_{\rho\ge1} E_{\mathrm{x}}^{\mathrm{cc}}(Q,\rho) - \rho R, \quad (6)\end{aligned}$$

where

$$d_s(x,\overline{x}) \triangleq -\log\sum_y W(y|x)\left(\frac{q(\overline{x},y)}{q(x,y)}\right)^s \quad (7)$$

$$E_{\mathrm{x}}^{\mathrm{cc}}(Q,\rho) \triangleq \sup_{s\ge0,a(\cdot)} \\ -\rho\sum_x Q(x)\log\sum_{\overline{x}} Q(\overline{x})\frac{e^{a(\overline{x})}}{e^{a(x)}}e^{-d_s(x,\overline{x})/\rho}. \quad (8)$$

The right-hand side of (5) can be considered a generalization of the exponent in [2] to the setting of mismatched decoding; the exponent for ML decoding is recovered by setting $s=\frac{1}{2}$. Theorem 1 shows that the exponents of [2] and [3] are equivalent even when Q is fixed; the equivalence for the optimal Q is well-known [3].

The right-hand side of (8) resembles Gallager's E_{x} function, which can be extended to the mismatched setting to obtain

$$E_{\mathrm{ex}}^{\mathrm{iid}}(Q,R) \triangleq \sup_{\rho\ge1} E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho) - \rho R, \quad (9)$$

where

$$E_{\mathrm{x}}^{\mathrm{iid}}(Q,\rho) \triangleq \sup_{s\ge0} -\rho\log\sum_{x,\overline{x}} Q(x)Q(\overline{x})e^{-d_s(x,\overline{x})/\rho}. \quad (10)$$

We immediately see that $E_{\mathrm{ex}}^{\mathrm{cc}} \ge E_{\mathrm{ex}}^{\mathrm{iid}}$. While equality holds under the optimal Q for ML decoding [3], the inequality can be strict for a suboptimal Q and/or a suboptimal decoding rule.

In this paper, we seek alternative derivations of the stronger exponent $E_{\mathrm{ex}}^{\mathrm{cc}}$ which are not sensitive to the assumption of finite alphabets, and which remain valid for channels with input constraints.

II. Analysis Using Finite-Length Bounds

Let $p_{e,m}(\mathcal{C})$ denote the error probability for a given codebook $\mathcal{C}$ given that the m-th codeword was sent, and let $p_e(\mathcal{C}) \triangleq \max_m p_{e,m}(\mathcal{C})$. We fix an arbitrary codeword distribution $P_{\boldsymbol{X}}$ and define

$$(\boldsymbol{X},\boldsymbol{Y},\overline{\boldsymbol{X}}) \sim P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}). \quad (11)$$

Stated in a general form, Gallager's analysis proves the existence of a codebook $\mathcal{C}$ of size $M > M'\frac{\eta}{1+\eta}$ such that

$$f\big(p_e(\mathcal{C})\big) \le (1+\eta)\mathbb{E}\big[f(p_{e,m}(\mathsf{C}))\big] \quad (12)$$

for any $\eta\ge0$ and non-negative function $f(\cdot)$, where C is a *random* codebook with M' codewords drawn independently from the distribution $P_{\boldsymbol{X}}$. In particular, we obtain

$$p_e(\mathcal{C}) \le \Big(2\mathbb{E}\big[p_{e,m}(\mathsf{C})^{1/\rho}\big]\Big)^\rho \quad (13)$$

by choosing $\eta=1$ and $f(\cdot)=(\cdot)^{1/\rho}$, where C contains $2M-1$ codewords. The following non-asymptotic bound follows from (13) using the union bound and the inequality

$$\Big(\sum_i a_i\Big)^{1/\rho} \le \sum_i a_i^{1/\rho} \quad (\rho\ge1). \quad (14)$$

Theorem 2. *For any pair (n,M) and codeword distribution $P_{\boldsymbol{X}}$, there exists a codebook $\mathcal{C}$ with M codewords of length n whose maximal error probability satisfies*

$$p_e(\mathcal{C}) \le \inf_{\rho\ge1} \\ \left(4(M-1)\mathbb{E}\left[\mathbb{P}\Big[q^n(\overline{\boldsymbol{X}},\boldsymbol{Y}) \ge q^n(\boldsymbol{X},\boldsymbol{Y})\,\Big|\,\boldsymbol{X},\overline{\boldsymbol{X}}\Big]^{1/\rho}\right]\right)^\rho. \quad (15)$$

The bound in (15) extends immediately to general channels and metrics (e.g. channels with memory), and can be considered an analog of the random-coding union (RCU) bound given by Polyanskiy *et al.* [9]. In the remainder of the section, we present the resulting exponents for various ensembles in the memoryless case.

i.i.d. ensemble: Choosing the i.i.d. distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i), \quad (16)$$

we can use Markov's inequality to weaken (15) and obtain the exponent $E_{\mathrm{ex}}^{\mathrm{iid}}(Q,R)$. This approach does not rely on the alphabets being finite, but it is unsuitable for input-constrained channels, since in all non-trivial cases there is a non-zero probability of violating the constraint.

Constant-composition ensemble: Suppose that $|\mathcal{X}|$ and $|\mathcal{Y}|$ are finite, and consider the constant-composition codeword distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|T^n(Q_n)|}\mathbb{1}\big\{\boldsymbol{x}\in T^n(Q_n)\big\}, \quad (17)$$

where Q_n is a type with the same support as Q such that $|Q_n(x)-Q(x)| \le \frac{1}{n}$ for all x. By expanding (15) in terms

[1] The notation $Q\times Q\times W$ denotes the distribution $Q(x)Q(\overline{x})W(y|x)$.

of types and applying standard properties [2, Ch. 10], we can derive the exponent

$$\sup_{\rho \geq 1} \min_{P_{X\overline{X}Y} \in \mathcal{T}^{\mathrm{cc}}(Q)} D(P_{X\overline{X}Y} \| P_{X\overline{X}} \times W) + \rho\big(I_P(X;\overline{X}) - R\big). \tag{18}$$

Using the minimax theorem [10], we recover $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (3). This provides a simple alternative proof to the one given in [3] based on graph decomposition techniques.

Cost-constrained ensemble: In the case of continuous alphabets and input constraints (see (2)), we can derive $E_{\mathrm{ex}}^{\mathrm{cc}}$ using the cost-constrained codeword distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu_n} \prod_{i=1}^{n} Q(x_i) \mathbb{1}\{\boldsymbol{x} \in \mathcal{D}_n\}, \tag{19}$$

where

$$\mathcal{D}_n \triangleq \bigg\{ \boldsymbol{x} \,:\, \frac{1}{n}\sum_{i=1}^{n} c(x_i) \leq \Gamma, \\ \bigg|\frac{1}{n}\sum_{i=1}^{n} a_l(x_i) - \phi_l\bigg| \leq \frac{\delta}{n},\, l = 1, \dots, L \bigg\}, \tag{20}$$

and where δ is a positive constant, $\{a_l(\cdot)\}_{l=1}^{L}$ are arbitrary *auxiliary cost functions* with means $\phi_l \triangleq \mathbb{E}_Q[a_l(X)]$, and μ_n is a normalizing constant.

One can show that $\mu_n \doteq 1$ provided that $\mathbb{E}_Q[c(X)] \leq \Gamma$. $\mathbb{E}_Q[c(X)^2] < \infty$ and $\mathbb{E}_Q[a_l(X)^2] < \infty$ for $l = 1, \cdots, L$ [11]. Assuming these conditions are satisfied, we can analyze (15) similarly to the case of random coding without expurgation [11] to obtain the exponent

$$E_{\mathrm{ex}}^{\mathrm{cost}}(Q, R, \{a_l\}) \triangleq \sup_{\rho \geq 1} E_{\mathrm{x}}^{\mathrm{cost}}(Q, \rho, \{a_l\}) - \rho R, \tag{21}$$

where[2]

$$E_{\mathrm{x}}^{\mathrm{cost}}(Q, R, \{a_l\}) \triangleq \sup_{s \geq 0, \{r_l\}, \{\overline{r}_l\}} \\ -\rho \log \sum_{x,\overline{x}} Q(x) Q(\overline{x}) \frac{e^{\sum_{l=1}^{L} \overline{r}_l (a_l(\overline{x}) - \phi_l)}}{e^{\sum_{l=1}^{L} r_l (a_l(x) - \phi_l)}} e^{-d_s(x,y)/\rho}, \tag{22}$$

and the constants $\{r_l\}$ and $\{\overline{r}_l\}$ are arbitrary. Roughly speaking, the additional factor in (22) compared to (10) is obtained using the fact that the empirical mean of each auxiliary cost is close to the true mean.

Finally, we claim that (22) reduces to (8) when $L = 2$ and $a_1(\cdot), a_2(\cdot)$ are optimized. This is easily shown by setting $r_l = \overline{r}_l = 1$ for $l = 1, 2$, choosing $a_2(\cdot)$ such that Jensen's inequality holds with equality when $\sum_x Q(x)$ is taken outside the logarithm, and using the definition of ϕ_l to write

$$-\sum_x Q(x) \log \frac{e^{-\phi_1}}{e^{a_2(x) - \phi_2}} = -\sum_x Q(x) \log e^{-a_1(x)}. \tag{23}$$

In summary, this derivation shows that, under mild technical assumptions, (6) is an achievable exponent even in the case of infinite or continuous alphabets, provided that Q satisfies $\mathbb{E}_Q[c(X)] \leq \Gamma$ in accordance with (2).

III. Analysis Using Enumerator Functions

In this section, we present an alternative method for deriving expurgated exponents which is based on statistical-mechanical methods (e.g. see [7], [8]). In [4], we provide two variations of this approach depending on whether the alphabets are discrete or continuous. We begin here by discussing the discrete case.

Applying the union bound to (13), we obtain

$$p_e \leq \left(2\mathbb{E}\left[\left(\sum_{\overline{m} \neq m} \mathbb{P}\left[\frac{q^n(\boldsymbol{X}^{(\overline{m})}, \boldsymbol{Y})}{q^n(\boldsymbol{X}^{(m)}, \boldsymbol{Y})} \geq 1 \,\Big|\, \mathsf{C} \right] \right)^{1/\rho} \right] \right)^{\rho}, \tag{24}$$

where $\big\{\boldsymbol{X}^{(j)}\big\}_{j=1}^{2M-1} \sim \prod_{j=1}^{M} P_{\boldsymbol{X}}(\boldsymbol{x}^{(j)})$ are the random codewords in C. For any codeword distribution $P_{\boldsymbol{X}}(\boldsymbol{x})$ depending only on the type of $\boldsymbol{x}$, we can perform an exponentially tight analysis of (24) using type enumerators [7]. For the constant-composition ensemble (see (17)), we obtain $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (3). Although the exponent is the same as that obtained via Theorem 2, the type enumerator approach guarantees exponential tightness starting from an earlier step.

On the other hand, for the i.i.d. ensemble (see (16)), we show in [4] that (24) yields an exponent which is strictly higher in general than that obtained via Theorem 2. It follows that the inequality in (14) is not exponentially tight in general, thus motivating the more refined analysis of (24).

In the remainder of the section, we consider a more general approach which remains applicable in the continuous case. We assume that each codeword must satisfy (2), and that

$$\lim_{\gamma \to \infty} \frac{1}{\gamma} \log\log \frac{1}{\pi(\gamma)} = 0, \tag{25}$$

where

$$\pi(\gamma) \triangleq \min_{(x,\overline{x})\,:\,c(x) \leq \gamma, c(\overline{x}) \leq \gamma} \mathbb{P}[Y_x \in \mathcal{E}(x, \overline{x})] \tag{26}$$

$$\mathcal{E}(x, \overline{x}) \triangleq \big\{ y \,:\, q(\overline{x}, y) \geq q(x, y) \big\}, \tag{27}$$

and in (26) we define $Y_x \sim W(\cdot|x)$. This assumption is mild and generally easily to verify. For example, for the power-constrained additive white Gaussian noise channel with ML decoding, $\pi(\cdot)$ only decays exponentially in γ, whereas (25) allows for nearly double-exponential rates of decay. See [4] for further discussion and examples.

The following theorem follows by applying (12) with a function of the form $f(\cdot) = f_n(\cdot) = \log(\cdot) + c_n$, where c_n is chosen such that $f_n(p_{e,m})$ is non-negative for all values of $p_{e,m}$ which can occur when (25) holds.

Theorem 3. *Fix $R > 0$ and consider a sequence of codebooks C_n containing $M_n' = \lfloor \exp(nR) \rfloor$ codewords which are generated independently according to $P_{\boldsymbol{X}}$. Under the assumption in* (25)*, there exists a sequence of codebooks $\mathcal{C}_n$ with M_n codewords such that*

$$\lim_{n \to \infty} \frac{1}{n} \log M_n = R \tag{28}$$

[2] In the case of continuous alphabets, the summations should be replaced by integrals.

and

$$p_e(\mathcal{C}_n) \,\dot{\le}\, \exp\Big(\mathbb{E}[\log p_{e,m}(\mathsf{C}_n)]\Big) \tag{29}$$

$$\le \exp\Big(\rho\mathbb{E}\Big[\log \mathbb{E}\big[p_{e,m}(\mathsf{C}_n)^{1/\rho} \,\big|\, \boldsymbol{X}^{(m)}\big]\Big]\Big), \tag{30}$$

where (30) *holds for any* $\rho > 0$.

Equation (30) can be thought of as improving on (13) to the fact that the expectation with respect to the transmitted codeword is outside the logarithm.

Applying the union bound and Markov's inequality to (30), we conclude that there exists a sequence of codebooks $\mathcal{C}_n$ of rate approaching R such that

$$p_e(\mathcal{C}_n) \,\dot{\le}\, \exp\Big(\mathbb{E}\big[\log A_n(R,\rho,\boldsymbol{X}^{(m)})\big]\Big), \tag{31}$$

where

$$A_n(R,\rho,\boldsymbol{X}^{(m)}) \triangleq \mathbb{E}\left[\bigg(\sum_{\overline{m}\ne m} e^{-d_s^n(\boldsymbol{X}^{(m)},\boldsymbol{X}^{(\overline{m})})}\bigg)^{1/\rho} \,\bigg|\, \boldsymbol{X}^{(m)}\right]^{\rho} \tag{32}$$

and $d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}}) \triangleq \sum_{i=1}^n d_s(x_i,\overline{x}_i)$. We fix $\delta > 0$ and write

$$\sum_{\overline{m}\ne m} e^{-d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})})} \le \sum_{k=0}^{\infty} e^{-nk\delta} N_m(k,\boldsymbol{x}), \tag{33}$$

where

$$N_m(k,\boldsymbol{x}) \triangleq \sum_{\overline{m}\ne m} \mathbb{1}\big\{nk\delta \le d_s^n(\boldsymbol{x},\boldsymbol{X}^{(\overline{m})}) < n(k+1)\delta\big\}. \tag{34}$$

The key observation which permits the subsequent analysis is that the maximum value of k for which $N_m(k,\boldsymbol{x}) \ne 0$ grows subexponentially in n; this can easily be verified using the assumption in (25). Applying this observation to (32) multiple times, we obtain

$$A_n(R,\rho,\boldsymbol{x}) \,\dot{\le}\, \max_{k\ge 0}\Big(\mathbb{E}\big[N_m(k,\boldsymbol{x})^{1/\rho}\big]\Big)^{\rho} e^{-nk\delta}. \tag{35}$$

We can further upper bound this expression by removing the lower inequality in the indicator function in (34). Letting $R(D,\boldsymbol{x})$ be any continuous function such that $\mathbb{P}\big[d_s^n(\boldsymbol{x},\overline{\boldsymbol{X}}) < nD\big] \,\dot{\le}\, e^{-nR(D,\boldsymbol{x})}$ uniformly in $\boldsymbol{x}$, it follows by treating the cases $R(D,\boldsymbol{x}) \le R$ and $R(D,\boldsymbol{x}) > R$ separately that

$$A_n(R,\rho,\boldsymbol{x}) \,\dot{\le}\, e^{-n\min\{E_1(R,\rho,\delta,\boldsymbol{x}),E_2(R,\delta,\boldsymbol{x})\}}, \tag{36}$$

where

$$E_1(R,\rho,\delta,\boldsymbol{x}) \triangleq \min_{k\,:\,R((k+1)\delta,\boldsymbol{x})\ge R} k\delta + \rho\big(R((k+1)\delta,\boldsymbol{x}) - R\big) \tag{37}$$

$$E_2(R,\delta,\boldsymbol{x}) \triangleq \min_{k\,:\,R((k+1)\delta,\boldsymbol{x})\le R} k\delta + R((k+1)\delta,\boldsymbol{x}) - R. \tag{38}$$

Finally, we obtain the following by taking $\delta \to 0$ and $\rho \to \infty$.

Theorem 4. *Under the assumption in* (25)*, the exponent*

$$E_{\mathrm{ex}}(R) \triangleq \mathbb{E}\left[\inf_{D\,:\,R(D,\boldsymbol{X})\le R} D + R(D,\boldsymbol{X}) - R\right] \tag{39}$$

is achievable for any continuous function $R(D,\boldsymbol{x})$ *such that* $\mathbb{P}\big[d_s^n(\boldsymbol{x},\overline{\boldsymbol{X}}) < nD\big] \,\dot{\le}\, e^{-nR(D,\boldsymbol{x})}$ *uniformly in* $\boldsymbol{x}$.

After a suitable modification of the definition of $d_s^n(\boldsymbol{x},\overline{\boldsymbol{x}})$, (39) extends immediately to general channels and metrics. The ability to simplify the exponent (e.g. to a single-letter expression) depends on the form of $R(D,\boldsymbol{x})$, which in turn depends strongly on the codeword distribution $P_{\boldsymbol{X}}$. In some cases, $P_{\boldsymbol{X}}$ can be chosen in such a way that $R(D,\boldsymbol{x})$ is the same for all $\boldsymbol{x}$ with $P_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, thus greatly simplifying (39).

Consider the cost-constrained ensemble given in (19) with $L = 1$, and assume analogously to Section III that $\mathbb{E}_Q[c(X)] \le \Gamma$, $\mathbb{E}[c(X)^2] < \infty$ and $\mathbb{E}[a_1(X)^2] < \infty$. Using standard Chernoff-type bounding techniques, we obtain

$$R(D,\boldsymbol{x}) = \sup_{t\ge 0,\overline{r}} \overline{r}\phi_1 - tD - \frac{1}{n}\sum_{i=1}^{n}\theta(x_i,\overline{r},t), \tag{40}$$

where

$$\theta(x,\overline{r},t) \triangleq \log \mathbb{E}_Q\big[e^{\overline{r}a_1(\overline{X}) - td_s(x,\overline{X})}\big]. \tag{41}$$

Substituting (40) into (39) and performing some manipulations, we obtain $E_{\mathrm{ex}}^{\mathrm{cc}}$ in the form given in (6), with the summations replaced by integrals where necessary. In contrast to Section III, we only require $L = 1$ instead of $L = 2$. However, this comes at the price of requiring (25) to hold.

References

[1] R. Gallager, *Information Theory and Reliable Communication.* John Wiley & Sons, 1968.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[3] ——, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.

[4] J. Scarlett, L. Peng, N. Merhav, A. Martinez, and A. Guillén i Fàbregas, "Expurgated random-coding ensembles: Exponents, refinements and connections," 2013, submitted to *IEEE Trans. Inf. Theory* [Online: http://arxiv.org/abs/1307.6679].

[5] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[6] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[7] N. Merhav, "Error exponents of erasure/list decoding revisited via moments of distance enumerators," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4439–4447, Oct. 2008.

[8] ——, "Statistical physics and information theory," *Foundations and Trends in Comms. and Inf. Theory*, vol. 6, no. 1-2, pp. 1–212, 2009.

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

[10] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci.*, vol. 39, pp. 42–47, 1953.

[11] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," submitted to *IEEE Trans. Inf. Theory* [Online: http://arxiv.org/abs/1303.6166].

Second-Order Rate of Constant-Composition Codes for the Gel'fand-Pinsker Channel

Jonathan Scarlett

Abstract—**This paper presents an achievable second-order coding rate for the discrete memoryless Gel'fand-Pinkser channel. The result is obtained using constant-composition random coding, and by using an asymptotically negligible fraction of the block to transmit the type of the state sequence.**

I. Introduction

In this paper, we present an achievable second-order coding rate [1]–[3] for channel coding with a random state known non-causally at the encoder, as studied by Gel'fand and Pinsker [4]. The alphabets of the input, output and state are denoted by $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{S}$ respectively, and each are assumed to be finite. The channel transition law is given by $W^n(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}) \triangleq \prod_{i=1}^n W(y_i|x_i,s_i)$, where n is the block length. The state sequence $\boldsymbol{S} = (S_1,\cdots,S_n)$ is assumed to be independent and identically distributed (i.i.d.) according to a distribution $\pi(s)$. The capacity is given by [4]

$$C = \max_{\mathcal{U},Q_{U|S},\phi(\cdot,\cdot)} I(U;Y) - I(U;S), \tag{1}$$

where the mutual informations are with respect to

$$P_{SUY}(s,u,y) = \pi(s)Q_{U|S}(u|s)W(y|\phi(u,s),s) \tag{2}$$

and the maximum is over all finite alphabets $\mathcal{U}$, conditional distributions $Q_{U|S}$ and functions $\phi : \mathcal{U}\times\mathcal{S}\to\mathcal{X}$.

We say that a triplet (n,M,ϵ) is achievable if there exists a code with block length n containing at least M messages and yielding an average error probability not exceeding ϵ, and we define $M^*(n,\epsilon) \triangleq \max\big\{M : (n,M,\epsilon) \text{ is achievable}\big\}$. Letting $P_{Y|U}$, P_Y, etc. denote the marginals of (2), we define the information densities

$$i(u,s) \triangleq \log\frac{Q_{U|S}(u|s)}{P_U(u)} \tag{3}$$

$$i(u,y) \triangleq \log\frac{P_{Y|U}(y|u)}{P_Y(y)} \tag{4}$$

with a slight abuse of notation.

Theorem 1. *Let $\mathcal{U}$, $Q_{U|S}$ and $\phi(\cdot,\cdot)$ by any set of capacity-achieving parameters in* (1)*, and let P_{SUY}, $i(u,s)$ and $i(u,y)$ be as given in* (2)*–*(4) *under these parameters. If $\mathbb{E}\big[\mathrm{Var}[i(U,Y)\,|\,U,S]\big] > 0$, then*

$$\log M^*(n,\epsilon) \geq nC - \sqrt{nV}\mathsf{Q}^{-1}(\epsilon) + O(\log n), \tag{5}$$

for $\epsilon\in(0,1)$, where

$$V \triangleq \mathbb{E}\big[\mathrm{Var}[i(U,Y)\,|\,U,S]\big] + \mathrm{Var}\big[\mathbb{E}[i(U,Y) - i(U,S)\,|\,S]\big] \tag{6}$$

$$= \mathrm{Var}[i(U,Y) - i(U,S)]. \tag{7}$$

Proof: We provide a number of preliminary results in Section II, and present the proof in Section III. ■

It should be noted that the equality in (7) holds under the capacity-achieving parameters, but more generally (7) is at least as high as (6), with strict inequality possible for suboptimal choices of $Q_{U|S}$.

To our knowledge, the only previous result on the second-order asymptotics for the present problem is that of Watanabe *et al.* [5] and Yassaee *et al.* [6], who used i.i.d. random coding. In [7], we show that for $\epsilon < \frac{1}{2}$ our second-order term is at least as good as that of [5], [6], with strict improvement possible. Furthermore, we show in [7] that Theorem 1 recovers, as a special case, the dispersion for channels with i.i.d. state known at both the encoder and decoder, which was derived in [8].

Notation: Bold symbols are used for vectors and matrices (e.g. $\boldsymbol{x}$), and the corresponding i-th entry of a vector is denoted with a subscript (e.g. x_i). The marginals of a joint distribution P_{XY} are denoted by P_X and P_Y. The empirical distribution (i.e. type [9, Ch. 2]) of a vector $\boldsymbol{x}$ is denoted by $\hat{P}_{\boldsymbol{x}}$. The set of all types of length n on an alphabet $\mathcal{X}$ is denoted by $\mathcal{P}_n(\mathcal{X})$. The set of all sequences of length n with a given type P_X is denoted by $T^n(P_X)$, and similarly for joint types. We make use of the standard asymptotic notations $O(\cdot)$ and $o(\cdot)$.

II. Preliminary Results

In this section, we present a number of preliminary results which will prove useful in the proof of Theorem 1. We assume that $\mathcal{U}$, $Q_{U|S}$ and $\phi(\cdot,\cdot)$ achieve the capacity in (1).

A. A Genie-Aided Setting

We prove Theorem 1 by first proving the following result for a genie-aided setting.

Theorem 2. *Theorem 1 holds true in the case that the empirical distribution $\hat{P}_{\boldsymbol{S}}$ of $\boldsymbol{S}$ is known at the decoder.*

To see that Theorem 2 implies Theorem 1, we use a technique which was proposed in [10]. We use the first $g(n) = K_0\log(n+1)$ symbols of the block to transmit the

J. Scarlett is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: jmscarlett@gmail.com).

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme (PEOPLE-2011-CIG) under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grant TEC2012-38800-C03-03.

type of the remaining $\tilde{n} = n - g(n)$ symbols. Using Gallager's random-coding bound [11, Sec. 5.6] and the fact that the number of such types is upper bounded by $(n+1)^{|\mathcal{S}|-1}$, it is easily shown that there exists a choice of K_0 such that the decoder estimates the state type correctly with probability $O\big(\frac{1}{n}\big)$. Thus, $\big(n - O(\log n), M, \epsilon - O\big(\frac{1}{n}\big)\big)$-achievability in the genie-aided setting implies (n, M, ϵ)-achievability in the absence of the genie. By performing a Taylor expansion of the square root and $\mathsf{Q}^{-1}(\cdot)$ function in (5), we obtain the desired result.

B. A Typical Set

We define a typical set of state types given by

$$\tilde{\mathcal{P}}_n = \left\{ P_S \in \mathcal{P}_n(\mathcal{S}) \,:\, \|P_S - \pi\|_\infty \le \sqrt{\frac{\log n}{n}} \right\}. \tag{8}$$

We will see the second-order performance is unaffected by types falling outside $\tilde{\mathcal{P}}_n$, due to the fact that [8, Lemma 22]

$$\mathbb{P}\big[\hat{P}_{\boldsymbol{S}} \notin \tilde{\mathcal{P}}_n\big] = O\Big(\frac{1}{n^2}\Big). \tag{9}$$

C. Approximations of Distributions

For each $P_S \in \mathcal{P}_n(\mathcal{S})$, we define an approximation $Q_{U|S,n}^{(P_S)}$ of $Q_{U|S}$ as follows. For each $s \in \mathcal{S}$ with $P_S(s) > 0$, let $Q_{U|S,n}^{(P_S)}(\cdot|s)$ be a type in $\mathcal{P}_{nP_S(s)}(\mathcal{U})$ whose probabilities are $\frac{1}{nP_S(s)}$-close to $Q_{U|S}$ in terms of L_∞ norm, and such that $Q_{U|S,n}^{(P_S)}(u|s) = 0$ wherever $Q_{U|S}(u|s) = 0$. If $P_S(s) = 0$ then $Q_{U|S,n}^{(P_S)}(\cdot|s)$ is arbitrary (e.g. uniform). Assuming without loss of generality that $\pi(s) > 0$ for all $s \in \mathcal{S}$, we have from (8) that $\min_s nP_S(s)$ grows linearly in n for all $P_S \in \tilde{\mathcal{P}}_n$. Thus,

$$\Big|Q_{U|S}(u|s) - Q_{U|S,n}^{(P_S)}(u|s)\Big| = O\Big(\frac{1}{n}\Big) \tag{10}$$

uniformly in $P_S \in \tilde{\mathcal{P}}_n$ and (s,u).

We will make use of the following joint distributions:

$$P_{SUY}^{(P_S)}(s,u,y) \triangleq P_S(s) Q_{U|S}(u|s) W(y|\phi(u,s),s) \tag{11}$$

$$P_{SUY,n}^{(P_S)}(s,u,y) \triangleq P_S(s) Q_{U|S,n}^{(P_S)}(u|s) W(y|\phi(u,s),s). \tag{12}$$

Using (10), we immediately obtain that

$$\Big|P_{SUY}^{(P_S)}(s,u,y) - P_{SUY,n}^{(P_S)}(s,u,y)\Big| = O\Big(\frac{1}{n}\Big) \tag{13}$$

uniformly in $P_S \in \tilde{\mathcal{P}}_n$ and (s,u,y).

D. A Taylor Expansion of the Mutual Information

Let $I^{(P_S)}(U;S)$ and $I^{(P_S)}(U;Y)$ denote mutual informations under the joint distribution $P_{USY}^{(P_S)}$ in (11), and define

$$I(P_S) \triangleq I^{(P_S)}(U;Y) - I^{(P_S)}(U;S). \tag{14}$$

We observe from (1) that $C = I(\pi)$. The following Taylor expansion (about $P_S = \pi$) is proved in [7]:

$$I(P_S) = \tilde{I}(P_S) + \Delta(P_S), \tag{15}$$

where

$$\tilde{I}(P_S) \triangleq \sum_s P_S(s) \sum_u Q_{U|S}(u|s) \times \left(\sum_y W(y|\phi(u,s),s) \log \frac{P_{Y|U}^{(\pi)}(y|u)}{P_Y^{(\pi)}(y)} - \log \frac{Q_{U|S}(u|s)}{P_U^{(\pi)}(u)} \right), \tag{16}$$

and

$$\max_{P_S \in \tilde{\mathcal{P}}_n} |\Delta(P_S)| \le \frac{K_1 \log n}{n} \tag{17}$$

for some constant K_1.

III. Proof of Theorem 1

As stated above, it suffices to prove Theorem 2. Thus, we assume that the state type P_S is known at the decoder.

1) Random-Coding Parameters: The parameters are the auxiliary alphabet $\mathcal{U}$, input distribution $Q_{U|S}$, function $\phi : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$, and number of auxiliary codewords $L^{(P_S)}$ for each state type $P_S \in \mathcal{P}_n(\mathcal{S})$. We assume that $\mathcal{U}$, $Q_{U|S}$ and ϕ are capacity-achieving.

2) Codebook Generation: For each state type $P_S \in \mathcal{P}_n(\mathcal{S})$ and each message m, we randomly generate an auxiliary codebook $\{\boldsymbol{U}^{(P_S)}(m,l)\}_{l=1}^{L^{(P_S)}}$, where each codeword is drawn independently according to the uniform distribution on the type class $T^n(P_{U,n}^{(P_S)})$ (see (12)). Each auxiliary codebook is revealed to the encoder and decoder.

3) Encoding and Decoding: Given the state sequence $\boldsymbol{S} \in T^n(P_S)$ and message m, the encoder sends

$$\phi^n(\boldsymbol{U}, \boldsymbol{S}) \triangleq \big(\phi(U_1,S_1), \cdots, \phi(U_n,S_n)\big), \tag{18}$$

where $\boldsymbol{U}$ is an auxiliary codeword $\boldsymbol{U}^{(P_S)}(m,l)$ with l chosen such that $(\boldsymbol{S},\boldsymbol{U}) \in T^n(P_{SU,n}^{(P_S)})$, with an error declared if no such auxiliary codeword exists. Given $\boldsymbol{y}$ and the state type P_S, the decoder estimates m according to the pair $(\tilde{m},\tilde{l})$ whose corresponding sequence $\boldsymbol{U}^{(P_S)}(\tilde{m},\tilde{l})$ maximizes

$$i_n^{(P_S)}(\boldsymbol{u},\boldsymbol{y}) \triangleq \sum_{i=1}^n i^{(P_S)}(u_i,y_i), \tag{19}$$

where

$$i^{(P_S)}(u_i,y_i) \triangleq \log \frac{P_{Y|U}^{(P_S)}(y|u)}{P_Y^{(P_S)}(y)} \tag{20}$$

with $P_{SUY}^{(P_S)}$ defined in (11). It should be noted that $P_{SUY}^{(\pi)}$ coincides with the distribution in (2), and hence $i^{(\pi)}(u,y)$ coincides with (4).

We consider the events

$$\mathcal{E}_1 \triangleq \Big\{\text{No } l \text{ yields } (\boldsymbol{S}, \boldsymbol{U}^{(P_S)}(m,l)) \in T^n(P_{SU,n}^{(P_S)})\Big\} \tag{21}$$

$$\mathcal{E}_2 \triangleq \Big\{\text{Decoder chooses a message } \tilde{m} \ne m\Big\}. \tag{22}$$

It follows from these definitions and (9) that the overall random-coding error probability $\overline{p}_e$ satisfies

$$\overline{p}_e \le \sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}\big[\hat{P}_{\boldsymbol{S}} = P_S\big]\Big(\mathbb{P}\big[\mathcal{E}_1 \,|\, \hat{P}_{\boldsymbol{S}} = P_S\big] + \mathbb{P}\big[\mathcal{E}_2 \,|\, \hat{P}_{\boldsymbol{S}} = P_S, \mathcal{E}_1^c\big]\Big) + O\Big(\frac{1}{n^2}\Big). \tag{23}$$

4) Analysis of $\mathcal{E}_1$: We study the probability of $\mathcal{E}_1$ conditioned on $\boldsymbol{S}$ having a given type $P_S \in \tilde{\mathcal{P}}_n$. Combining (13) with a standard property of types [12, Eq. (18)], each of the auxiliary codewords induces the joint type $P_{SU,n}^{(P_S)}$ with probability at least $p_0(n)^{-1}e^{-nI^{(P_S)}(U;S)}$, where $I^{(P_S)}(U;S)$ is defined in Section II-D, and $p_0(n)$ is polynomial in n. Since the codewords are independent, we have

$$\mathbb{P}\big[\mathcal{E}_1 \,|\, \hat{P}_{\boldsymbol{S}} = P_S\big] \le \big(1 - p_0(n)^{-1}e^{-nI^{(P_S)}(U;S)}\big)^{L^{(P_S)}} \tag{24}$$

$$\le \exp\Big(- p_0(n)^{-1}e^{-n\left(I^{(P_S)}(U;S) - R_L^{(P_S)}\right)}\Big), \tag{25}$$

where (25) follows using $1-\alpha \le e^{-\alpha}$ and defining

$$R_L^{(P_S)} \triangleq \frac{1}{n}\log L^{(P_S)}. \tag{26}$$

Choosing

$$R_L^{(P_S)} = I^{(P_S)}(U;S) + K_2\frac{\log n}{n} \tag{27}$$

with K_2 equal to one plus the degree of the polynomial $p_0(n)$, we obtain from (25) that

$$\mathbb{P}\big[\mathcal{E}_1 \,|\, P_S\big] \le e^{-\psi n} \tag{28}$$

for some $\psi > 0$ and sufficiently large n.

5) Analysis of $\mathcal{E}_2$: We study the probability of $\mathcal{E}_2$ conditioned on $\boldsymbol{S}$ having a given type $P_S \in \tilde{\mathcal{P}}_n$, and also conditioned on $\mathcal{E}_1^c$. By symmetry, all $(\boldsymbol{s},\boldsymbol{u}) \in T^n(P_{SU,n}^{(P_S)})$ are equally likely, and hence the conditional distribution given $\hat{P}_{\boldsymbol{S}} = P_S$ and $\mathcal{E}_1^c$ of the state sequence $\boldsymbol{S}$, auxiliary codeword $\boldsymbol{U}$, and received sequence $\boldsymbol{Y}$ is given by

$$(\boldsymbol{S},\boldsymbol{U},\boldsymbol{Y}) \sim P_{\boldsymbol{SU}}^{(P_S)}(\boldsymbol{s},\boldsymbol{u})W^n(\boldsymbol{y}|\phi^n(\boldsymbol{u},\boldsymbol{s}),\boldsymbol{s}), \tag{29}$$

where $P_{\boldsymbol{SU}}^{(P_S)}$ is uniform on the type class:

$$P_{\boldsymbol{SU}}^{(P_S)}(\boldsymbol{s},\boldsymbol{u}) \triangleq \frac{1}{\big|T^n(P_{SU,n}^{(P_S)})\big|}\mathbb{1}\Big\{(\boldsymbol{s},\boldsymbol{u}) \in T^n(P_{SU,n}^{(P_S)})\Big\}. \tag{30}$$

Let $P_{\boldsymbol{Y}}^{(P_S)}(\boldsymbol{y}) \triangleq \sum_{\boldsymbol{u},\boldsymbol{s}} P_{\boldsymbol{US}}^{(P_S)}(\boldsymbol{u},\boldsymbol{s})W^n(\boldsymbol{y}|\phi^n(\boldsymbol{u},\boldsymbol{s}),\boldsymbol{s})$ be the corresponding output distribution. Using a standard change of measure from constant-composition to i.i.d. (e.g. see [9, Ch. 2]), we can easily show that

$$P_{\boldsymbol{Y}}^{(P_S)}(\boldsymbol{y}) \le p_1(n)\prod_{i=1}^n P_Y^{(P_S)}(y_i), \tag{31}$$

where $p_1(n)$ is polynomial in n.

Recall that the decoder maximizes $i_n^{(P_S)}$ given in (19). Using a well-known threshold-based non-asymptotic bound [2], we have for any $\gamma^{(P_S)}$ that

$$\mathbb{P}\big[\mathcal{E}_2 \,|\, \hat{P}_{\boldsymbol{S}} = P_S, \mathcal{E}_1^c\big] \le \mathbb{P}\Big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y}) \le \gamma^{(P_S)}\Big]$$

$$+ ML^{(P_S)}\mathbb{P}\Big[i_n^{(P_S)}(\overline{\boldsymbol{U}},\boldsymbol{Y}) > \gamma^{(P_S)}\Big], \tag{32}$$

where $\overline{\boldsymbol{U}} \sim P_{\boldsymbol{U}}^{(P_S)}$ independently of $(\boldsymbol{S},\boldsymbol{U},\boldsymbol{Y})$. Using the change of measure given in (31), we can apply standard steps (e.g. see [3]) to upper bound the second term in (32) by $p_2(n)ML^{(P_S)}e^{-\gamma^{(P_S)}}$, where $p_2(n)$ is polynomial in n. We can ensure that this term is $O\big(\frac{1}{n}\big)$ by choosing $\gamma^{(P_S)} = \log ML^{(P_S)} + K_3\log n$, where K_3 is one higher than the degree of $p_2(n)$. Under this choice, and defining $K_4 \triangleq K_2 + K_3$, we obtain from (27) and (32) that

$$\mathbb{P}\big[\mathcal{E}_2 \,|\, \hat{P}_{\boldsymbol{S}} = P_S\big] \le \mathbb{P}\Big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y}) \le \log M$$

$$+ nI^{(P_S)}(U;S) + K_4\log n\Big] + O\Big(\frac{1}{n}\Big). \tag{33}$$

6) Application of the Berry-Esseen Theorem: Combining (28) and (33), we have for all $P_S \in \tilde{\mathcal{P}}_n$ that

$$\mathbb{P}\big[\mathcal{E}_1 \cup \mathcal{E}_2 \,|\, \hat{P}_{\boldsymbol{S}} = P_S\big] \le \mathbb{P}\Big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y}) \le \log M$$

$$+ nI^{(P_S)}(U;S) + K_4\log n\Big] + O\Big(\frac{1}{n}\Big). \tag{34}$$

In order to apply the Berry-Esseen theorem to the right-hand side of (34), we first compute the mean and variance of $i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y})$, defined according to (19) and (29). The required third moment can easily be uniformly bounded in terms of the alphabet sizes [13, Appendix D]. We will use the fact that, by the symmetry of the constant-composition distribution in (30), the statistics of $i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y})$ are unchanged upon conditioning on $(\boldsymbol{S},\boldsymbol{U}) = (\boldsymbol{s},\boldsymbol{u})$ for some $(\boldsymbol{s},\boldsymbol{u}) \in T^n(P_{SU,n}^{(P_S)})$. Using the joint distribution $P_{SUY,n}^{(P_S)}$ defined in (12), it follows that

$$\mathbb{E}\big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y})\big] = n\sum_{u,y} P_{UY,n}^{(P_S)}(u,y)i^{(P_S)}(u,y) \tag{35}$$

$$= nI^{(P_S)}(U;Y) + O(1), \tag{36}$$

where (35) follows by expanding the expectation as a sum from 1 to n, and (36) follows from (13) and the definitions of $i^{(P_S)}(u,y)$ and $I^{(P_S)}(U;Y)$. A similar argument yields

$$\mathrm{Var}\big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y})\big] = n\mathbb{E}\Big[\mathrm{Var}\big[i^{(P_S)}(U,Y) \,|\, U,S\big]\Big] + O(1) \tag{37}$$

$$\triangleq nV(P_S) + O(1). \tag{38}$$

It should be noted that $V(P_S)$ is bounded away for zero for $P_S \in \tilde{\mathcal{P}}_n$ and sufficiently large n, since $V(\pi) > 0$ by assumption in Theorem 1. Furthermore, the $O(1)$ terms in (36) and (38) are uniform in $P_S \in \tilde{\mathcal{P}}_n$.

Using the definition of $I(P_S)$ in (14), we choose

$$\log M = nI(\pi) - K_4\log n - \beta_n, \tag{39}$$

where β_n will be specified later, and will behave as $O(\sqrt{n})$. Combining (34), (36), (38) and (39), we have

$$\mathbb{P}\big[\mathcal{E}_1 \cup \mathcal{E}_2 \,|\, \hat{P}_{\boldsymbol{S}} = P_S\big]$$

$$\le \mathbb{P}\Big[i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y}) \le nI(\pi) + nI^{(P_S)}(U;S) - \beta_n\Big] + O\Big(\frac{1}{n}\Big). \tag{40}$$

$$\le \mathsf{Q}\bigg(\frac{\beta_n + nI(P_S) - nI(\pi) + K_5}{\sqrt{nV(P_S) + K_6}}\bigg) + O\bigg(\frac{1}{\sqrt{n}}\bigg) \tag{41}$$

where (41) follows by conditioning on $(\boldsymbol{S},\boldsymbol{U}) = (\boldsymbol{s},\boldsymbol{u})$ for some $(\boldsymbol{s},\boldsymbol{u}) \in T^n(P_{SU,n}^{(P_S)})$ (recall that this does not change the statistics of $i_n^{(P_S)}(\boldsymbol{U},\boldsymbol{Y})$), applying the Berry-Esseen theorem for independent and non-identically distributed variables [14, Sec. XVI.5], and introducing the constants K_5 and K_6 to represent the uniform $O(1)$ terms in (36) and (38).

7) Averaging Over the State Type: Substituting (41) into (23), we have

$$\overline{p}_e \le \sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\bigg(\frac{\beta + nI(P_S) - nI(\pi)}{\sqrt{nV(P_S)}}\bigg) + O\bigg(\frac{1}{\sqrt{n}}\bigg), \quad (42)$$

where we have factored the constants K_5 and K_6 into the remainder term using standard Taylor expansions along with the assumption $\beta_n = O(\sqrt{n})$; see [7] for details. Analogously to [8, Lemmas 17-18], we simplify (42) using two lemmas.

Lemma 1. *For any $\beta_n = O(\sqrt{n})$, we have*

$$\sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\Bigg(\frac{\beta_n + nI(P_S) - nI(\pi)}{\sqrt{nV(P_S)}}\Bigg) \le \sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\Bigg(\frac{\beta_n + nI(P_S) - nI(\pi)}{\sqrt{nV(\pi)}}\Bigg) + O\bigg(\frac{\log n}{\sqrt{n}}\bigg) \quad (43)$$

Proof: This follows using standard Taylor expansions along with the definition of $\tilde{\mathcal{P}}_n$ in (8) and the fact that $V(P_S)$ is continuously differentiable at $P_S = \pi$; see [7]. ■

Lemma 2. *For any β_n, we have*

$$\sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\bigg(\frac{\beta_n + nI(P_S) - nI(\pi)}{\sqrt{nV(\pi)}}\bigg) \le \mathsf{Q}\Bigg(\frac{\beta_n}{\sqrt{nV}}\Bigg) + O\bigg(\frac{\log n}{\sqrt{n}}\bigg), \quad (44)$$

where V is defined in (6).

Proof: Using the expansion of $I(P_S)$ in terms of $\tilde{I}(P_S)$ and $\Delta(P_S)$ given in (15), along with the property given in (17), we can easily show that the left-hand side of (44) is upper bounded by

$$\sum_{P_S \in \tilde{\mathcal{P}}_n} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\Bigg(\frac{\beta_n - nI(\pi) + n\tilde{I}(P_S)}{\sqrt{nV(\pi)}}\Bigg) + O\bigg(\frac{\log n}{\sqrt{n}}\bigg). \quad (45)$$

Since $\tilde{I}(P_S)$ is written in the form $\sum_s P_S(s)\psi(s)$, a trivial generalization of [8, Lemma 18] gives

$$\sum_{P_S} \mathbb{P}[\hat{P}_{\boldsymbol{S}} = P_S]\,\mathsf{Q}\Bigg(\frac{\beta_n + n\tilde{I}(P_S) - n\tilde{I}(\pi)}{\sqrt{nV(\pi)}}\Bigg) = \mathsf{Q}\Bigg(\frac{\beta_n}{\sqrt{n\big(V(\pi) + V^*(\pi)\big)}}\Bigg) + O\bigg(\frac{1}{\sqrt{n}}\bigg), \quad (46)$$

where $V^*(\pi) \triangleq \mathrm{Var}_\pi[\psi(S)]$. Using (16), we see that $\psi(S) = \mathbb{E}[i^{(\pi)}(U,Y) - i^{(\pi)}(U,S)\,|\,S]$, and it follows that $V(\pi) + V^*(\pi)$ is equal to V, defined in (6). The proof is concluded by expanding the summation in (45) to be over all types, and substituting (46). ■

Using (42) along with Lemmas 1 and 2, we have

$$\overline{p}_e \le \mathsf{Q}\bigg(\frac{\beta_n}{\sqrt{nV}}\bigg) + O\bigg(\frac{\log n}{\sqrt{n}}\bigg). \quad (47)$$

Setting $\overline{p}_e = \epsilon$ and solving for β_n, we obtain

$$\beta_n = \sqrt{nV}\mathsf{Q}^{-1}(\epsilon) + O(\log n). \quad (48)$$

Consistent with (42) and Lemma 1, we have $\beta_n = O(\sqrt{n})$. Substituting (48) into (39) yields the desired result with V of the form given in (6).

By analyzing the Karush-Kuhn-Tucker (KKT) corresponding to the maximization in (1), it can be shown that the equality in (7) holds under any $Q_{U|S}$ which maximizes the objective for a given pair $(\mathcal{U}, \phi)$ [7]. Since the parameters are capacity-achieving by assumption, this completes the proof.

Acknowledgment

I would like to thank Vincent Tan for many helpful comments and suggestions.

References

[1] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3rd Prague Conf. on Inf. Theory*, 1962, pp. 689–723, English Translation: http://www.math.wustl.edu/~luthy/strassen.pdf.

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

[3] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[4] S. I. Gelfand and M. S. Pinsker, "Coding for channel with random parameters," *Prob. Inf. Transm.*, vol. 9, no. 1, pp. 19–31, 1980.

[5] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, "Non-asymptotic and second-order achievability bounds for coding with side-information," 2013, http://arxiv.org/abs/1301.6467.

[6] M. H. Yassaee, M. R. Aref, and A. Gohari, "A technique for deriving one-shot achievability results in network information theory," http://arxiv.org/abs/1303.0696.

[7] J. Scarlett, "On the dispersions of the Gel'fand-Pinsker channel and dirty paper coding," 2013, submitted to *IEEE Trans. Inf. Theory* [arxiv: http://arxiv.org/abs/1309.6200].

[8] M. Tomamichel and V. Y. F. Tan, "ϵ-capacities and second-order coding rates for channels with general state," [Online: http://arxiv.org/abs/1305.6789].

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[10] A. Somekh-Baruch and N. Merhav, "On the random coding error exponents of the single-user and the multiple-access Gel'fand-Pinsker channels," in *IEEE Int. Symp. Inf. Theory*, Chicago, IL, June 2004.

[11] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[12] ——, "Fixed composition arguments and lower bounds to error probability," http://web.mit.edu/gallager/www/notes/notes5.pdf.

[13] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," 2012, arXiv:1201.3901v2 [cs.IT].

[14] W. Feller, *An introduction to probability theory and its applications*, 2nd ed. John Wiley & Sons, 1971, vol. 2.

An Alternative Coding Theorem for Posterior Matching via Extrinsic Jensen–Shannon Divergence

Tara Javidi
Electrical and Computer Engineering
University of California San Diego
Email: tjavidi@ucsd.edu

Michèle Wigger
Communications and Electronics
Telecom ParisTech
Email: michele.wigger@telecom-paristech.fr

Mohammad Naghshvar
Qualcomm Technologies, Inc.
Corporate R&D
Email: mnaghshvar@qti.qualcomm.com

***Abstract*—This paper considers the problem of coding over a discrete memoryless channel (DMC) with noiseless feedback. The paper provides a stochastic control view of a variable-length version of the posterior matching scheme which is analyzed via a recently proposed symmetrized divergence, termed Extrinsic Jensen–Shannon (EJS) divergence. In particular, under the variable-length posterior matching scheme, the EJS divergence can be lower bounded by the Shannon capacity of the DMC, which can be used for a relatively simple proof that the variable-length posterior matching scheme achieves capacity.**

I. Introduction

In [1], [2], see also [3], a sequential, one-phase scheme for transmission over a BSC with noiseless feedback was proposed. This scheme is briefly explained next. Each message is represented as a subinterval of size $\frac{1}{M}$ of the unit interval. After each transmission and given the channel output, the posterior probability of all subintervals are updated. In the next time slot, the transmitter sends 0 if the true message's corresponding subinterval is below the current median, or 1 if it is above. If the current median lies within the true message's subinterval, then the transmitter sends 0 with probability equal to the fraction of the interval above the median and 1 otherwise. As the rounds of transmission proceed, the posterior probability of the true message's subinterval most likely grows larger than $\frac{1}{2}$, which pushes the median within the message's subinterval and thus leads to a randomized encoding. Although this simple one-phase scheme was believed to achieve the capacity of a BSC, a rigorous proof remained illusive prior to the work by Shayevitz and Feder [3]. They generalized the described scheme to arbitrary DMCs (satisfying some mild conditions) and proved that their general scheme, named *posterior matching scheme*, achieves capacity [3]. Recently, Li and El Gamal proposed a related scheme [4] with a greatly improved error-exponent, i.e. with exponentially smaller probability of error than the posterior matching scheme.

In [5], we introduced the *Extrinsic Jensen–Shannon (EJS) divergence* as a tool to analyze error exponents and achievable rates for variable-length schemes. In this paper we show that this tool allows for a relatively simple proof that a variable-length version of the posterior matching scheme achieves the capacity of DMCs.

We finish this section with some notation.

Notation: Let $[x]^+ = \max\{x, 0\}$. The i^{th} element of vector $\boldsymbol{v}$ is denoted by v_i. The notations A^t and a^t stand for the tuples $[A_0, \ldots, A_t]$ and $[a_0, \ldots, a_t]$, respectively, for positive integer t. For any set $\mathcal{S}$, $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$ and $\mathcal{S}^t$ its t-fold Cartesian product. All logarithms are in base 2. The entropy function on a vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_M] \in [0,1]^M$ is defined as $H(\boldsymbol{\rho}) := \sum_{i=1}^M \rho_i \log \frac{1}{\rho_i}$, with the convention that $0 \log \frac{1}{0} = 0$. We denote the conditional probability $P(Y|X = x)$ by P_x.

II. Preliminaries: Symmetric Divergences

We first recall that the *Kullback–Leibler (KL) divergence* between two probability distributions P_Y and P'_Y over a finite set $\mathcal{Y}$ is defined as $D(P_Y \| P'_Y) := \sum_{y \in \mathcal{Y}} P_Y(y) \log \frac{P_Y(y)}{P'_Y(y)}$ with the convention $0 \log \frac{0}{a} = 0$ and $b \log \frac{b}{0} = \infty$ for $a, b \in [0,1]$ with $b \neq 0$. The KL divergence is *not* symmetric, i.e., in general $D(P_Y \| P'_Y) \neq D(P'_Y \| P_Y)$.

The *J divergence* [6] and *L divergence* [7] symmetrize the KL divergence:

$$J(P_1, P_2) := D(P_1 \| P_2) + D(P_2 \| P_1), \tag{1}$$

$$L(P_1, P_2) := D\Big(P_1 \| \frac{1}{2}P_1 + \frac{1}{2}P_2\Big) + D\Big(P_2 \| \frac{1}{2}P_1 + \frac{1}{2}P_2\Big). \tag{2}$$

The *Jensen–Shannon (JS) divergence* [7], [8] is defined similarly to the L divergence but for general $M \geq 2$ probability distributions. Given M probability distributions $P_1, P_2 \ldots, P_M$ over a set $\mathcal{Y}$ and a vector of a priori weights $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots, \rho_M]$, where $\boldsymbol{\rho} \in [0,1]^M$ and $\sum_{i=1}^M \rho_i = 1$, the JS divergence is defined as [7], [8]:

$$\begin{aligned} JS(\boldsymbol{\rho}; P_1, \ldots, P_M) &:= \sum_{i=1}^M \rho_i D\Big(P_i \| \sum_{j=1}^M \rho_j P_j\Big) \\ &= I(\theta; Y) \end{aligned} \tag{3}$$

where θ is a random variable that takes values in $\{1, 2, \ldots, M\}$ and has probability mass function $\boldsymbol{\rho}$ and $Y \sim P_\theta$ (which implies that $\Pr(Y = y) = \sum_{i=1}^M \rho_i P_i(y)$).

Similarly, one can consider the *Extrinsic Jensen–Shannon (EJS) divergence* [5] which extends the J divergence to general $M \geq 2$ probability distributions. For distributions $P_1, \ldots, P_M$ and an M-dimensional weight vector $\boldsymbol{\rho}$,

$$EJS(\boldsymbol{\rho}; P_1, \ldots, P_M) := \sum_{i=1}^M \rho_i D\Big(P_i \| \sum_{j \neq i} \frac{\rho_j}{1-\rho_i} P_j\Big), \tag{4a}$$

when $\rho_i < 1$ for all $i \in \{1, \ldots, M\}$, and

$$EJS(\boldsymbol{\rho}; P_1, \ldots, P_M) := \max_{j \neq i} D(P_i \| P_j) \tag{4b}$$

when $\rho_i = 1$ for some $i \in \{1, \ldots, M\}$.

III. Coding over DMC with Noiseless Feedback

A. The Problem Setup

Consider the problem of variable-length coding over a discrete memoryless channel (DMC) with noiseless feedback as depicted in Fig. 1. The DMC is described by finite input and output sets $\mathcal{X}$ and $\mathcal{Y}$, and a collection of conditional probabilities $P(Y|X)$. To simplify notation, and without loss of generality, we assume that

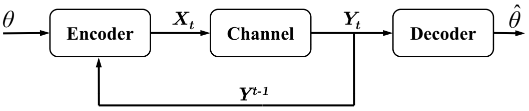


Fig. 1. A noisy memoryless channel with a noiseless causal feedback link.

$$\mathcal{X} = \{0, 1, \ldots, |\mathcal{X}| - 1\} \tag{5}$$

and

$$\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}. \tag{6}$$

Let τ denote the total transmission time (or equivalently the total length of the code). In this paper, our focus is on variable-length coding, i.e., the case where τ is a random stopping time decided at the receiver as a function of the observed channel outputs. (A specific stopping rule is described later in this section.) Thanks to the noiseless feedback, the transmitter is also informed of the channel outputs and the stopping time.

The transmitter wishes to communicate a message θ to the receiver, where the message is uniformly distributed over a message set

$$\Omega := \{1, 2, \ldots, M\}. \tag{7}$$

To this end, it produces channel inputs X_t for $t = 0, 1, \ldots, \tau - 1$, which it can compute as a function of the message θ and (thanks to the feedback) also of the past channel outputs $Y^{t-1} := [Y_0, Y_1, \ldots, Y_{t-1}]$:

$$X_t = e_t(\theta, Y^{t-1}), \quad t = 0, 1, \ldots, \tau - 1, \tag{8}$$

for some encoding function $e_t \colon \Omega \times \mathcal{Y}^t \to \mathcal{X}$.

To describe the encoding process, we shall also use the functions $\{\gamma_{y^{t-1}}\}$ for $y^{t-1} \in \mathcal{Y}^t$ and $t \in \{0, 1, \ldots, \tau - 1\}$ where

$$\gamma_{y^{t-1}} \colon \Omega \to \mathcal{X} \tag{9a}$$
$$i \mapsto e_t(i, y^{t-1}). \tag{9b}$$

Where it is clear from the context and to simplify notation, we omit the subscript y^{t-1} and simply write γ.

We will particularly be interested in *randomized* encoding rules. In this case the encoding is described by the *random* encoding functions $\{\Gamma_{y^{t-1}}\}$ whose realizations $\gamma_{y^{t-1}}$ are of the form in (9). Again, for notational convenience we omit the subscript y^{t-1} where it is clear from the context.

After observing the τ channel outputs $Y_0, Y_1, \ldots, Y_{\tau-1}$, the receiver performs optimum maximum-likelihood decoding and produces as its guess the message with the highest posterior:

$$\hat{\theta} = \arg\max_{i \in \Omega} \rho_i(\tau), \tag{10}$$

where for each positive t and each $i \in \Omega$:

$$\rho_i(t) := \Pr(\theta = i | Y^{t-1}). \tag{11}$$

The probability of error is

$$\text{Pe} := \Pr(\hat{\theta} \neq \theta). \tag{12}$$

For a fixed DMC and for a given $\epsilon > 0$, the goal is to find an encoding rule as in (8) and a stopping rule τ such that combined with the decoding rule in (10) the probability of error satisfies $\text{Pe} \leq \epsilon$ and the expected number of channel uses $\mathbb{E}[\tau]$ is minimized.

Throughout the paper we assume that the receiver applies the following possibly suboptimal stopping rule

$$\tau := \min\{t : \max_{i \in \Omega} \rho_i(t) \geq 1 - \epsilon\}, \tag{13}$$

where $\epsilon > 0$ is the desired probability of error.

The main interest in this paper is in achieving the *capacity* of DMCs with a variable-length scheme. The capacity is defined as follows. If for any small numbers $\delta > 0$, $0 \leq \epsilon < 1$ and all sufficiently large positive integers ℓ an encoding scheme γ (or Γ) can transmit one out of M_ℓ equiprobable messages so that with the ML decoder in (10) and an appropriate stopping rule τ,

$$\text{Pe} \leq \epsilon \tag{14a}$$
$$M_\ell \geq 2^{\ell(R-\delta)} \tag{14b}$$
$$\mathbb{E}[\tau] \leq \ell, \tag{14c}$$

for some positive real number R, then we say that the scheme achieves rate R. The capacity is the supremum over all achievable rates and is given by

$$C := \max_{P_X} I(X; Y), \tag{15}$$

as in the case of fixed-length coding.

B. Stochastic Control View

The problem of coding with noiseless feedback is a decentralized team problem with two agents (the encoder and the decoder) and non-classical information structure [9]. Appealing to [10], the problem can be interpreted as a special case of active hypothesis testing in which a (fictitious) Bayesian decision-maker is responsible to enhance his information about the correct message in a speedy manner by sequentially sampling from conditionally independent observations at the output of the channel (given the input). Here the (fictitious) decision maker has access to the channel output symbols causally (common observations) and is responsible to control the conditional distribution of the observations given the true message (private observation) by selecting encoding functions for the encoder which map the message θ to the input symbols

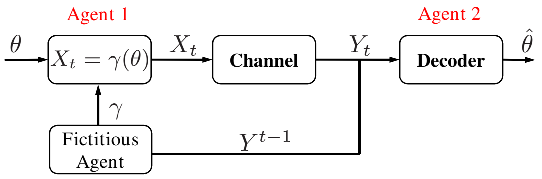


Fig. 2. Two-agent problem with common and private observations from the point of view of the fictitious agent.

of the channel. In other words, as also observed in [11], the problem can be viewed as a (centralized) partially observable Markov decision problem (POMDP) with (static) state space Ω and the observation space $\mathcal{Y}$. Let $\mathcal{E} := \{\gamma(\cdot) : \Omega \to \mathcal{X}\}$ be the set of all mappings from Ω to $\mathcal{X}$. The action space (for the fictitious agent) becomes $\mathcal{E} \cup \{T\}$ where T denotes the termination of the transmission phase, hence the realization of the stopping time τ.

Casting the problem as a POMDP allows for the structural characterization of the information state, also known as sufficient statistics: The decision maker's posteriors about the messages collectively,

$$\boldsymbol{\rho}(t) := [\rho_1(t), \rho_2(t), \ldots, \rho_M(t)], \tag{16}$$

form a sufficient statistics for our (fictitious) Bayesian decision maker. Furthermore, this (fictitious) decision maker's posterior at any time t coincides with the receiver's posterior and, thanks to the perfect feedback, is available to the transmitter. In other words, the selection of encoding and decoding rules as a function of this posterior incurs no loss of optimality [12].

We also note that the dynamics of the information state, i.e. the posterior, follows Bayes' rule. More specifically, given an encoding function γ at time t and an information state $\boldsymbol{\rho}$, the conditional distribution of the next channel output Y_t, given the past observation Y^{t-1}, is

$$P_{\boldsymbol{\rho}}(y) = \sum_{i=1}^{M} \rho_i P(Y = y | X = \gamma(i)).$$

Similarly, given also the output symbol $Y_t = y$, according to Bayes' rule, the posterior at time $t+1$ is:

$$\boldsymbol{\rho}(t+1) = \left[\frac{\rho_1 P_{\gamma(1)}(y)}{P_{\boldsymbol{\rho}}(y)}, \ldots, \frac{\rho_M P_{\gamma(M)}(y)}{P_{\boldsymbol{\rho}}(y)} \right].$$

This stochastic control view of the problem, suggests an achievability analysis which generalizes the approach of [11] beyond mutual information and is based on symmetric divergence associate with the belief state $\boldsymbol{\rho}$ and $\{P_x\}_{x\in\mathcal{X}}$. In the sections that follow, we utilize this approach with respect to the EJS divergence induced by the posterior matching. This allows us to provide a concise achievability analysis for variable-length posterior matching.

IV. Main Result

Consider the variable-length version of the posterior matching encoding in [3]:

At each time $t = 0, 1, \ldots, \tau - 1$, if $\theta = i$ and given the posterior vector $\boldsymbol{\rho}(t)$, the input $X(t)$ takes value in the set

$$\mathcal{X}_i(t) := \Big\{ x \in \mathcal{X} \colon \sum_{i'=1}^{i-1} \rho_{i'}(t) < \sum_{x' \leq x} \pi^\star_{x'} \\ \text{and } \sum_{x' < x} \pi^\star_{x'} \leq \sum_{i'=1}^{i} \rho_{i'}(t) \Big\};$$

where each value $x \in \mathcal{X}_i(t)$ is taken with probability

$$\Pr\big(X(t) = x | \theta = i, Y^{t-1} = y^{t-1}\big) \\ = \frac{\min\Big\{ \sum\limits_{i'=1}^{i} \rho_{i'}(t), \sum\limits_{x' \leq x} \pi^\star_{x'} \Big\} - \max\Big\{ \sum\limits_{i'=1}^{i-1} \rho_{i'}(t), \sum\limits_{x' < x} \pi^\star_{x'} \Big\}}{\rho_i(t)}.$$

We show that the described posterior matching encoding Γ^{PM} combined with the ML decoding in (10) and stopping rule (13) achieves capacity for all DMCs satisfying a mild condition. Let C_1 be the KL divergence between the two most distinguishable inputs of the DMC:

$$C_1 := \max_{x, x' \in \mathcal{X}} D(P(Y|X = x) \| P(Y|X = x')). \tag{18}$$

Theorem 1. *The described posterior matching encoding Γ^{PM} combined with the optimal ML decoder* (10) *and stopping rule* (13) *achieve the capacity of any DMC where C and C_1 are positive and finite.*[1]

Proof: For a fixed encoding rule γ and given a sequence of channel outputs y^{t-1} with corresponding posteriors $\boldsymbol{\rho}(t)$, we define $EJS(\boldsymbol{\rho}(t), \gamma)$ to be the EJS divergence between the conditional output distributions $P_{\gamma(1)}, \ldots, P_{\gamma(M)}$ with weight vector $\boldsymbol{\rho}(t)$:

$$EJS(\boldsymbol{\rho}(t), \gamma) := EJS\big(\boldsymbol{\rho}(t); P_{\gamma(1)}, \ldots, P_{\gamma(M)}\big). \tag{19}$$

For a randomized encoding function Γ, we use

$$EJS(\boldsymbol{\rho}(t), \Gamma) := \sum_{\gamma \in \mathcal{E}} \Pr(\Gamma = \gamma | Y^{t-1} = y^{t-1}) EJS(\boldsymbol{\rho}(t), \gamma)$$

where recall that $\mathcal{E}$ denotes the set of all possible encoding functions. Let $\tilde{\rho} := 1 - (1 + \max\{\log M, \log \frac{1}{\epsilon}\})^{-1}$.

Our proof is based on the following theorem from [5]:

Theorem 2 (Corollary 2 in [5])**.** *Consider a DMC with $C > 0$ and $C_1 < \infty$ and a variable-length encoding Γ combined with the ML decoding in* (10) *and the stopping rule* (13)*. If for any time $t < \tau$ and for any posterior vector $\boldsymbol{\rho}(t)$,*

$$EJS(\boldsymbol{\rho}(t), \Gamma) \geq C, \tag{20a}$$

then the scheme achieves the capacity C of the channel. Furthermore, if also,

$$EJS(\boldsymbol{\rho}(t), \Gamma) \geq \tilde{\rho} C_1 \qquad \text{if } \max_{i \in \Omega} \rho_i(t) \geq \tilde{\rho}, \tag{20b}$$

then it also achieves the channel's optimal reliability function.

[1] Notice that $C \leq C_1$ and $C_1 < \infty$ if, and only if, $P(Y = y|X = x)$ is positive for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Theorem 1 can thus be shown by proving that Condition (20a) is satisfied for the posterior matching encoding $\Gamma = \Gamma^{\mathrm{PM}}$:

$$EJS(\boldsymbol{\rho}(t), \Gamma^{\mathrm{PM}}) \geq C. \tag{21}$$

Fix a time instant t and $Y^{t-1} = y^{t-1}$. For ease of notation, in the following we drop the time index t for $\rho_i(t)$ and simply write ρ_i. Let

$$\lambda_\gamma := \Pr(\Gamma^{\mathrm{PM}} = \gamma | Y^{t-1} = y^{t-1}), \qquad \gamma \in \mathcal{E}.$$

Define for each $i \in \Omega$ and $x \in \mathcal{X}$:

$$\Lambda_{i,x} := \sum_{\gamma\colon \gamma(i)=x} \lambda_\gamma = \Pr(X = x | \theta = i, Y^{t-1} = y^{t-1}) \tag{22}$$

and

$$\hat{\rho}_{i,x} := \rho_i \Lambda_{i,x} = \Pr(X = x, \theta = i | Y^{t-1} = y^{t-1}). \tag{23}$$

For a fixed posterior distribution, the various messages are mapped into inputs of the channel independently of each other and hence, for $x, x' \in \mathcal{X}$ and $i, j \in \Omega$ where $i \neq j$

$$\sum_{\gamma\colon \substack{\gamma(i)=x \\ \gamma(j)=x'}} \lambda_\gamma = \Lambda_{i,x}\Lambda_{j,x'}. \tag{24}$$

Let $\pi^\star_0, \ldots, \pi^\star_{|\mathcal{X}|-1}$ denote the capacity-achieving input distribution, i.e., the maximizer of (15). Rearranging terms and using Jensen's inequality and the convexity of the KL divergence, we have

$$\begin{aligned}
&EJS(\boldsymbol{\rho}(t), \Gamma^{\mathrm{PM}}) \\
&= \sum_{\gamma\in\mathcal{E}} \lambda_\gamma \sum_{i=1}^{M} \rho_i D\Big(P_{\gamma(i)} \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} P_{\gamma(j)}\Big) \\
&= \sum_{i=1}^{M} \rho_i \sum_{x\in\mathcal{X}} \sum_{\gamma\colon \gamma(i)=x} \lambda_\gamma D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} P_{\gamma(j)}\Big) \\
&\geq \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \rho_i \Lambda_{i,x} D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} \sum_{\gamma\colon \gamma(i)=x} \frac{\lambda_\gamma}{\Lambda_{i,x}} P_{\gamma(j)}\Big) \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \sum_{j\neq i} \frac{\rho_j}{1-\rho_i} \sum_{x'\in\mathcal{X}} \sum_{\gamma\colon \substack{\gamma(i)=x \\ \gamma(j)=x'}} \frac{\lambda_\gamma}{\Lambda_{i,x}} P_{x'}\Big) \\
&\overset{(a)}{=} \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{j\neq i}\sum_{x'\in\mathcal{X}} \rho_j \Lambda_{j,x'} P_{x'}}{1-\rho_i}\Big) \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{x'\in\mathcal{X}}(\pi^\star_{x'} P_{x'} - \hat{\rho}_{i,x'} P_{x'})}{1-\rho_i}\Big), \\
&= \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} D\Big(P_x \Big\| \frac{\sum_{x'\in\mathcal{X}}(\pi^\star_{x'} P_{x'} - \hat{\rho}_{i,x'} P_{x'})}{1-\rho_i}\Big) \\
&\qquad + \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} \frac{\rho_i}{1-\rho_i} D\Big(P_x \Big\| \frac{\sum_{x'} \hat{\rho}_{i,x'} P_{x'}}{\rho_i}\Big) \\
&\qquad - \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \hat{\rho}_{i,x} \frac{\rho_i}{1-\rho_i} D\Big(P_x \Big\| \frac{\sum_{x'} \hat{\rho}_{i,x'} P_{x'}}{\rho_i}\Big) \\
&\geq \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \frac{\hat{\rho}_{i,x}}{1-\rho_i} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \pi^\star_{x'} P_{x'}\Big) \\
&\qquad - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} \sum_{x\in\mathcal{X}} \Lambda_{i,x} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \Lambda_{i,x'} P_{x'}\Big) \\
&\overset{(b)}{\geq} \sum_{i=1}^{M} \sum_{x\in\mathcal{X}} \frac{\hat{\rho}_{i,x}}{1-\rho_i} C - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} C \\
&= \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i} C - \sum_{i=1}^{M} \frac{\rho_i^2}{1-\rho_i} C \\
&= C
\end{aligned} \tag{25}$$

where (a) follows from (24); and where (b) follows from [13, Theorem 4.5.1] because $\hat{\rho}_{i,x} > 0$ only if $\pi^\star_x > 0$ and from the fact that $\sum_{x\in\mathcal{X}} \Lambda_{i,x} D\Big(P_x \Big\| \sum_{x'\in\mathcal{X}} \Lambda_{i,x'} P_{x'}\Big) = I(X;Y) \leq C$ when X denotes an input with probability mass function $\{\Lambda_{i,x}\}_{x\in\mathcal{X}}$ and Y the output produced by the channel. ■

FUTURE WORK

In future work, using large-deviation analysis, we plan to extend our EJS-divergence based proof technique to the original fixed-length posterior matching scheme.

REFERENCES

[1] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, July 1963.

[2] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problemy Peredachi Informatsii*, vol. 10, no. 3, pp. 51–61, 1974.

[3] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1186–1222, March 2011.

[4] C. T. Li and A. E. Gamal, "An efficient feedback coding scheme with low error probability for discrete memoryless channels," Nov. 2013, available on http://arxiv.org/pdf/1311.0100v2.

[5] M. Naghshvar, T. Javidi, and M. Wigger, "Extrinsic Jensen–Shannon divergence: Applications to variable-length coding," available on arXiv: 1307.0067.

[6] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. London. Ser. A.*, vol. 186, pp. 453–461, 1946.

[7] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145 –151, January 1991.

[8] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inform. Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[9] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.

[10] A. Mahajan, A. Nayyar, and D. Teneketzis, "Identifying tractable decentralized problems on the basis of information structures," in *Proceedings of the 46th Allerton conference on communication, control, and computing*, 2008, pp. 1440–1449.

[11] T. P. Coleman, "A stochastic control viewpoint on 'posterior matching'-style feedback communication schemes," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2009, pp. 1520–1524.

[12] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[13] R. G. Gallager, *Information theory and reliable communication*. John Wiley & Sons, Inc., New York, 1968.

Duality with Linear-Feedback Schemes for the Scalar Gaussian MAC and BC

Selma Belhadj Amor
Telecom ParisTech
Paris, France
belhadjamor@telecom-paristech.fr

Yossef Steinberg
Technion—Israel Inst. of Technology
Haifa, Israel
ysteinbe@ee.technion.ac.il

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

***Abstract*—We show that with perfect feedback and when restricting to linear-feedback schemes, the regions achieved over the two-user scalar Gaussian memoryless MAC and over the two-user scalar Gaussian memoryless BC coincide, if the MAC and the BC have equal channel coefficients and if the same (sum-)power constraint P is imposed on their inputs. Since the achievable region for the MAC is well known (it equals Ozarow's perfect-feedback capacity region under a sum-power constraint), we can characterize the region that is achievable over the scalar Gaussian BC with linear-feedback schemes.**

I. Introduction

Feedback is known to increase the capacity of multi-user channels such as the multi-access channel (MAC) and the broadcast channel (BC). But for most multi-user channels the exact capacity region is open even with perfect output-feedback. A notable exception is the two-user memoryless Gaussian MAC whose capacity was derived by Ozarow [1]. Ozarow's capacity-achieving scheme is a *linear-feedback scheme*, i.e., a scheme where at each time the transmitters send linear combinations of the past feedback signals and of code symbols that only depend on their own message. Kramer extended this scheme to $K \geq 3$ users [2]. Under symmetric power constraints P for all users, Kramer's scheme achieves the largest sum-rate among all *linear-feedback schemes* [3], and it achieves the sum-capacity when P is sufficiently large [2]. (It is yet unknown whether the scheme achieves the sum-capacity also when P is small.)

For the memoryless Gaussian BC the capacity region with perfect feedback is unknown even with two receivers. Achievable regions based on linear-feedback schemes have been proposed in [2], [4], [5], [6], [7], [8]. Non-linear feedback schemes have been proposed in [14], [9], [10]. The best known achievable regions are due to linear-feedback schemes.

The linear-feedback schemes in [5], [6], [7] are designed based on control-theoretic considerations. For some setups, e.g., uncorrelated noises of equal variance [7], these schemes achieve the same sum-rate over the Gaussian BC under power constraint P as Ozarow's scheme achieves over the Gaussian MAC under a *sum-power* constraint P. Thus, there is a duality in terms of sum-rate between the BC-schemes in [5], [6], [7] and Ozarow's MAC-scheme [1].

In this paper, we prove a duality between *arbitrary* linear-feedback schemes over the two-user scalar Gaussian MAC and BC, similar to the MIMO (nofeedback) MAC-BC duality in [11], [15]. Specifically, we show that the regions achieved by linear-feedback schemes over the two-user scalar Gaussian MAC under sum-power constraint P and over the two-user scalar Gaussian BC with uncorrelated noises under the same power constraint P coincide, if the scalar channel coefficients of the MAC and the BC are equal. Since the set of achievable regions over the Gaussian MAC using linear-feedback schemes is known—it equals Ozarow's achievable region under a sum-power constraint—our result allows to characterize the region that is achievable with linear-feedback schemes over the two-user scalar Gaussian BC with uncorrelated noises. We can also identify the optimal linear-feedback schemes over the scalar Gaussian BC and show that for equal noise-variances the control-theoretic schemes in [5], [6], [7] achieve the largest sum-rate among all linear-feedback schemes.

II. Gaussian MAC with Feedback

Consider the two-user memoryless scalar Gaussian MAC with perfect output-feedback in Figure 1. At each time $t \in \mathbb{N}$, if $x_{1,t}$ and $x_{2,t}$ denote the real symbols sent by Transmitters 1 and 2, the receiver observes the real channel output

$$Y_t = h_1 x_{1,t} + h_2 x_{2,t} + Z_t, \tag{1}$$

where h_1 and h_2 are constant nonzero channel coefficients and $\{Z_t\}$ is a sequence of independent and identically distributed (i.i.d.) zero-mean unit-variance[1] Gaussian random variables.

The goal of communication is that Transmitters 1 and 2 convey their independent messages M_1 and M_2 to the common receiver. The messages M_1 and M_2 are independent of the noises $\{Z_t\}$ and uniformly distributed over the sets $\mathcal{M}_1 \triangleq \{1, \ldots, \lfloor 2^{nR_1} \rfloor\}$ and $\mathcal{M}_2 \triangleq \{1, \ldots, \lfloor 2^{nR_2} \rfloor\}$, where R_1 and R_2 denote the rates of transmission and n the blocklength.

The two transmitters observe perfect feedback from the channel outputs. Thus, the time-t channel input at Transmitter $i \in \{1,2\}$ can depend on all previous channel outputs Y^{t-1} and its message M_i:

$$X_{i,t} = f_{i,t}^{(n)}(M_i, Y^{t-1}), \quad t \in \{1, \ldots, n\}, \tag{2}$$

[1] Assuming unit-variance entails no loss in generality because otherwise the receiver can simply scale its outputs appropriately.

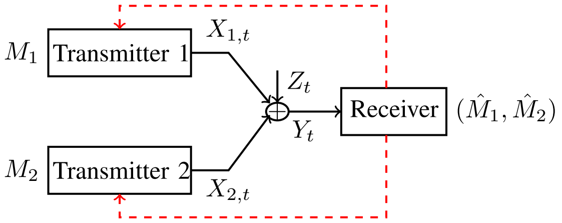


Fig. 1. Two-user Gaussian MAC with feedback.

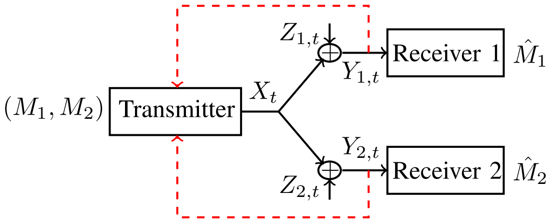


Fig. 2. Two-user Gaussian BC with feedback.

for some encoding function $f_{i,t}^{(n)}\colon \mathcal{M}_i \times \mathbb{R}^{t-1} \to \mathbb{R}$. The channel inputs $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ have to satisfy an expected average *sum-power constraint*

$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{E}[X_{1,t}^2]+\mathbf{E}[X_{2,t}^2]\right) \le P, \tag{3}$$

where the expectation is over the messages and the realizations of the channel. The receiver produces a guess

$$(\hat{M}_1^{(n)}, \hat{M}_2^{(n)}) = \Phi^{(n)}(Y^n)$$

by means of a decoding function $\Phi^{(n)}\colon \mathbb{R}^n \to \mathcal{M}_1 \times \mathcal{M}_2$.

The average probability of error is

$$P_{\text{e,MAC}}^{(n)} \triangleq \Pr\left[(\hat{M}_1, \hat{M}_2) \neq (M_1, M_2)\right]. \tag{4}$$

We say that a rate-pair (R_1, R_2) is achievable over the Gaussian MAC with feedback under a sum-power constraint P, if there exists a sequence of encoding and decoding functions $\left\{\{f_{1,t}^{(n)}\}_{t=1}^{n}, \{f_{2,t}^{(n)}\}_{t=1}^{n}, \Phi^{(n)}\right\}_{n=1}^{\infty}$ as described above, satisfying (3) and such that the average probability of error $P_{\text{e,MAC}}^{(n)}$ tends to zero as the blocklength n tends to infinity. The closure of the union of all achievable regions is called *capacity region*.

In the present paper we focus on *linear-feedback schemes for the MAC*, where the channel inputs can be written as

$$\mathbf{X}_i = \mathbf{V}_i + \mathsf{C}_i\mathbf{Y}, \qquad i \in \{1,2\}, \tag{5}$$

where $\mathbf{Y} \triangleq \left(Y_1, \ldots, Y_n\right)^{\mathsf{T}}$ is the channel output vector, C_1 and C_2 are n-by-n strictly lower-triangular matrices and $\mathbf{V}_i$ is an n-dimensional information-carrying vector $\mathbf{V}_i = \varphi_i(M_i)$. The mapping $\varphi_i\colon \mathcal{M}_i \to \mathbb{R}^n$ as well as the decoder mapping $\Phi^{(n)}$ can be arbitrary (also non-linear). The strict-lower-triangularity of the matrices C_1 and C_2 ensures that the feedback is used in a strictly causal way. (Notice that any nofeedback scheme is of the form in (5) with $\mathsf{C}_1 = \mathsf{C}_2 = 0$.)

The set of all rate-pairs achieved by linear-feedback schemes is called *linear-feedback capacity region* and is denoted $\mathcal{C}_{\text{MAC}}^{\text{linfb}}(h_1, h_2; P)$. The largest sum-rate achieved by a linear-feedback scheme is called *linear-feedback sum-capacity* and is denoted $C_{\text{MAC},\Sigma}^{\text{linfb}}(h_1, h_2, P)$. Since Ozarow's capacity-achieving scheme [1] is a linear-feedback scheme,[2] the general feedback capacity region and the linear-feedback capacity region coincide. They are both given by

$$\mathcal{C}_{\text{MAC}}^{\text{linfb}}(h_1, h_2; P) = \bigcup_{\substack{P_1, P_2 \ge 0:\\ P_1+P_2=P}} \; \bigcup_{\rho\in[0,1]} \mathcal{R}_{\text{Oz}}^{\rho}(h_1, h_2; P_1, P_2) \tag{6}$$

where $\mathcal{R}_{\text{Oz}}^{\rho}(h_1, h_2; P_1, P_2)$ is the set of all nonnegative rate-pairs (R_1, R_2) satisfying

$$R_i \le \frac{1}{2}\log\left(1 + h_i^2 P_i(1-\rho^2)\right), \quad i \in \{1,2\}, \tag{7a}$$

$$R_1 + R_2 \le \frac{1}{2}\log\left(1 + h_1^2P_1 + h_2^2P_2 + 2\sqrt{h_1^2h_2^2P_1P_2}\rho\right). \tag{7b}$$

The linear-feedback sum-capacity is given in Equation (8) on top of the next page, where $\rho^\star(h_1, h_2; P_1, P_2)$ is the unique solution in $[0,1]$ to the following quartic equation in ρ

$$\begin{aligned}(1 + h_1^2P_1 + h_2^2P_2 + 2\sqrt{h_1^2h_2^2P_1P_2}\rho) &= \\ (1 + h_1^2P_1(1-\rho^2))&(1 + h_2^2P_2(1-\rho^2)).\end{aligned} \tag{9}$$

III. Gaussian BC with Feedback

Consider the two-user scalar Gaussian BC with perfect output-feedback in Figure 2. We now have one transmitter and two receivers. At each time $t \in \mathbb{N}$, if $x_t \in \mathbb{R}$ denotes the transmitter's channel input, Receiver $i \in \{1,2\}$ observes the real channel output

$$Y_{i,t} = h_i x_t + Z_{i,t}, \tag{10}$$

where h_1 and h_2 are constant non-zero channel coefficients and $\{Z_{1,t}\}_{t=1}^{n}$ and $\{Z_{2,t}\}_{t=1}^{n}$ model the additive noise at Receivers 1 and 2. The noise sequences $\{Z_{1,t}\}_{t=1}^{n}$ and $\{Z_{2,t}\}_{t=1}^{n}$ are independent and each consists of i.i.d. centered Gaussian random variables of unit variance[3].

The goal of the communication is that the transmitter conveys Message M_1 to Receiver 1 and Message M_2 to Receiver 2. The transmitter observes perfect output-feedback from both receivers. Thus, the time-t channel input X_t can depend on all previous channel outputs Y_1^{t-1} and Y_2^{t-1} and the messages M_1 and M_2:

$$X_t = g_t^{(n)}(M_1, M_2, Y_1^{t-1}, Y_2^{t-1}), \quad t \in \{1, \ldots, n\}, \tag{11}$$

for some encoding function $g_t^{(n)}\colon \mathcal{M}_1 \times \mathcal{M}_2 \times \mathbb{R}^{t-1} \times \mathbb{R}^{t-1} \to \mathbb{R}$. We impose an *expected average block-power constraint*

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}[X_t^2] \le P, \tag{12}$$

where the expectation is over the messages and the realizations of the channel. Each Receiver $i \in \{1,2\}$ produces the guess

$$\hat{M}_i^{(n)} = \phi_i^{(n)}(Y_i^n)$$

[2] Notice that also Ozarow's rate-splitting scheme has the form in (5) because the feedback signals are combined linearly with code-symbols.

[3] As for the MAC, assuming $Z_{1,t}$ and $Z_{2,t}$ have unit variance entails no loss in generality.

$$C^{\text{linfb}}_{\text{MAC},\Sigma}(h_1,h_2;P) = \sup_{\substack{P_1,P_2\geq 0:\\ P_1+P_2=P}} \frac{1}{2}\log\left(1+h_1^2P_1+h_2^2P_2+2\sqrt{h_1^2h_2^2P_1P_2}\cdot\rho^\star(h_1,h_2;P_1,P_2)\right) \tag{8}$$

for some decoding function $\phi_i^{(n)}\colon \mathbb{R}^n \to \mathcal{M}_i$.

The average probability of error is

$$P^{(n)}_{\text{e,BC}} \triangleq \Pr\big[(\hat{M}_1 \neq M_1) \text{ or } (\hat{M}_2 \neq M_2)\big]. \tag{13}$$

A rate-pair (R_1, R_2) is achievable over the Gaussian BC with feedback and power constraint P, if there exists a sequence of encoding and decoding functions $\big\{\{g_t^{(n)}\}_{t=1}^n, \phi_1^{(n)}, \phi_2^{(n)}\big\}_{n=1}^\infty$ as described above, satisfying the power constraint (12) and such that the average probability of error $P^{(n)}_{\text{e,BC}}$ tends to zero as n tends to infinity.

Also here we restrict attention to *linear-feedback schemes for the BC* where the transmitter's channel input vector $\mathbf{X} \triangleq \big(X_1,\dots,X_n\big)^{\mathsf{T}}$ can be written as:

$$\mathbf{X} = \mathbf{W} + \mathsf{B}_1\mathbf{Z}_1 + \mathsf{B}_2\mathbf{Z}_2, \tag{14}$$

where $\mathbf{Z}_i \triangleq \big(Z_{i,1},\dots,Z_{i,n}\big)^{\mathsf{T}}$ represents the noise vector at Receiver i, B_1 and B_2 are strictly lower-triangular matrices, and $\mathbf{W}$ is an n-dimensional information-carrying vector

$$\mathbf{W} = \xi(M_1, M_2) \tag{15}$$

The mapping $\xi\colon \mathcal{M}_1\times\mathcal{M}_2 \to \mathbb{R}^n$ and the decoding operations $\phi_1^{(n)}$ and $\phi_2^{(n)}$ can be arbitrary.

Taking a linear combination of the information-carrying vector $\mathbf{W}$ and the past noise vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$ is equivalent to taking a (different) linear combination of $\mathbf{W}$ and the past outputs $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Thus, the strict lower-triangularity of B_1 and B_2 ensures that the feedback is used strictly causally.

Linear-feedback capacity and linear-feedback sum-capacity for the BC are defined analogously as for the MAC. We denote them by $\mathcal{C}^{\text{linfb}}_{\text{BC}}(h_1,h_2;P)$ and $C^{\text{linfb}}_{\text{BC},\Sigma}(h_1,h_2;P)$.

IV. Main Results

We first present multi-letter expressions for $\mathcal{C}^{\text{linfb}}_{\text{MAC}}(h_1,h_2;P)$ and $\mathcal{C}^{\text{linfb}}_{\text{BC}}(h_1,h_2;P)$. They are used to prove Theorem 1 ahead.

Definition 1. *Given $\eta \in \mathbb{N}$ and η-by-η matrices D_1 and D_2, let Q_1 and Q_2 be the positive square roots of the (positive-definite) η-by-η matrices*

$$\mathsf{M}_1 \triangleq (\mathsf{I}_\eta + h_1\mathsf{D}_1)^{\mathsf{T}}(\mathsf{I}_\eta + h_1\mathsf{D}_1) + h_1^2\mathsf{D}_2^{\mathsf{T}}\mathsf{D}_2 \tag{16a}$$

$$\mathsf{M}_2 \triangleq h_2^2\mathsf{D}_1^{\mathsf{T}}\mathsf{D}_1 + (\mathsf{I}_\eta + h_2\mathsf{D}_2)^{\mathsf{T}}(\mathsf{I}_\eta + h_2\mathsf{D}_2) \tag{16b}$$

and let $\mathcal{R}_{\text{MAC}}(\eta,\mathsf{D}_1,\mathsf{D}_2,h_1,h_2;P)$ denote the (private messages) nofeedback capacity [12] of the MIMO MAC

$$\tilde{\mathbf{Y}}^{\text{MAC}} \triangleq h_1\mathsf{Q}_1^{-1}\mathbf{V}_1 + h_2\mathsf{Q}_2^{-1}\mathbf{V}_2 + \tilde{\mathbf{Z}} \tag{17}$$

when the η-by-1 input vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ have to satisfy

$$\text{tr}(\mathsf{K}_{\mathbf{V}_1} + \mathsf{K}_{\mathbf{V}_2}) \leq \max\{0, \eta P - \text{tr}(\mathsf{D}_1\mathsf{D}_1^{\mathsf{T}}) - \text{tr}(\mathsf{D}_2\mathsf{D}_2^{\mathsf{T}})\}, \tag{18}$$

where $\mathsf{K}_{\mathbf{V}_i}$ denotes the covariance matrix of $\mathbf{V}_i$ and in (17) $\tilde{\mathbf{Z}}$ is a centered Gaussian vector of identity covariance matrix I_η.

Proposition 1. *The linear-feedback capacity of the scalar Gaussian MAC under sum-power constraint P satisfies*

$$\mathcal{C}^{\text{linfb}}_{\text{MAC}}(h_1,h_2;P) = \text{cl}\left(\bigcup_{\eta,\mathsf{D}_1,\mathsf{D}_2} \frac{1}{\eta}\mathcal{R}_{\text{MAC}}(\eta,\mathsf{D}_1,\mathsf{D}_2,h_1,h_2;P)\right) \tag{19}$$

where the union is over all positive integers η and strictly lower-triangular η-by-η matrices D_1 and D_2.

Observe that [1] and Proposition 1 imply that the right-hand sides (RHSs) of (19) and (6) coincide.

Proof: For fixed η, D_1, and D_2, the rate region $\frac{1}{\eta}\mathcal{R}_{\text{MAC}}(\eta,\mathsf{D}_1,\mathsf{D}_2,h_1,h_2;P)$ is achieved by coding over blocks of η channel uses. Choose the η channel inputs of Transmitter i for a given block as

$$\mathbf{X}_i = \mathsf{Q}_i^{-1}\mathbf{V}_i + \mathsf{C}_i\mathbf{Y}, \quad i \in \{1,2\}, \tag{20}$$

where

$$\mathsf{C}_i \triangleq \mathsf{D}_i(\mathsf{I}_\eta + h_1\mathsf{D}_1 + h_2\mathsf{D}_2)^{-1}, \tag{21}$$

and where $\mathbf{Y}$ denotes the η-by-1 vector consisting of the channel outputs in this block and $\mathbf{V}_i$ is an η-by-1 vector that depends on Message M_i, and over which we can code. The corresponding output vector is

$$\mathbf{Y} = (\mathsf{I}_\eta - h_1\mathsf{C}_1 - h_2\mathsf{C}_2)^{-1}(h_1\mathsf{Q}_1^{-1}\mathbf{V}_1 + h_2\mathsf{Q}_2^{-1}\mathbf{V}_2 + \mathbf{Z}). \tag{22}$$

By coding over the vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ of the different blocks, we can achieve the capacity of the MIMO MAC in (22), which equals the capacity of the MIMO MAC in (17). Algebraic manipulations show that the inputs in a given block (20) are sum-power constrained to P, if (18) holds and if $\eta P - \text{tr}(\mathsf{D}_1\mathsf{D}_1^{\mathsf{T}}) - \text{tr}(\mathsf{D}_2\mathsf{D}_2^{\mathsf{T}})$ is positive.

More details and the converse are omitted for brevity. ■

Definition 2. *Given $\eta \in \mathbb{N}$ and η-by-η matrices B_1 and B_2, let S_1 and S_2 be the positive square roots of the η-by-η matrices*

$$\mathsf{A}_1 \triangleq (\mathsf{I}_\eta + h_1\mathsf{B}_1)(\mathsf{I}_\eta + h_1\mathsf{B}_1)^{\mathsf{T}} + h_1^2\mathsf{B}_2\mathsf{B}_2^{\mathsf{T}} \tag{23a}$$

$$\mathsf{A}_2 \triangleq h_2^2\mathsf{B}_1\mathsf{B}_1^{\mathsf{T}} + (\mathsf{I}_\eta + h_2\mathsf{B}_2)(\mathsf{I}_\eta + h_2\mathsf{B}_2)^{\mathsf{T}} \tag{23b}$$

and let $\mathcal{R}_{\text{BC}}(\eta,\mathsf{B}_1,\mathsf{B}_2,h_1,h_2;P)$ denote the (private-messages) nofeedback capacity of the MIMO BC [13]

$$\tilde{\mathbf{Y}}_i^{\text{BC}} \triangleq h_i\mathsf{S}_i^{-1}\mathbf{W} + \tilde{\mathbf{Z}}_i, \quad i \in \{1,2\}, \tag{24}$$

when the η-by-1 input vector $\mathbf{W}$ has to satisfy

$$\text{tr}(\mathsf{K}_{\mathbf{W}}) \leq \max\{0, \eta P - \text{tr}(\mathsf{B}_1\mathsf{B}_1^{\mathsf{T}}) - \text{tr}(\mathsf{B}_2\mathsf{B}_2^{\mathsf{T}})\}, \tag{25}$$

where $\mathsf{K}_{\mathbf{W}}$ denotes the covariance matrix of $\mathbf{W}$ and where in (24) $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$ denote independent centered Gaussian vectors of identity covariance matrix I_η.

Proposition 2. *The linear-feedback capacity region of the Gaussian BC with feedback is:*

$$\mathcal{C}^{\text{linfb}}_{\text{BC}}(h_1,h_2;P) = \text{cl}\left(\bigcup_{\eta,\mathsf{B}_1,\mathsf{B}_2} \frac{1}{\eta}\mathcal{R}_{\text{BC}}(\eta,\mathsf{B}_1,\mathsf{B}_2,h_1,h_2;P)\right) \tag{26}$$

where the union is over all positive integers η and strictly lower-triangular η-by-η matrices B_1 and B_2.

Proof: For fixed η, B_1, and B_2, the rate region $\frac{1}{\eta}\mathcal{R}_{\text{BC}}(\eta, \mathsf{B}_1, \mathsf{B}_2, h_1, h_2; P)$ is achieved by coding over blocks of η channel uses, if the channel inputs in a block are

$$\mathbf{X} = \mathbf{W} + \mathsf{B}_1\mathbf{Z}_1 + \mathsf{B}_2\mathbf{Z}_2, \tag{27}$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ denote the block's η-by-1 noise vectors at Receivers 1 and 2 and $\mathbf{W}$ is an η-by-1 input vector that depends on the messages (M_1, M_2). Receiver i's outputs $\mathbf{Y}_i$ in a block are then given by

$$\mathbf{Y}_i = h_i\mathbf{W} + h_i\mathsf{B}_1\mathbf{Z}_1 + h_i\mathsf{B}_2\mathbf{Z}_2 + \mathbf{Z}_i, \quad i \in \{1, 2\}. \tag{28}$$

By coding over the inputs $\mathbf{W}$ of the different blocks, we can achieve the capacity of the MIMO BC in (28), which coincides with the capacity of the MIMO BC in (24).

More details and the converse are omitted for brevity. ■

Theorem 1. *The linear-feedback capacity regions of the scalar Gaussian BC under power constraint P and of the scalar Gaussian MAC under sum-power constraint P coincide:*

$$\mathcal{C}^{\text{linfb}}_{\text{MAC}}(h_1, h_2; P) = \mathcal{C}^{\text{linfb}}_{\text{BC}}(h_1, h_2; P). \tag{29}$$

Corollary 1.

$$C^{\text{linfb}}_{\text{MAC},\Sigma}(h_1, h_2; P) = C^{\text{linfb}}_{\text{BC},\Sigma}(h_1, h_2; P). \tag{30}$$

Corollary 2. *If $h_1 = h_2 = h$, then*

$$C^{\text{linfb}}_{\text{BC},\Sigma}(h, h; P) = \frac{1}{2}\log\left(1 + h^2P + h^2P \cdot \rho^\star(h, h; P, P)\right) \tag{31}$$

and thus the control-theoretic scheme in [7] achieves the linear-feedback sum-capacity.

Proof: By a symmetry argument. Omitted. ■

Proof of Theorem 1: We show that

$$\mathcal{R}_{\text{MAC}}(\eta, \mathsf{D}_1, \mathsf{D}_2, h_1, h_2; P) = \mathcal{R}_{\text{BC}}(\eta, \mathsf{B}_1, \mathsf{B}_2, h_1, h_2; P), \tag{32}$$

coincide if

$$\mathsf{B}_i = \bar{D}_i, \quad i \in \{1, 2\}, \tag{33}$$

where for a matrix A, $\bar{\mathsf{A}} \triangleq \mathsf{E}_\eta\mathsf{A}^\intercal\mathsf{E}_\eta$ is called its *reversed image* and E_η is the *exchange matrix* which is 0 everywhere except on the counter-diagonal where it is 1. The theorem then follows by Propositions 1 and 2, and since the mapping in (33) is one-to-one over the set of strictly lower-triangular matrices.

Under (33), the RHSs of the power constraints (18) and (25) coincide. Moreover, under power constraint (18), the MIMO MAC in (17) has the same capacity as the MIMO MAC[4]

$$\bar{\mathbf{Y}}^{\text{MAC}} \triangleq \mathsf{E}_\eta\tilde{\mathbf{Y}}^{\text{MAC}} = h_1\bar{\mathsf{Q}}_1^{-1}\bar{\mathbf{V}}_1 + h_2\bar{\mathsf{Q}}_2^{-1}\bar{\mathbf{V}}_2 + \bar{\mathbf{Z}}, \tag{34}$$

where $\bar{\mathbf{Z}} \triangleq \mathsf{E}_\eta\tilde{\mathbf{Z}}$ and where $\bar{\mathbf{V}}_i \triangleq \mathsf{E}_\eta\mathbf{V}_i$ has to satisfy the power constraint (18) when $\bar{\mathbf{V}}_i$ replaces $\mathbf{V}_i$. Now, Equality (32) follows by the MIMO MAC-BC (nofeedback) duality in [11], [15] and because under (33) the MAC $\bar{\mathbf{Y}}^{\text{MAC}}$ is dual to the BC in (24). In fact, under (33), $h_1\bar{\mathsf{Q}}_1^{-1} = h_1\mathsf{S}_1^{-\intercal}$ and $h_2\bar{\mathsf{Q}}_2^{-1} = h_2\mathsf{S}_2^{-\intercal}$. ■

Remark 1. *The optimal MAC scheme is described in [1]. From this we can deduce the optimal MAC-parameters $\mathbf{V}_1$, $\mathbf{V}_2$, C_1, and C_2 describing the block inputs in* (20). *(A different set of parameters is required to approach each point on the boundary of the capacity region.) Now, by* (32) *and comparing* (21) *and* (33), *from these parameters we can deduce the optimal BC-parameters B_1 and B_2 describing the block inputs in* (27). *Finally, the results in [13] tell us how to code and decode over the resulting MIMO BC in* (28).

Remark 2. *Theorem 1 extends to the scalar Gaussian MAC and BC with $K \geq 3$ users.*

Acknowledgements

The work of S. Belhadj Amor and M. Wigger has been supported by the city of Paris under the "Emergences" program. The work of Y. Steinberg has been supported by the Israel Science Foundation (grant no. 684/11).

References

[1] L. Ozarow, "The capacity of the white Gaussian multiple access channel with feedback," *IEEE Trans. on Inf. Th.*, vol. 30, no. 4, pp. 623–629, 1984.

[2] G. Kramer, "Feedback strategies for white Gaussian interference networks," *IEEE Trans. on Inf. Th.*, vol. 48, no. 6, pp. 1423–1438, 2002.

[3] E. Ardestanizadeh, M. Wigger, Y.H. Kim, and T. Javidi, "Linear-feedback sum-capacity for Gaussian multiple access channels," *IEEE Trans. on Inf. Th.*, vol. 58, no. 1, pp. 224–236, 2012.

[4] L. Ozarow and S. Leung-Yan-Cheong, "An achievable region and outer bound for the Gaussian broadcast channel with feedback," *IEEE Trans. on Inf. Th.*, vol. 30, no. 4, pp. 667–671, 1984.

[5] N. Elia, "When Bode meets Shannon: control-oriented feedback communication schemes," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, 2004.

[6] S. Vishwanath, W. Wu, and A. Arapostathis, "Gaussian interference networks with feedback: duality, sum capacity and dynamic team problems," in *Proc. 44th Ann. Allerton Conf.* 2005.

[7] E. Ardestanizadeh, P. Minero, and M. Franceschetti, "LQG control approach to Gaussian broadcast channels with feedback," *IEEE Trans. on Inf. Th.*, vol. 58, no. 8, pp. 5267–5278, 2012.

[8] M. Gastpar, A. Lapidoth, Y. Steinberg, and M. Wigger, "New achievable rates for the Gaussian broadcast channel with feedback," in *Proceedings of ISWCS* 2011, pp. 579–583.

[9] O. Shayevitz and M. Wigger, "On the capacity of the discrete memoryless broadcast channel with feedback," *IEEE Trans. on Inf. Th.*, vol. 59, no. 3, pp. 1329–1345, 2013.

[10] R. Venkataramanan and S.S. Pradhan, "An achievable rate region for the broadcast channel with feedback," submitted to *IEEE Trans. on Inf. Th.*, May 2011, available at http://arxiv.org/abs/1105.2311.

[11] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. on Inf. Theory*, vol. 49, no. 10, pp. 2658-2668, 2003.

[12] R.S. Cheng and S. Verdu, "Gaussian multiaccess channels with ISI: capacity region and multiuser water-filling," *IEEE Trans. on Inf. Theory*, vol. 39, no. 3, pp. 773–785, 1993.

[13] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. on Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[14] Y. Wu and M. Wigger, "Any positive feedback rate increases the capacity of strictly less-noisy broadcast channels," in *Proceedings of ITW* 2013.

[15] P. Viswanath and D.N.C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. on Inf. Theory*, vol. 49, no. 8, pp. 1912-1921, 2003.

[4] Multiplying $\tilde{\mathbf{Y}}^{\text{MAC}}$ by E_η from the left only reverses the order of receiving antennas.

International Zurich Seminar on Communications (IZS), February 26 – 28, 2014

Analysis of Mismatched Estimation Errors Using Gradients of Partition Functions

Wasim Huleihel and Neri Merhav
Technion - Israel Institute of Technology
Department of Electrical Engineering
Haifa 32000, ISRAEL
E-mail: {wh@tx, merhav@ee}.technion.ac.il

***Abstract*—We consider the problem of signal estimation (denoising) from a statistical-mechanical perspective, in continuation to a recent work on the analysis of mean-square error (MSE) estimation using a direct relationship between optimum estimation and certain partition functions. Accordingly, we derive a single-letter expressions of the MMSE and mismatched MSE of a codeword (from a randomly selected code), corrupted by a Gaussian vector channel, and we provide several examples to demonstrate phase transitions in the behavior of the MSE.**

I. Introduction

The connections and the interplay between information theory, statistical physics and signal estimation have been known for several decades, and they are still being studied from a variety of aspects, see, for example [1-3] and many references therein.

Recently, in [2], the well known I-MMSE relation [3], which relates the mutual information and the derivative of the minimum mean-square error (MMSE), was further explored using a statistical physics perspective. One of the main contributions in [2] is the demonstration of the usefulness of statistical-mechanical tools (in particular, utilizing the fact that the mutual information can be viewed as the partition function of a certain physical system) in assessing MMSE via the I-MMSE relation of [3]. More recently, Merhav [1] proposed a more flexible method, whose main idea is that, for the purpose of evaluating the covariance matrix of the MMSE estimator, one may use other information measures, which have the form of a partition function and hence can be analyzed using methods of statistical physics (see, e.g., [4] and many references therein). The main advantage of the proposed approach over the I-MMSE relations, is its full generality: Any joint probability function $P(\boldsymbol{x},\boldsymbol{y})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ designate the channel input to be estimated and the channel output, respectively, can be handled (for example, the channel does not have to be additive or Gaussian). Moreover, using this approach, any mismatch, both in the source and the channel, can be considered.

This paper is a further development of [1] in the above described direction. Particularly, in [1, Section IV. A], the problem of mismatched estimation of a codeword, transmitted over an additive white Gaussian (AWGN) channel, was considered. It was shown that the mismatched MSE exhibits phase transitions at some rate thresholds, which depend upon the real and the mismatched parameters of the problem, and the behavior of the receiver. To wit, the mismatched MSE acts inherently differently for a *pessimistic* and *optimistic* receivers, where in the example considered in [1, Section IV. A] pessimism literally means that the estimator assumes that the channel is worse than it really is (in terms of signal-to-noise ratio (SNR)), and the vice versa for optimism. In this paper, we extend the above described model to a much more general one; the Gaussian vector channel, which has a plenty of applications in communications and signal processing. It is important to emphasize that compared to [1, 2], the mathematical analysis is much more complicated (consisting of some new concepts), and the notions of pessimism and optimism described above, also play a significant role in this model, although their physical meanings in general are not obvious. Moreover, in contrast to previous work on mismatched estimation, the case of channel mismatch is explored, namely, the receiver has a wrong assumption on the channel.

II. Model and Problem Formulation

Let $\mathcal{C} = \{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{M-1}\}$ denote a codebook of size $M = e^{nR}$, which is selected at random (and then revealed to the estimator) in the following manner: Each $\boldsymbol{x}_i$ is drawn independently under the uniform distribution over the surface of the n-dimensional hyperesphere, which is centered at the origin, and whose radius is $\sqrt{nP_x}$. Finally, let $\boldsymbol{X}$ assume a uniform distribution over $\mathcal{C}$. We consider the Gaussian vector channel model

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{N}$ are random vectors in $\mathbb{R}^n$, designating the channel output vector, the transmitted codeword and the noise vector, respectively. It is assumed that the components of the noise vector, $\boldsymbol{N}$, are i.i.d., zero-mean, Gaussian random variables with variance $1/\beta$, where β is a given positive constant designating the signal-to-noise ratio (SNR) (for $P_x = 1$), or the inverse temperature in the statistical-mechanical jargon. We further assume that $\boldsymbol{X}$ and $\boldsymbol{N}$ are statistically independent. Finally, the channel matrix, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, is assumed to be a given deterministic Toeplitz matrix, whose entries are given by the coefficients of the impulse response of a given linear system. Specifically, let $\{h_k\}$ denote the generating sequence (or impulse response) of $\boldsymbol{A}$, so that $\boldsymbol{A} = \{a_{i,j}\}_{i,j} = \{h_{i-j}\}_{i,j}$, and let $H(\omega)$ designate the frequency response (Fourier transform) of $\{h_k\}$.

There are several motivations for codeword estimation. One example is that of a user that, in addition to its desired signal, receives also a relatively strong interference signal, which carries digital information intended to other users, and which comes from a codebook whose rate exceeds the capacity of this crosstalk channel between the interferer and our user, so that the user cannot fully decode this interference. Nevertheless, our user would like to estimate the interference as accurately as possible for the purpose of cancellation. Furthermore, we believe that the tools/concepts developed in this paper for handling matched and mismatched problems, can be used in other applications in signal processing and communication. Such examples are denoising, mismatched decoding, blind deconvolution, and many other applications. Note that although the aforementioned examples are radically different (in terms of their basic models and systematization), they will all suffer from mismatch when estimating the input signals.

As was mentioned previously, we analyze the problem of mismatched codeword estimation which is formulated as follows: Consider a mismatched estimator which is the conditional mean of $\boldsymbol{X}$ given $\boldsymbol{Y}$, based on an incorrect joint distribution $P'(\boldsymbol{x},\boldsymbol{y})$, whereas the true joint distribution continues to be $P(\boldsymbol{x},\boldsymbol{y})$. Accordingly, the *mismatched MSE* is defined as

$$\mathrm{mse}(\boldsymbol{X}\mid\boldsymbol{Y}) \triangleq \boldsymbol{E}\left\|\boldsymbol{X}-\boldsymbol{E}'\left\{\boldsymbol{X}\mid\boldsymbol{Y}\right\}\right\|^2 \tag{2}$$

where $\boldsymbol{E}'\{\boldsymbol{X}\mid\boldsymbol{Y}\}$ is the conditional expectation with respect to (w.r.t.) the mismatched measure P'. In this paper, the following mismatch mechanism is assumed: The input measure is matched, i.e., $P(\boldsymbol{x}) = P'(\boldsymbol{x})$ (namely, the mismatched estimator knows the true code), both conditional measures ("channels") $P(\cdot\mid\boldsymbol{x})$ and $P'(\cdot\mid\boldsymbol{x})$ are Gaussian, but are associated with different channel matrices. More precisely, while the true channel matrix (under P) is $\boldsymbol{A}$, the assumed channel matrix (under P') is $\boldsymbol{A}'$, another Toeplitz matrix, generated by the impulse response $\{h_k'\}$, whose frequency response is $H'(\omega)$. It should be pointed out, however, that the analysis in this paper can be easily carried out also for the case of mismatch in the input distribution, or mismatch in the noise distribution, which has been already considered in [1]. In the matched case, $P = P'$, we use the notation $\mathrm{mmse}(\boldsymbol{X}\mid\boldsymbol{Y}) = \mathrm{mse}(\boldsymbol{X}\mid\boldsymbol{Y})$.

A very important function, which is pivotal to the derivation of both the estimator and the MSE is the *partition function*, defined as follows.

Definition 1 (Partition Function) Let $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_n)^T$ be a column vector of n real-valued parameters. The partition function w.r.t. the joint distribution $P(\boldsymbol{x},\boldsymbol{y})$, denoted by $Z(\boldsymbol{y},\boldsymbol{\lambda})$, is defined as

$$Z(\boldsymbol{y},\boldsymbol{\lambda}) \triangleq \sum_{\boldsymbol{x}\in\mathcal{C}} \exp\left\{\boldsymbol{\lambda}^T\boldsymbol{x}\right\} P(\boldsymbol{x},\boldsymbol{y}) \,. \tag{3}$$

Accordingly, under the above described model, we have that

$$P'(\boldsymbol{y}\mid\boldsymbol{x}) = \frac{1}{(2\pi/\beta)^{n/2}} \exp\left[-\beta\left\|\boldsymbol{y}-\boldsymbol{A}'\boldsymbol{x}\right\|^2/2\right], \tag{4}$$

and so, the mismatched partition function is given by

$$Z'(\boldsymbol{y},\boldsymbol{\lambda}) \triangleq \sum_{\boldsymbol{x}\in\mathcal{C}} \exp\left\{\boldsymbol{\lambda}^T\boldsymbol{x}\right\} P'(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

$$= (2\pi/\beta)^{-n/2} \sum_{\boldsymbol{x}\in\mathcal{C}} e^{-nR} \exp\left[-\beta\left\|\boldsymbol{y}-\boldsymbol{A}'\boldsymbol{x}\right\|^2/2+\boldsymbol{\lambda}^T\boldsymbol{x}\right]. \tag{6}$$

The role of $\boldsymbol{\lambda}$ in the above partition function is in computing the conditional mean estimator and the MSE. Indeed, it is easy to see that the gradient of $\ln Z'(\boldsymbol{y},\boldsymbol{\lambda})$ w.r.t. $\boldsymbol{\lambda}$, computed at $\boldsymbol{\lambda}=\boldsymbol{0}$, simply gives the conditional mean estimator, i.e.,

$$\boldsymbol{E}'\left\{\boldsymbol{X}\mid\boldsymbol{Y}=\boldsymbol{y}\right\} = \nabla_{\boldsymbol{\lambda}} \ln Z'(\boldsymbol{y},\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=0} \tag{7}$$

where $\nabla_{\boldsymbol{\lambda}}$ denotes the gradient operator w.r.t. $\boldsymbol{\lambda}$. Also, in the matched case, it can be verified that the expectation of the Hessian of $\ln Z(\boldsymbol{y},\boldsymbol{\lambda})$ w.r.t. $\boldsymbol{\lambda}$, computed at $\boldsymbol{\lambda}=\boldsymbol{0}$, gives the MMSE, i.e.,

$$\mathrm{mmse}(\boldsymbol{X}\mid\boldsymbol{Y}) = \mathrm{tr}\left\{\boldsymbol{E}\left(\nabla^2_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=0}\right)\right\}. \tag{8}$$

where $\nabla^2_{\boldsymbol{\lambda}}$ denotes the Hessian operator w.r.t. $\boldsymbol{\lambda}$. Using (7), the mismatched MSE can be calculated as

$$\begin{aligned}\mathrm{mse}(\boldsymbol{X}\mid\boldsymbol{Y}) &\triangleq \sum_{i=1}^n \boldsymbol{E}\left\{\left(X_i-\boldsymbol{E}'\left\{X_i\mid\boldsymbol{Y}\right\}\right)^2\right\}\\ &= \sum_{i=1}^n\left[\boldsymbol{E}\left\{X_i^2\right\}+\boldsymbol{E}\left\{\left[\frac{\partial \ln Z'(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\right]^2\Bigg|_{\boldsymbol{\lambda}=0}\right\}\right.\\ &\quad\left.-2\boldsymbol{E}\left\{\frac{\partial \ln Z(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\Bigg|_{\boldsymbol{\lambda}=0}\cdot\frac{\partial \ln Z'(\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\lambda_i}\Bigg|_{\boldsymbol{\lambda}=0}\right\}\right].\end{aligned} \tag{9}$$

All the above relations (and further) can be found in [1].

III. Main Result and Discussion

In this section, our main results are presented and discussed. Due to space limitation, the proofs of all the following results are omitted and can be found in [5]. The asymptotic MMSE is given in the following theorem.

Theorem 1 (Asymptotic MMSE) Consider the model defined in Section II, and assume that the sequence $\{h_k\}$ is square summable. Then, the asymptotic MMSE is given by

$$\lim_{n\to\infty}\frac{\mathrm{mmse}(\boldsymbol{X}\mid\boldsymbol{Y})}{n} = \begin{cases}\frac{1}{2\pi}\int_0^{2\pi}\frac{P_x}{1+|H(\omega)|^2P_x\beta}\mathrm{d}\omega, & R>R_c\\ 0, & R\leq R_c\end{cases} \tag{10}$$

where $R_c \triangleq \frac{1}{4\pi}\int_0^{2\pi}\ln\left(1+|H(\omega)|^2P_x\beta\right)\mathrm{d}\omega$.

From the above result, it can be seen that for $R<R_c$ the MMSE essentially vanishes since the correct codeword can be reliably decoded, whereas for $R>R_c$ the MMSE is simply the estimation error which results by the Wiener filter that would have been applied had the input been a zero-mean, i.i.d. Gaussian process, with variance $1/\beta$. Accordingly, it can be shown that (as a byproduct of the analysis) the MMSE estimator is exactly the Wiener filter. In the jargon of statistical mechanics of spin arrays (see for example [4, Ch. 6]), the range of rates $R\leq R_c$, correspond to the ordered phase (or ferromagnetic phase) in which the partition function is dominated by the correct codeword (and hence so is the posterior), while the range of rates $R > R_c$ corresponds to the paramagnetic phase, in which the partition function is dominated by an exponential number of wrong codewords.

In contrast to the MMSE, unfortunately, the mismatched MSE does not lend itself to a simple closed-form expression. This complexity stems from the complicated dependence of the partition function on $\boldsymbol{\lambda}$. Nevertheless, despite of the non-trivial expressions, it should be emphasized that the obtained MSE expression has a single letter formula, and thus, practically, it can be easily calculated at least numerically. Due to the complicated expressions obtained for the MSE, in the following, we only present the general structure/behavior (in the sense of phase transitions) of the MSE without presenting the absolute error itself. It is shown in [5] that the MSE takes the following form: For $R_d \geq 0$ the MSE is given by

$$\lim_{n\to\infty} \frac{\text{mse}\left(\boldsymbol{X} \mid \boldsymbol{Y}\right)}{n} = \begin{cases} 0, & R \leq R_s \\ E_p, & R > R_s \end{cases}, \tag{11}$$

and for $R_d < 0$ it is given by

$$\lim_{n\to\infty} \frac{\text{mse}\left(\boldsymbol{X} \mid \boldsymbol{Y}\right)}{n} = \begin{cases} 0, & R \leq R_g \\ E_g, & R_g < R \leq R_e \\ E_p, & R > R_e \end{cases} \tag{12}$$

where the various parameters (R_d, R_s, etc.) in the above expressions are not presented here due to space limitations, but can be found in [5]. Thus, it can be seen that in the mismatched case, there is additional intermediate range (when $R_d < 0$), which in statistical mechanics jargon is analogous to the glassy phase (or "frozen" phase), in which the partition function is dominated by a sub-exponential number of wrong codewords. Intuitively, in this range, we may have the illusion that there is relatively little uncertainty about the transmitted codeword, but this is wrong due to the mismatch (as the main support of the mismatched posterior belongs to incorrect codewords). In Section IV, we will relate each one of the two cases $R_d \geq 0$ and $R_d < 0$, to "pessimistic" and "optimistic" behaviors of the receiver, which were already mentioned in the Introduction.

In the following, we state a few general qualitative properties of the various quantities appearing in the obtained results. Similarly to [1], it turns out that the absolute error E_p is independent on R, while E_g depends on R non-trivially. Accordingly, unlike the matched and pessimistic mismatched cases, the MSE is not piecewise constant in the whole range of rates when the estimator is optimistic. Also, as the SNR increases, the absolute errors E_g and E_p decrease, while the critical ferromagnetic rate (R_s if $R_d \geq 0$ and R_g otherwise) increases, as should be expected. Finally, while in the matched case the MMSE is independent of the filter/channel phase (readily seen from Theorem 1), in the mismatched case, this conclusion is not true anymore. This fact is demonstrated in Section IV.

Finally, note that it is tempting to think that there should not be a range of rates for which the MSE (MMSE) vanishes, as we deal with an estimation problem rather than a decoding problem. Nonetheless, since codewords are being estimated, and there are a finite number of them, for low enough rates (up to some critical rate) the posterior is dominated by the correct codeword, and thus asymptotically, the estimation can be regarded as a maximum a posteriori probability (MAP) estimation, and so the error vanishes. In the same breath, note that this is not the case if mismatch in the input distribution is considered. For example, if the receiver's assumption on the transmitted energy is wrong, then no matter how low the rate is, there will always be an inherent error which stems from the fallacious averaging over a hypersphere with wrong radius (wrong codebook). Precisely, in this case, the estimated codeword will differ from the real one by an inevitable scaling of $\sqrt{P_x'/P_x}$, where P_x' is the mismatched power.

Remark 1 Although we have assumed that the transmitted codeword has a flat spectrum, the analysis can readily be extended to any input spectral density $S_x(\omega)$.

IV. Examples

In this section, we provide two examples in order to illustrate the theoretical results presented in the previous section. In particular, we present and explore the phase diagrams and the MSE's as functions of the rate and some parameters of the mismatched channel. The main goal in these examples is further understanding of the role of the true and the mismatched probability measures in creating phase transitions. Further examples can be found in [5].

Example 1 Let $H(\omega)$ be a multiband filter given by

$$H(\omega) = \begin{cases} 1, & \left|\omega \pm \frac{3\pi}{8}\right| \leq \frac{\pi}{8} \text{ or } \left|\omega \pm \frac{7\pi}{8}\right| \leq \frac{\pi}{8} \\ 0, & \text{else} \end{cases}, \tag{13}$$

and let the mismatched filter be given by a band-pass filter

$$H'(\omega) = \begin{cases} 1, & \omega_L \leq |\omega| \leq \omega_R \\ 0, & \text{else} \end{cases}, \tag{14}$$

with constant bandwidth, $\omega_R - \omega_L = \pi/8$, i.e., smaller than the real one. In the numerical calculations, we again chose $\beta = P_x = 1$. Figures 1 and 2 show, respectively, the phase diagram and the MSE as functions of R and ω_L. First, observe that for $\omega_R < \pi/4$, which means that $H'(\omega)$ and $H(\omega)$ are equal to one over non intersecting frequency ranges, there is no ferromagnetic phase, as expected. Accordingly, for $\omega_R > \pi/4$, the ferromagnetic phase begins to play a role, and it can be seen that for $\pi/4 + \pi/8 < \omega_R < \pi/2$, which means maximal intersection between the two filters, the range of rates for which the ferromagnetic phase dominates the partition function is maximal. Since the matched filter has two bands, obviously, the same behavior appears also in the second band. Thus, in this example, we actually obtain two disjoint glassy (and ferromagnetic) regions, which correspond to the two bands of the matched filter. Also, as shown in Fig. 2, in the ranges where no ferromagnetic phase exists, the MSE within the paramagnetic phase is larger than the MSE within the regions where ferromagnetic phase does exists, as one would expect.

Remark 2 Example 1 essentially demonstrates that there can be arbitrarily many phase transitions. Generally speaking, for a matched multiband filter with N disjoint bands, and a

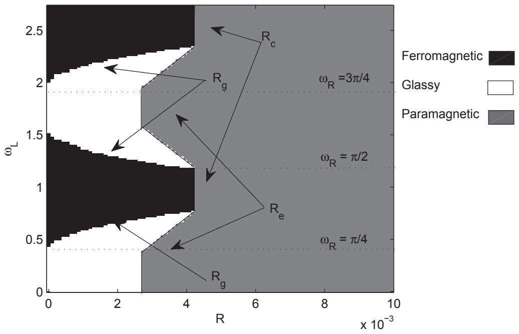


Fig. 1. Example 1: Phase diagram in the plane of R vs. ω_L.

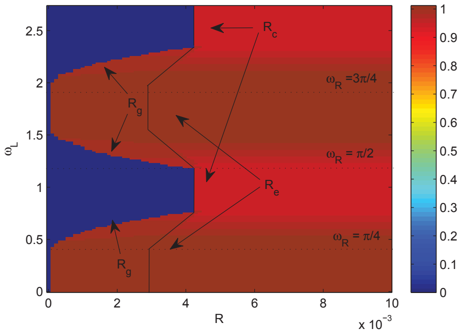


Fig. 2. Example 1: Mismatched MSE as a function of R and ω_L.

mismatched bandpass filter (with small enough bandwidth), there are N disjoint glassy and ferromagnetic phases.

Example 2 Let $H(z)$ be given by

$$H(z) = z - 2\cos(0.8\pi) + z^{-1}$$
$$= z \cdot \left(1 - e^{j0.8\pi} z^{-1}\right)\left(1 - e^{-j0.8\pi} z^{-1}\right) \quad (15)$$

and let the mismatched filter be given as

$$H'(z) = H(z)\, z^{-d} \quad (16)$$

where $d \in \mathbb{Z}$ is a mismatched delay. As before, in the numerical calculations, we chose $\beta = P_x = 1$. Figures 3 and 4 show, respectively, the phase diagram and the MSE as functions of R and d. First, we see that R_e is constant, which makes sense since it can be shown that R_e is independent of the delay [5]. Also, for all $d \neq 0$ there is a glassy phase, which means that for all $d \neq 0$, $R_d \leq 0$. More importantly, it can be observed that the MSE vanishes (or equivalently, the ferromagnetic phase dominates the partition function) only in case that $d = 0$, namely, zero delay. This is a reasonable result, as a delay of one sample (linear phase) is enough to cause a serious degradation in the MSE. Actually, for any fixed rate the error is constant, independently of the delay, due to the fact that the MSE takes into account all the possible codewords in the codebook. Finally, note that the MSE is larger in the glassy region than in the paramagnetic region[1]. This is also

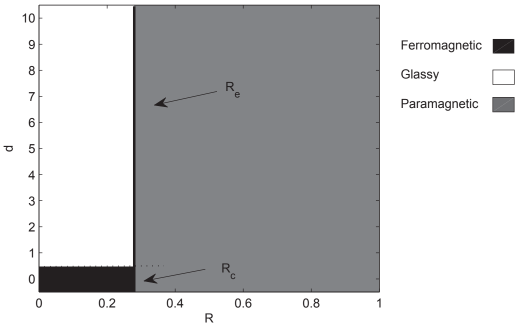


Fig. 3. Example 2: Phase diagram in the plane of R vs. d.

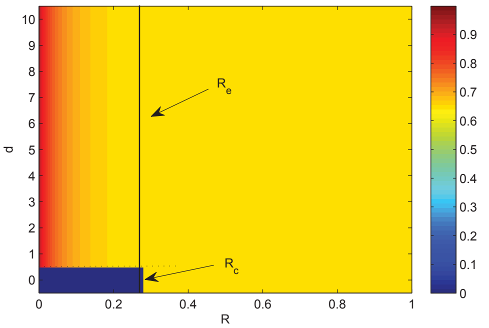


Fig. 4. Example 2: Mismatched MSE as a function of R and d.

a reasonable result: As the rate increases, and hence more codewords are possible, since the MSE estimator is actually a weighted average (w.r.t. the posterior) over the codewords, the MSE can only decrease (each codeword in the codebook contributes approximately the same estimation error). Accordingly, for small codebooks (low rates) the MSE is larger, since the averaging is performed over “fewer” codewords.

ACKNOWLEDGMENT

This research was partially supported by The Israeli Science Foundation (ISF), grant no. 412/12.

REFERENCES

[1] N. Merhav, “Optimum estimation via gradients of partition functions and information measures: A statistical-mechanical perspective,” *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3887–3898, June 2011.

[2] N. Merhav, D. Guo, and S. Shamai, “Statistical physics of signal estimation in Gaussian noise: theory and examples of phase transitions,” *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1400–1416, Mar. 2010.

[3] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[4] A. Mézard, M. Montanari, *Information, Physics and Computation*. Oxford, U.K.: Oxford Univ. Press., 2009.

[5] W. Huleihel and N. Merhav, “Analysis of mismatched estimation errors using gradients of partition functions,” *submitted to IEEE Trans. Inf. Theory*, June 2013. [Online]. Available: http://arxiv.org/pdf/1306.0094.pdf

[1] Note that the MSE, in contrast to the MMSE, must not be monotonically increasing as a function of the rate.

On an Extremal Data Processing Inequality for long Markov Chains

Thomas A. Courtade
University of California, Berkeley
Department of Electrical Engineering and Computer Science
Email: courtade@eecs.berkeley.edu

Jiantao Jiao and Tsachy Weissman
Stanford University
Department of Electrical Engineering
Email: {jiantao, tsachy}@stanford.edu

***Abstract*—We pose the following extremal conjecture: Let X, Y be jointly Gaussian random variables with linear correlation ρ. For any random variables U, V for which U, X, Y, V form a Markov chain, in that order, we conjecture that:**

$$2^{-2[I(X;V)+I(Y;U)]} \geq (1-\rho^2)2^{-2I(U;V)} + \rho^2 2^{-2[I(X;U)+I(Y;V)]}.$$

By letting V be constant, we see that this inequality generalizes a well-known extremal result proved by Oohama in his work on the quadratic Gaussian one-helper problem. If valid, the conjecture would have some interesting consequences. For example, the converse for the quadratic Gaussian two-encoder source coding problem would follow from the converse for multiterminal source coding under logarithmic loss, thus unifying the two results under a common framework.

Although the conjecture remains open, we discuss both analytical and numerical evidence supporting its validity.

I. Introduction

This paper is a brief exposition on the following conjecture, its potential applications, and evidence supporting its validity. To this end, we propose:

Conjecture 1. *Suppose X, Y are jointly Gaussian, each with unit variance and correlation ρ. Then, for any U, V satisfying $U - X - Y - V$, the following inequality holds:*

$$2^{-2[I(Y;U)+I(X;V|U)]} \geq (1-\rho^2) + \rho^2 2^{-2[I(X;U)+I(Y;V|U)]}. \tag{1}$$

In the statement of Conjecture 1, we employ the conventional notation $U-X-Y-V$ to denote that U, X, Y, V form a Markov chain, in that order. Throughout this paper, X, Y will have the distribution given in the statement of the conjecture.

Our interest in Conjecture 1 stems from previous work by two of the present authors on multiterminal source coding under logarithmic loss [1]. In order to illustrate the connection between these problems, define $\mathcal{R} \subset \mathbb{R}^2$ as follows. Let $(R, I) \in \mathcal{R}$ if and only if there exists Q independent of X, Y, and U, V satisfying

$$R \geq I(X,Y;U,V|Q) \tag{2}$$
$$I \leq I(X;U,V|Q) + I(Y;U,V|Q), \tag{3}$$

and, conditioned on Q, the Markov relation $U - X - Y - V$.

Next, let P_{XY} denote the joint distribution of X, Y, and assume $(X^n, Y^n) \sim \prod_{i=1}^n P_{XY}(x_i, y_i)$. For functions

$$\mathsf{f}_x : X^n \mapsto \mathsf{f}_x(X^n) \in \{1, 2, \ldots, 2^{nR_x}\} \tag{4}$$
$$\mathsf{f}_y : Y^n \mapsto \mathsf{f}_y(Y^n) \in \{1, 2, \ldots, 2^{nR_y}\}, \tag{5}$$

define

$$I(n, \mathsf{f}_x, \mathsf{f}_y) \triangleq \frac{1}{n}\Big(I(X^n; \mathsf{f}_x(X^n), \mathsf{f}_y(Y^n)) + I(Y^n; \mathsf{f}_x(X^n), \mathsf{f}_y(Y^n))\Big),$$

$$\mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\Big(X_i - \mathbb{E}[X_i|\mathsf{f}_x(X^n), \mathsf{f}_y(Y^n)]\Big)^2\right], \tag{6}$$

and $\mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y)$ in an analogous manner. These quantities satisfy the inequality

$$-\frac{1}{2}\log\Big(\mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y)\Big) - \frac{1}{2}\log\Big(\mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y)\Big) \leq I(n, \mathsf{f}_x, \mathsf{f}_y), \tag{7}$$

which easily follows by convexity, the maximum entropy property of Gaussian random variables, and the memoryless property of X^n, Y^n.

An immediate consequence of the converse for multiterminal source coding under logarithmic loss is that $(R_x + R_y, I(n, \mathsf{f}_x, \mathsf{f}_y)) \in \mathcal{R}$, which easily follows from [1] and the corresponding entropy characterization result [2, Theorem 2].

Now, to show an interesting application of Conjecture 1, assume (1) holds. Combined with the fact that $(R_x + R_y, I(n, \mathsf{f}_x, \mathsf{f}_y)) \in \mathcal{R}$, elementary manipulations on (1) and (7) would reveal that

$$R_1 + R_2 \geq \frac{1}{2}\log\left[\frac{(1-\rho^2)\beta(D)}{2D}\right], \tag{8}$$

where we have defined

$$D \triangleq \mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y) \times \mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y), \text{ and} \tag{9}$$

$$\beta(\xi) \triangleq 1 + \sqrt{1 + \frac{4\rho^2\xi}{(1-\rho^2)^2}} \tag{10}$$

This work was supported by the NSF Center for Science of Information under grant agreement CCF-0939370.

for notational convenience. We note that (8) is precisely the sum-rate constraint for the quadratic Gaussian two-encoder source coding problem first established in the seminal work [3] by Wagner et al.

Thus, while we have only sketched the argument here, we hope the reader is convinced that the sum-rate constraint for the quadratic Gaussian two-encoder source coding problem would follow in a relatively straightforward manner from known results on compression under logarithmic loss and the conjectured extremal inequality (1). In fact, the entire converse (not only the sum-rate constraint) for the quadratic Gaussian two-encoder source coding problem would follow from Conjecture 1 and the characterization of the rate-distortion region for compression under log loss. Details are omitted due to space constraints.

On the term "Data Processing"

As the title suggests, we refer to (1) as a *data processing inequality* since it gives the upper bound

$$\begin{aligned} &I(Y;U) + I(X;V|U) \\ &\quad \le -\frac{1}{2}\log\left[1-\rho^2+\rho^2 2^{-2[I(X;U)+I(Y;V|U)]}\right]. \end{aligned} \tag{11}$$

By straightforward calculus, a simple corollary is, for example, the upper bound

$$I(Y;U) \le \rho^2 I(X;U), \tag{12}$$

which falls into the category of so-called *strong data processing inequalities* (cf. [4], [5]). Since (1) is met with equality when U, X, Y, V are jointly Gaussian, (11) would provide the best possible data processing inequality of the form

$$I(Y;U) + I(X;V|U) \le \psi(I(X;U) + I(Y;V|U)), \tag{13}$$

under our assumptions on U, X, Y, V.

II. Observations on Conjecture 1

There are many equivalent forms of Conjecture 1. It seems particularly useful to consider dual forms of Conjecture 1. For instance, one such form is stated as follows:

Conjecture 2. *Let X, Y be jointly Gaussian, each with unit variance and correlation ρ. For $\lambda > 1/\rho^2$, the infimum of*

$$I(X;U) + I(Y;V|U) - \lambda\Big(I(Y;U) + I(X;V|U)\Big) \tag{14}$$

taken over all U, V satisfying $U - X - Y - V$ is attained when U, X, Y, V are jointly Gaussian.

Note that we only conjecture that the minimum of (14) is attained by U, V which are jointly Gaussian with X, Y. Clearly, since mutual information is invariant under one-to-one transformations, there are minimizers of (14) which are non-Gaussian.

Let $F^\star_\lambda$ be the infimum of the functional (14) for fixed $\lambda > 1/\rho^2$. If Conjecture 1 were to hold, then straightforward computations reveal that $F^\star_\lambda$ would be given by

$$F^\star_\lambda = \frac{1}{2}\left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right) - \lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right]. \tag{15}$$

It is interesting to note that we also have[1]

$$\begin{aligned} &\inf_{U:U-X-Y}\Big\{I(X;U) - \lambda I(Y;U)\Big\} \\ &\quad = \frac{1}{2}\left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right) - \lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right]. \end{aligned} \tag{16}$$

Since (14) can be rewritten as

$$\begin{aligned} &\Big(I(X;U) - \lambda I(Y;U)\Big) + \Big(I(Y;V) - \lambda I(X;V)\Big) \\ &\quad + (\lambda-1)I(U;V) \end{aligned} \tag{17}$$

by Markovity, the conjecture implies an unexpected conservation property: either U and V can be optimized jointly in minimizing (14), or we can set V to be constant and only optimize over U (or vice versa). Assuming the conjecture is valid, both approaches yield the same optimal value, which suggests one should eliminating one of the variables is a viable proof strategy. Unfortunately, this has proved difficult to do. In any case, (16) and (17) yield the lower bound

$$F^\star_\lambda \ge \left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right) - \lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right], \tag{18}$$

which reveals why we need only consider $\lambda > 1/\rho^2$ in Conjecture 2: for $\lambda \le 1/\rho^2$, the infimum of (14) is zero.

Moving on, if we were to assume the conjecture were true, and let optimizing $U^\star, V^\star$ be of the form

$$U^\star = \rho_u X + Z_u \tag{19}$$

$$V^\star = \rho_v X + Z_v, \tag{20}$$

where $Z_u \sim N(0, 1-\rho_u^2)$ and $Z_v \sim N(0, 1-\rho_v^2)$ are independent additive Gaussian noises, then the parameters ρ_u, ρ_v should satisfy the following equation, which gives an intuitive sense for the tension between the conjectured optimizers $U^\star$ and $V^\star$:

$$(1-\rho^2)(1-\rho^2\rho_u^2\rho_v^2) = \rho^2(\lambda-1)(1-\rho_u^2)(1-\rho_v^2). \tag{21}$$

In particular, for given ρ, λ, there is a continuously parametrized family of conjectured optimizers.

III. Analytical Evidence Supporting Conjecture 1

There are several partial results which suggest the validity of Conjecture 1. To this end, note that Conjecture 1 generalizes the following well-known consequence of the conditional entropy power inequality to a longer Markov chain.

Lemma 1 (From [6])**.** *Suppose X, Y are jointly Gaussian, each with unit variance and correlation ρ. For any U satisfying $U - X - Y$, the following inequality holds:*

$$2^{-2I(Y;U)} \ge 1 - \rho^2 + \rho^2 2^{-2I(X;U)}. \tag{26}$$

Proof: Consider any U satisfying $U - X - Y$. Let Y_u, X_u denote the random variables X, Y conditioned on $U = u$. By Markovity and definition of X, Y, we have that $Y_u = \rho X_u + Z$,

[1] This is a consequence of Lemma 1 in Section III.

```
Given $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$, initialize $P_{UVXY} := P^{(0)}_{U|X}P^{(0)}_{V|Y}P_{XY}$
for $i = 1, 2, ...$ do
    $P^{(i)}_{U|X}(u|x) := \frac{\exp\left\{\lambda \int P_{Y|X}(y|x)\log\left(P_{U|Y}(u|y)\right)dy - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(u|v)\right)dv\right\}}{\int \exp\left\{\lambda \int P_{Y|X}(y|x)\log\left(P_{U|Y}(s|y)\right)dy - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(s|v)\right)dv\right\}ds}$   (22)
    $P_{UVXY} \leftarrow P^{(i)}_{U|X}P_{VXY}$   (23)

    $P^{(i)}_{V|Y}(v|y) := \frac{\exp\left\{\lambda \int P_{X|Y}(x|y)\log\left(P_{V|X}(v|x)\right)dx - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(v|u)\right)du\right\}}{\int \exp\left\{\lambda \int P_{X|Y}(x|y)\log\left(P_{V|X}(s|x)\right)dx - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(s|u)\right)du\right\}ds}$   (24)
    $P_{UVXY} \leftarrow P^{(i)}_{V|Y}P_{UXY}$   (25)
```

Algorithm 1: Iterative procedure for solving the Euler-Lagrange equations (35)-(36).

where $Z \sim N(0, 1-\rho^2)$ is independent of X_u. Hence, the conditional entropy power inequality implies that

$$2^{2h(Y|U)} \geq \rho^2 2^{2h(X|U)} + 2\pi e(1-\rho^2) \tag{27}$$

$$= 2\pi e\rho^2 2^{-2I(X;U)} + 2\pi e(1-\rho^2). \tag{28}$$

From here, the lemma easily follows. ■

Lemma 1 can be applied to prove the following special case of Conjecture 1. This result subsumes many special cases that could be analyzed.

Proposition 1. *Suppose X, Y are jointly Gaussian, each with unit variance and correlation ρ. Let U be a random variable for which $X|\{U=u\} \sim N(\mathbb{E}[X|U=u], \sigma^2)$ for all u. If $U - X - Y - V$, then* (1) *holds.*

Proof: Since $X|\{U=u\} \sim N(\mathbb{E}[X|U=u], \sigma^2)$, we have $h(X|U) = h(X|u) = \frac{1}{2}\log(2\pi e\sigma^2)$, and therefore

$$I(X;U) = -\frac{1}{2}\log\sigma^2. \tag{29}$$

By Markovity, it is easy to see that $\mathrm{Var}(Y|U=u) = \rho^2\sigma^2 + (1-\rho^2)$, and hence

$$I(Y;U) = -\frac{1}{2}\log\left(\rho^2\sigma^2 + (1-\rho^2)\right). \tag{30}$$

Let X_u, Y_u, V_u denote the random variables X, Y, V conditioned on $U = u$, respectively. Define $\rho_{XY|u}$ to be the correlation coefficient between X_u and Y_u. It is readily verified that

$$\rho_{XY|u} = \frac{\rho\sigma}{\sqrt{\rho^2\sigma^2 + (1-\rho^2)}}, \tag{31}$$

which does not depend on the particular value of u. By plugging (29)-(31) into (1), we see that (1) is equivalent to

$$2^{-2I(X;V|U)} \geq (1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2I(Y;V|U)}. \tag{32}$$

For every u, X_u, Y_u are jointly Gaussian with correlation coefficient $\rho_{XY|u}$ and $X_u - Y_u - V_u$ form a Markov chain, hence Lemma 1 implies

$$2^{-2I(X_u;V_u)} \geq (1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2I(Y_u;V_u)}. \tag{33}$$

The desired inequality (32) follows by convexity of

$$\log\left[(1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2z}\right] \tag{34}$$

as a function of z. ■

IV. Numerical Evidence Supporting Conjecture 1

Conjecture 2 is amenable to numerical experiments. Dispensing with technicalities in favor of a cleaner exposition, some insight can be gained by deriving the Euler-Lagrange equations corresponding to the functional (14) and attempting to solve them. To this end, the Euler-Lagrange equations are given by:

$$\begin{aligned}&\log P_{U|X}(u|x)\\&= \lambda\int P_{Y|X}(y|x)\log\left(P_{U|Y}(u|y)\right)dy\\&\quad - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(u|v)\right)dv - g(x),\end{aligned} \tag{35}$$

$$\begin{aligned}&\log P_{V|Y}(v|y)\\&= \lambda\int P_{X|Y}(x|y)\log\left(P_{V|X}(v|x)\right)dx\\&\quad - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(v|u)\right)du - h(y),\end{aligned} \tag{36}$$

where the functions $g(x)$ and $h(y)$ serve for the purpose of normalization so that $\int P_{U|X}(u|x)du = 1$ and $\int P_{V|Y}(v|y)dv = 1$, for each x and y, respectively. Note that (35) should hold for all x, u, and (36) should hold for all y, v.

Though characterizing the family of solutions to the nonlinear system of equations given by (35) and (36) may be difficult, it may be possible to compute a particular solution satisfying (35) and (36). In this case, the iterative procedure given by Algorithm 1 is a natural candidate for computing a stationary point. In fact, Algorithm 1 has the desirable property of monotone convergence, which we discuss in the next subsection.

A. Monotone Convergence of Algorithm 1

Let $I^{(i)}(X;U)$, $I^{(i)}(Y;V)$, *etc.* be mutual informations evaluated for the joint distribution $P^{(i)}_{UVXY} = P^{(i)}_{V|Y}P^{(i)}_{U|X}P_{XY}$, and define the corresponding functional:

$$\begin{aligned} F_\lambda(i) \triangleq\ & I^{(i)}(X;U) - \lambda I^{(i)}(Y;U) \\ & + I^{(i)}(Y;V|U) - \lambda I^{(i)}(X;V|U). \end{aligned} \tag{37}$$

Although a proof is omitted due to space constraints, for any $i \geq 1$, we have the inequality

$$\begin{aligned} & F_\lambda(0) - F_\lambda(i) \\ & \geq \sum_{j=1}^{i} \Bigg[D\left(P^{(j-1)}_{U|X} \middle\| P^{(j)}_{U|X}\right) + D\Big(P^{(j)}_{U} \Big\| P^{(j-1)}_{U}\Big) \\ & \qquad + D\left(P^{(j-1)}_{V|Y} \middle\| P^{(j)}_{V|Y}\right) + D\Big(P^{(j)}_{V} \Big\| P^{(j-1)}_{V}\Big) \Bigg]. \end{aligned} \tag{38}$$

Since $F_\lambda(i)$ is bounded from below according to (18), the sum on the right hand side of (38) must converge (assuming the initial test channels $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$ satisfy $F_\lambda(0) < \infty$). In particular, (38) implies that $F_\lambda(i)$ decreases monotonically and converges to some limit, say $F_\lambda(\infty) \triangleq \lim_{i\to\infty} F_\lambda(i)$.

B. Numerical Experiments

Of course, given the infinite-dimensional nature of the problem, it is impractical to implement Algorithm 1 as stated. However, it is a simple matter to quantize the variables U, X, Y, V to a finite number of values. In this case, the integrals in updates (22) and (24) become sums over their respective variables, and (35) and (36) become KKT conditions for the corresponding discretized optimization problem.

The monotone convergence property discussed in the previous section carries over to the discretized variation of Algorithm 1. Therefore, by Pinsker's inequality, there exists a distribution[2]: $Q_{UVXY} = Q_{U|X}Q_{V|Y}P_{XY}$ such that $P^{(i)}_{UVXY} \xrightarrow{TV} Q_{UVXY}$ and hence, by continuity of mutual information, $F_\lambda(i) \searrow I_Q(X;U) - \lambda I_Q(Y;U) + I_Q(Y;V|U) - \lambda I_Q(X;V|U)$, where $I_Q(\cdot;\cdot)$ indicates mutual information evaluated with respect to the distribution Q_{UVXY}. Note that Q_{UVXY} will be a stationary point of the KKT conditions.

The plot shown in Figure 1 is a typical example of the evolution of $F_\lambda(i)$ when running the discretized variation of Algorithm 1. In particular, over thousands of trials with randomly instantiated test channels $P^{(0)}_{U|X}$ and $P^{(0)}_{V|Y}$, $F_\lambda(i)$ has always converged to the conjectured minimum value given by (15). Moreover, this convergence takes place quite rapidly (usually within a few iterations), as exemplified in Figure 1.

The fact that Algorithm 1 converges monotonically, combined with the empirical observation that it converges to the conjectured optimum without exception, suggests that traditional perturbation techniques for proving entropy power inequalities which construct a monotone path from any starting point to a global optimum (see, e.g., [7], [8]) could be adapted to our setting. Unfortunately, despite several attempts, the technical issue of preserving the long Markov chain has proven to be a significant barrier in doing so.

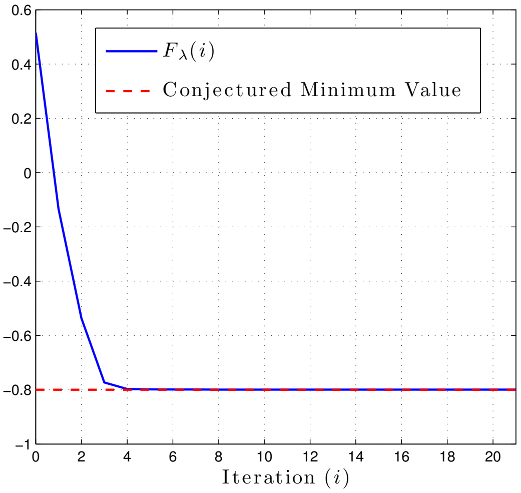


Fig. 1: Evolution of $F_\lambda(i)$ for $\rho = 0.5$, $\lambda = 3/\rho^2$. The variables U, X, Y, V were quantized to 101 evenly spaced values on the interval $[-6, 6]$, and $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$ were randomly instantiated.

V. Concluding Remarks

In summary, Conjecture 1 represents an elegant and natural extension of Lemma 1. Given the widespread use of EPIs in proving converse results, we believe the conjectured extremal inequality (1) could be a useful tool with many applications. As a motivating example, we described an application to the quadratic Gaussian two-encoder source coding problem.

Acknowledgment

The authors gratefully acknowledge conversations with Vignesh Ganapathi-Subramanian, Alon Kipnis, Chandra Nair, and Kartik Venkat which contributed to this work.

References

[1] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[2] T. Courtade, "Information masking and amplification: The source coding setting," in *Proc. of IEEE Int. Symp. on Information Theory (ISIT)*, 2012, pp. 189–193.

[3] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-encoder source-coding problem," *Information Theory, IEEE Transactions on*, vol. 54, no. 5, pp. 1938–1961, 2008.

[4] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint arXiv:1304.6133*, 2013.

[5] M. Raginsky, "Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels," in *Proc. of IEEE Int. Symp. on Information Theory (ISIT)*, 2013, pp. 419–423.

[6] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. on Information Theory*, vol. 43, no. 6, pp. 1912–1923, 1997.

[7] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. on Information Theory*, vol. 11, no. 2, pp. 267–271, 1965.

[8] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information-theoretic problems," *IEEE Trans. on Information Theory*, vol. 53, no. 5, pp. 1839–1851, 2007.

[2] Abusing notation for simplicity, we use P_{XY} to represent the distribution of the jointly Gaussian variables X, Y and their quantized counterparts.

The Likelihood Encoder

Paul Cuff
Department of Electrical Engineering
Princeton University

For source coding we demonstrate the use of a new likelihood encoder by applying it to obtain simple achievability proofs of rate-distortion theory, distributed source coding, and source coding for secrecy. The encoder selects codewords based on likelihoods with respect to a selected joint distribution. As opposed to other traditional encoders used for source coding proofs, this encoder (and proof) does not use an arbitrary threshold, such as the ϵ that defines a jointly-typical set. The likelihood encoder is stochastic. It chooses a codeword proportional to its likelihood. The induced performance is easily analyzed—so much so that the analysis for rate-distortion theory does not require defining error events, multiterminal settings do not require a Markov lemma, and secrecy can be easily analyzed.

As stated above and in [1], the likelihood encoder is defined by a codebook and a joint distribution. Let X^n be the source, with $\mathcal{C} = \{u^n(m)\}$ the codebook and $\bar{P}_{X,U}$ the joint distribution, consistent with the source distribution. The likelihood encoder selects an index in the codebook proportional to the likelihood as follows:

$$P_{M|X^n}(m|x^n) = \frac{\mathcal{L}(m|x^n)}{\sum_{m'} \mathcal{L}(m'|x^n)}$$

where

$$\mathcal{L}(m|x^n) = \prod_{t=1}^{n} \bar{P}_{X|U}(x_t|u_t(m)).$$

The reason for using such an encoder is that it induces a joint distribution P_{X^n,U^n} consistent with a memoryless channel, where $U^n = u^n(m)$ is the reconstruction. That is, let Q_{X^n,U^n} be the distribution created by picking a codeword from the codebook uniformly at random and passing it through a memoryless channel specified by $\bar{P}_{X|U}$. Then, by the construction of the likelihood encoder, $P_{U^n|X^n} = Q_{U^n|X^n}$.

The analysis tool needed to accompany the likelihood encoder is a soft covering lemma. This lemma states that an output distribution of a channel is nearly i.i.d. in total variation if the input is produced by selecting uniformly at random from a codebook of rate greater than the mutual information (in expectation over randomly generated codebooks). This tool appears in the literature first by Wyner [2], was popularized by Han and Verdú as a notion of channel resolvability [3], and is improved upon in [4].

The soft covering lemma allows us to assert that Q_{X^n} is approximately i.i.d. if the rate of the source coding codebook is large enough. For a source distiribution P_{X^n} that is also i.id., we find that P and Q are nearly identical as joint distributions according to total variation. This allows us to replace P with Q for the analysis, and Q is very well behaved.

The details of the rate-distortion proof can be found in [1]. Also, the above steps are the main pieces of the proof for distributed source coding and for secrecy analysis.

I. Acknowledgment

This work is supported by the National Science Foundation through grant CCF-1116013 and by the Air Force Office of Scientific Research through grant FA9550-12-1-0196.

References

[1] E. Song and P. Cuff. “The likelihood encoder for source coding.” *Proc. IEEE Information Theory Workshop,* Seville, Spain, 2013.

[2] A. Wyner. “The common information of two dependent random variables.” *IEEE Trans. Inf. Theory,* 21(2):163-79, 1975.

[3] T. Han and S. Verdú. “Approximation theory of output statistics.” *IEEE Trans. Inf. Theory,* 39(3):752-72, 1993.

[4] P. Cuff. “Distributed Channel Synthesis.” *IEEE Trans. Inf. Theory,* 59(11):7071-96, Nov., 2013.

Universal Decoding for Arbitrary Channels Relative to a Given Class of Decoding Metrics

Neri Merhav
Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

***Abstract*—We consider the problem of universal decoding for arbitrary, finite–alphabet unknown channels in the random coding regime. For a given random coding distribution and a given class of metric decoders, we propose a generic universal decoder whose average error probability is, within a sub–exponential multiplicative factor, no larger than that of the best decoder within this class of decoders. Since the optimal, maximum likelihood (ML) decoder of the underlying channel is not necessarily assumed to belong to the given class of decoders, this setting suggests a common generalized framework for: (i) mismatched decoding, (ii) universal decoding for a given family of channels, and (iii) universal coding and decoding for deterministic channels using the individual–sequence approach. The proof of our universality result is fairly simple, and it is demonstrated how some earlier results on universal decoding are obtained as special cases. We also demonstrate how our method extends to more complicated scenarios, like incorporation of noiseless feedback, the multiple access channel, and continuous alphabet channels.**

I. Introduction

The topic of universal coding and decoding under channel uncertainty has received very much attention in the last four decades. In [5], Goppa offered the *maximum mutual information* (MMI) decoder, which decides in favor of the codeword having the maximum empirical mutual information with the channel output sequence. Goppa showed that for discrete memoryless channels (DMC's), MMI decoding achieves capacity. Csiszár and Körner [2] have shown that the random coding error exponent of the MMI decoder achieves the optimum random coding error exponent. Csiszár [1] proved that for any modulo–additive DMC and the uniform random coding distribution over linear codes, the optimum random coding error exponent is universally achieved by a decoder that minimizes the empirical entropy of the difference between the output sequence and the input sequence. In [10] an analogous result was derived for a certain parametric class of memoryless Gaussian channels with an unknown interference signal.

For channels with memory, Ziv [16] studied universal decoding problem for finite–alphabe unifilar finite–state channels. For uniform random coding over a given set, he proved that a decoder based on the Lempel–Ziv algorithm asymptotically achieves the error exponent associated with ML decoding. In [6], Lapidoth and Ziv have extended this result to non–unifilar finite–state channels. In [3], Feder and Lapidoth furnished sufficient conditions for families of channels with memory to have universal decoders in the random coding error exponent sense. In [4], Feder and Merhav proposed a competitive minimax criterion, according to which, an optimum decoder is sought in the quest for minimizing the worst–case regret in the error exponent sense.

More recently, interesting attempts (see, e.g., [8], [9], [12], [14]) were made to devise coding and decoding strategies that avoid any probabilistic assumptions concerning the channel. In [8], the notion of *empirical rate functions* has been established and investigated for a given input distribution and a given posterior probability function of the channel input sequence given the output sequence. In [12], porosity–achieving universal encoders and decoders were devised for modulo additive channels with deterministic noise sequences and noiseless feedback.

In this paper, we take a different approach. We consider universal decoding for *arbitrary* unknown channels in the random coding regime. For a given random coding distribution and a given class of metric decoders, we propose a generic universal decoder whose average error probability is exponentially no larger than that of the best decoder in this class. The proof of our universality result is fairly simple and general, and it is demonstrated how some earlier mentioned results on universal decoding are obtained as special cases.

Finally, we demonstrate how our method extends to more complicated scenarios. The first extension corresponds to incorporation of noiseless feedback. This extension is fairly straightforward, but its main importance is in allowing adaptation of the random coding distribution to the channel statistical characteristics. The second extension is to the problem of universal decoding for multiple access channels (MAC's) with respect to a given class of decoding metrics. This extension is not trivial since the universal decoding metric has to confront three different types of error events. In particular, it turns out that the resulting universal decoding metric is surprisingly different from those of earlier works on universal decoding for the MAC [7], [3, Section VIII], [13], mostly because the problem setting here is different from those of these earlier works (in the sense that the universality here is relative to a given class of decoders while the underlying channel is arbitrary, and not relative to a given class of channels). A third possible extension, that refers to the continuous alphabet case, is discussed briefly along with an example.

II. Notation Conventions

Scalar random variables (RV's) will be denoted by capital letters, their sample values are denoted by the respective lower case letters, and their alphabets are denoted by the respective calligraphic letters. A similar convention applies to random vectors of dimension n and their sample values, which will be denoted with the same symbols in the bold face font. The set of all n–vectors with components taking values in a certain alphabet, will be denoted as the same alphabet superscripted by n. Channels and sources will be denoted generically by the letter P and Q, respectively. For example, the channel input probability distribution function will be denoted by $Q(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^n$, and the conditional probability distribution of the channel output vector $\boldsymbol{y} \in \mathcal{Y}^n$ given the input vector $\boldsymbol{x} \in \mathcal{X}^n$, will be denoted by $P(\boldsymbol{y}|\boldsymbol{x})$. Information theoretic quantities like entropies and conditional entropies, will be denoted following the standard conventions of the information theory literature, e.g., $H(\boldsymbol{X})$, $H(\boldsymbol{X}|\boldsymbol{Y})$, etc. The expectation operator will be denoted by $\boldsymbol{E}\{\cdot\}$ and the cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

For a given sequence $\boldsymbol{x} \in \mathcal{X}^n$, $\mathcal{X}$ being a finite alphabet, $\hat{P}_{\boldsymbol{x}}$ denotes the empirical distribution on $\mathcal{X}$ extracted from $\boldsymbol{x}$, in other words, $\hat{P}_{\boldsymbol{x}}$ is the vector $\{\hat{P}_{\boldsymbol{x}}(x),\ x \in \mathcal{X}\}$, where $\hat{P}_{\boldsymbol{x}}(x)$ is the relative frequency of the letter x in the vector $\boldsymbol{x}$. The type class of $\boldsymbol{x}$, denoted $T_{\boldsymbol{x}}$, is the set of all sequences $\boldsymbol{x}' \in \mathcal{X}^n$ with $\hat{P}_{\boldsymbol{x}'} = \hat{P}_{\boldsymbol{x}}$. Similarly, for a pair of sequences $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$, the empirical distribution $\hat{P}_{\boldsymbol{xy}}$ is the matrix of relative frequencies $\{\hat{P}_{\boldsymbol{xy}}(x,y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ and the type class $T_{\boldsymbol{xy}}$ is the set of pairs $(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{X}^n \times \mathcal{Y}^n$ with $\hat{P}_{\boldsymbol{x}'\boldsymbol{y}'} = \hat{P}_{\boldsymbol{xy}}$. For a given $\boldsymbol{y}$, $T_{\boldsymbol{x}|\boldsymbol{y}}$ denotes the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, which is the set of vectors $\{\boldsymbol{x}'\}$ such that $(\boldsymbol{x}', \boldsymbol{y}) \in T_{\boldsymbol{xy}}$. Information measures induced by empirical distributions, i.e., empirical information measures, will be denoted with a hat and a subscript that indicates the sequence(s) from which they are induced. For example, $\hat{H}_{\boldsymbol{x}}(X)$ is the empirical entropy extracted from $\boldsymbol{x} \in \mathcal{X}^n$, namely, the entropy of a random variable X whose distribution is $\hat{P}_{\boldsymbol{x}}$. Similarly, $\hat{H}_{\boldsymbol{xy}}(X|Y)$ and $\hat{I}_{\boldsymbol{xy}}(X;Y)$ are, respectively, the empirical conditional entropy of X given Y, and the empirical mutual information between X and Y, extracted from $(\boldsymbol{x}, \boldsymbol{y})$, and so on. For two sequences of positive numbers, $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Similarly, $a_n \stackrel{\cdot}{\le} b_n$ means that $\limsup_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} \le 0$, and so on. The functions $\log(\cdot)$ and $\exp(\cdot)$, throughout this paper, will be defined to the base 2, unless otherwise indicated. The operation $[\cdot]_+$ will mean positive clipping, that is $[x]_+ = \max\{0, x\}$.

III. Problem Formulation

Consider a random selection of a codebook $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subseteq \mathcal{X}^n$, where $M = 2^{nR}$, R being the coding rate in bits per channel use. The marginal probability distribution function of each codeword $\boldsymbol{x}_i$ is denoted by $Q(\boldsymbol{x}_i)$. It will be assumed that the various codewords are *conditionally pairwise independent.*[1] Let $P(\boldsymbol{y}|\boldsymbol{x})$ be the conditional probability distribution of the channel output vector $\boldsymbol{y} \in \mathcal{Y}^n$ given the channel input vector $\boldsymbol{x} \in \mathcal{X}^n$. We make no assumptions at all concerning the channel. We will assume that both the channel input alphabet $\mathcal{X}$ and the channel output alphabet $\mathcal{Y}$ are finite. Finally, we define a class of *decoding metrics*, as a class of real functions, $\mathcal{M} = \{m_\theta(\boldsymbol{x}, \boldsymbol{y}),\ \theta \in \Theta,\ \boldsymbol{x} \in \mathcal{X}^n,\ \boldsymbol{y} \in \mathcal{Y}^n\}$, where Θ is an index set. The decoder associated with m_θ, which will be denoted by $\mathcal{D}_\theta$, decides in favor of the message $i \in \{1, \ldots, M\}$ which maximizes $m_\theta(\boldsymbol{x}_i, \boldsymbol{y})$. The message i is assumed to be uniformly distributed over $\{1, 2, \ldots, M\}$. It should be emphasized that the ML decoding metric for the underlying channel $P(\boldsymbol{y}|\boldsymbol{x})$, may not necessarily belong to $\mathcal{M}$. The average error probability associated with $\mathcal{D}_\theta$, is denoted by $\bar{P}_{e,\theta}(R,n)$. While the decoder $\mathcal{D}_\theta$, that minimizes $\bar{P}_{e,\theta}(R,n)$ within $\mathcal{M}$, depends on the unknown underlying channel, our goal is to devise a universal decoder $\mathcal{U}$, with a decoding metric $U(\boldsymbol{x}, \boldsymbol{y})$, independent of the underlying channel $P(\boldsymbol{y}|\boldsymbol{x})$, whose average error probability would be essentially as small as $\min_\theta \bar{P}_{e,\theta}(R,n)$, whatever the underlying channel may be. By “essentially as small”, we mean that the average error probability associated with the universal decoder, denoted $\bar{P}_{e,u}(R,n)$, would not exceed $\min_\theta \bar{P}_{e,\theta}(R,n)$ in the exponential sense.

IV. Main Result

Let us define

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) \stackrel{\triangle}{=} \{\boldsymbol{x}' :\ \ \forall\theta \in \Theta\ \ m_\theta(\boldsymbol{x}', \boldsymbol{y}) = m_\theta(\boldsymbol{x}, \boldsymbol{y})\}\,. \tag{1}$$

Our universal decoding metric is defined as

$$U(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\triangle}{=} -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]. \tag{2}$$

Clearly, $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ are equivalence classes for every $\boldsymbol{y} \in \mathcal{Y}^n$, and so $\mathcal{X}^n$ can be partitioned into a disjoint union of them. Let $K_n(\boldsymbol{y})$ denote the number of equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ for a given $\boldsymbol{y}$. Also define $K_n \stackrel{\triangle}{=} \max_{\boldsymbol{y}\in\mathcal{Y}^n} K_n(\boldsymbol{y})$ and $\Delta_n \stackrel{\triangle}{=} \frac{\log K_n}{n}$. Our main result is the following theorem (the proof appears in the full version on the paper [11]):

Theorem 1: Under the above assumptions, the universal decoding metric defined in eq. (2) satisfies:

$$\bar{P}_{e,u}(R,n) \le 2 \cdot 2^{n\Delta_n} \cdot \min_{\theta\in\Theta} \bar{P}_{e,\theta}(R,n). \tag{3}$$

The theorem is meaningful when $\Delta_n \to 0$ as $n \to \infty$, which means that the number of various equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ grows sub-exponentially, uniformly in $\boldsymbol{y}$. In this case, whenever $\min_{\theta\in\Theta} \bar{P}_{e,\theta}(R,n)$ decays exponentially with n, then so does $\bar{P}_{e,u}(R,n)$, and at least as fast. Therefore, a sufficient condition for the existence of a universal decoder is $\lim_{n\to\infty} \Delta_n = 0$. The behavior of Δ_n for large n is a measure of the richness of $\mathcal{M}$. The larger is $\mathcal{M}$, the smaller are the equivalence classes, and then their total number becomes larger, and so does Δ_n. Universality is enabled, using this method, as long as the set Θ is not too rich, so that Δ_n still vanishes as n grows. When Q is invariant within $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ (i.e., $\boldsymbol{x}' \in \mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ implies $Q(\boldsymbol{x}') = Q(\boldsymbol{x})$), we have $U(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}) + \log|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|]$. The choice of Q that is invariant within $T(\boldsymbol{x}|\boldsymbol{y})$ is convenient, because it is easier to evaluate the log–cardinality of $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ than to evaluate its probability under Q.

Example 1. Let Q be the uniform distribution across a single type class, $T_{\boldsymbol{x}}$, and let $\mathcal{M}$ be the class of additive decoding metrics $m_\theta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n \theta(x_i, y_i)$, where $\{\theta(x, y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ are arbitrary real–valued matrices. Here, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) =$

[1] By “conditionally pairwise independent”, we mean that for every three randomly chosen codewords, we have that any two of them are conditionally independent given the third one.

$T_{\boldsymbol{x}|\boldsymbol{y}}$, the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$. Since the number of conditional type classes is polynomial in n, then $\Delta_n \to 0$. In this case, $U(\boldsymbol{x},\boldsymbol{y}) = \hat{I}_{\boldsymbol{xy}}(X;Y) + o(n)$. and so, the proposed universal decoder essentially coincides with the MMI decoder. If, on the other hand, $Q(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$, then $U(\boldsymbol{x},\boldsymbol{y}) = \hat{I}_{\boldsymbol{xy}}(X;Y) + D(\hat{P}_{\boldsymbol{x}}\|Q) + o(n)$, where $D(\hat{P}_{\boldsymbol{x}}\|Q)$ is the divergence between $\hat{P}_{\boldsymbol{x}}$ and Q. This concludes Example 1.

One of the elegant points in [16] is that the universality of the proposed decoding metric is proved without recourse to an explicit derivation of the random coding error exponent of the optimum decoder. The proof of Theorem 1 has the same feature. However, thanks to Shulman's lower bound [15] on the probability of a union of events, this proof is both simpler and more general in several respects: (i) it allows a general Q, not just the uniform distribution, (ii) it requires only conditionally pairwise independence between codewords, not mutual independences, and (iii) it assumes nothing concerning the underlying channel. Indeed, it will be seen shortly how Ziv's universal decoding metric is obtained as a special case.

In some situations, it may not be a trivial task to evaluate $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]$. Suppose, however, that one can uniformly lower bound $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})] = \exp\{-nU(\boldsymbol{x},\boldsymbol{y})\}$ by $\exp\{-nU'(\boldsymbol{x},\boldsymbol{y})\}$, for some function $U'(\boldsymbol{x},\boldsymbol{y})$ which is computable and suppose that $U'(\cdot,\cdot)$ satisfies

$$\max_{\boldsymbol{y}\in\mathcal{Y}^n} \sum_{\boldsymbol{x}\in\mathcal{X}^n} Q(\boldsymbol{x})2^{nU'(\boldsymbol{x},\boldsymbol{y})} \le 2^{n\Delta_n'} \tag{4}$$

where $\Delta_n' \to 0$. We argue (see [11] for a proof) that in such a case, $U'(\cdot,\cdot)$ can replace $U(\cdot,\cdot)$ as a universal decoding metric and Theorem 1 remains valid. The price of passing from U to U' might be in a slowdown of the convergence of Δ_n' vs. Δ_n. For example, U' might correspond to more refined equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$.

Example 2. As an example of the usefulness of this observation, let us refer to Ziv's universal decoding metric [16]. In particular, let $\mathcal{M}$ be the class of decoding metrics defined as follows: For a given $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$, let $\boldsymbol{s} = (s_1,\ldots,s_n) \in \mathcal{S}^n$ ($\mathcal{S}$ being a finite set), be a sequence generated recursively according to $s_{i+1} = g(x_i, y_i, s_i)$, $i = 1,\ldots,n-1$, where s_1 is a fixed initial state and $g : \mathcal{X}\times\mathcal{Y}\times\mathcal{S} \to \mathcal{S}$ is a given next–state function. Now define $m_\theta(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n \theta(x_i,y_i,s_i)$. Suppose that Q is the uniform distribution over $\mathcal{X}^n$. Then $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]$ is proportional to $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$, but the problem is that here, there is no apparent single–letter expression for the exponential growth rate of $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$ in general. Fortunately enough, however, $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$, in this case, can be lower bounded [16, Lemma 1] by $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})| \ge 2^{LZ(\boldsymbol{x}|\boldsymbol{y})-no(n)}$, where $LZ(\boldsymbol{x}|\boldsymbol{y})$ denotes the length of the conditional Lempel–Ziv code of $\boldsymbol{x}$ when $\boldsymbol{y}$ is given as side information at both encoder and decoder. Consequently, one can upper bound $U(\boldsymbol{x},\boldsymbol{y})$ by $U'(\boldsymbol{x},\boldsymbol{y}) = \log|\mathcal{X}| - \frac{LZ(\boldsymbol{x}|\boldsymbol{y})}{n} + o(n)$ as our decoding metric. Indeed, eq. (4) is satisfied by this choice of U'. This explains why Ziv's decoder, which selects the message i with the minimum of $LZ(\boldsymbol{x}_i|\boldsymbol{y})$, is universally asymptotically optimum in the random coding exponent sense. Note that the assumption that Q is uniform is not really essential here. In fact, Q can also be any exchangeable probability distribution. Moreover, if s_i includes a component, say, σ_i, that is fed merely by $\{x_i\}$ (but not $\{y_i\}$), then it is enough that Q would be invariant within conditional types of $\boldsymbol{x}$ given $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_n)$. In such a case, we would have $U'(\boldsymbol{x}.\boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}) + LZ(\boldsymbol{x}|\boldsymbol{y})]$.

V. Extensions

We now demonstrate how our method extends to more involved scenarios.

A. Feedback

In the paradigm of random coding in the presence of feedback, it is convenient to think of an independent random selection of symbols of $\mathcal{X}$ along a tree whose branches are labeled by $\{y_1\},\{y_1,y_2\},\ldots,\{y_1,\ldots,y_{n-1}\}$, for all possible outcomes of these vectors. Accordingly, the random coding distribution $Q(\boldsymbol{x})$ is replaced by $Q(\boldsymbol{x}|\boldsymbol{y}) \triangleq \prod_{i=1}^n Q(x_i|x^{i-1},y^{i-1})$. Each message $i \in \{1,2,\ldots,M\}$ is represented by a complete tree of depth n and $|\mathcal{Y}|^{n-1}$ leaves. Theorem 1 and its proof remain intact with $Q(\cdot)$ being replaced by $Q(\cdot|\boldsymbol{y})$ in all places. Thus, the universal decoding metric is redefined as $U(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|\boldsymbol{y}]$, the relevant expectations are redefined w.r.t. $P(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^n [Q(x_i|x^{i-1},y^{i-1})P(y_i|x^i,y^{i-1})]$, and in condition (4), $Q(\boldsymbol{x})$ is replaced by $Q(\boldsymbol{x}|\boldsymbol{y})$. One might limit the structure of the feedback, for example, by letting each $Q(\cdot|x^{i-1},y^{i-1})$ depend on (x^{i-1},y^{i-1}) only via a state variable t_i fed by these two sequences, i.e., $t_i = g(t_{i-1},x_{i-1},y_{i-1})$, that is

$$Q(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^n Q(x_i|x^{i-1},y^{i-1}) = \prod_{i=1}^n Q(x_i|t_i). \tag{5}$$

In the above example of decoding metrics corresponding to finite–state channels, one can refine the equivalence classes to include the information about t_i and then Q would be invariant within a type class $T_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{t}}$, where $\boldsymbol{t} = (t_1,\ldots,t_n)$. In this case, the decoding metric U' would become $U'(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}|\boldsymbol{y}) + LZ(\boldsymbol{x}|\boldsymbol{y})]$, where $Q(\boldsymbol{x}|\boldsymbol{y})$ is understood to be defined by (5).

B. The Multiple Access Channel

Consider a MAC with two inputs, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and one output $\boldsymbol{y}$. User no. i generates $M_i = 2^{nR_i}$ mutually independent codewords, $\boldsymbol{x}_i(1),\ldots,\boldsymbol{x}_i(M_i)$, using a random coding distribution Q_i, $i = 1,2$. We define a class $\mathcal{M} = \{m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}),\ \theta \in \Theta\}$. Decoder $\mathcal{D}_\theta$ picks the pair of messages $(\boldsymbol{x}_1(i),\boldsymbol{x}_2(j))$, maximizes $m_\theta(\boldsymbol{x}_1(i),\boldsymbol{x}_2(j),\boldsymbol{y})$. We assume that the random coding ensemble and the class of decoders are such that for every given $\boldsymbol{X}_1(i) = \boldsymbol{x}_1(i), \boldsymbol{X}_2(j) = \boldsymbol{x}_2(j)$ and $\boldsymbol{Y} = \boldsymbol{y}$, $m_\theta(\boldsymbol{X}_1(i'),\boldsymbol{X}_2(j'),\boldsymbol{y})$ and $m_\theta(\boldsymbol{X}_1(i''),\boldsymbol{X}_2(j''),\boldsymbol{y})$ are conditionally independent whenever $(i',j') \ne (i,j)$, $(i'',j'') \ne (i,j)$ and $(i'',j'') \ne (i',j')$. While this requirement is easily satisfied when $i' \ne i$, $i'' \ne i$, $i'' \ne i'$, $j' \ne j$, $j'' \ne j$, and $j'' \ne j'$ all hold (as all codewords are assumed to be drawn by independent random selection), it is less obvious when some of these indices concide. Still, this requirement is satisfied, for example, if $\mathcal{X}_1 = \mathcal{X}_2 = \{0,1,\ldots,K-1\}$ Q_1 and Q_2 are both uniform across the alphabet, and $m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ depends on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ only via $\boldsymbol{x}_1 \oplus \boldsymbol{x}_2$, where $\oplus$ denotes addition modulo K. Decoding metrics with this property are motivated by classes of multiple access channels, $P(\boldsymbol{y}|\boldsymbol{x}_1,\boldsymbol{x}_2)$, in which the users interfere with each other additively, i.e., $P(\boldsymbol{y}|\boldsymbol{x}_1,\boldsymbol{x}_2)=W(\boldsymbol{y}|\boldsymbol{x}_1\oplus\boldsymbol{x}_2)$. Still, the dependence of $\boldsymbol{y}$ on $\boldsymbol{x}_1\oplus\boldsymbol{x}_2$ can be arbitrary.

We now define three kinds of equivalence classes: $\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})$ is the set of $(\boldsymbol{x}_1',\boldsymbol{x}_2')$ such that $\forall\theta\in\Theta$, $m_\theta(\boldsymbol{x}_1',\boldsymbol{x}_2',\boldsymbol{y})=m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$, $\mathcal{T}(\boldsymbol{x}_1|\boldsymbol{x}_2,\boldsymbol{y})$ is the set of $\boldsymbol{x}_1'$ such that $\forall\theta\in\Theta$, $m_\theta(\boldsymbol{x}_1',\boldsymbol{x}_2,\boldsymbol{y})==m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ and $\mathcal{T}(\boldsymbol{x}_2|\boldsymbol{x}_1,\boldsymbol{y})$ is defined similarly with '1' and '2' swapped. We also assume, as before, that for every $\boldsymbol{y}$, the number of different type classes $\{\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})\}$ is upper bounded by $2^{n\Delta_n}$. Next, define the following functions:

$$U_0(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log\{(Q_1\times Q_2)[\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})]\} \quad (6)$$

$$U_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log Q_1[\mathcal{T}(\boldsymbol{x}_1|\boldsymbol{x}_2,\boldsymbol{y})] \quad (7)$$

$$U_2(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log Q_2[\mathcal{T}(\boldsymbol{x}_2|\boldsymbol{x}_1,\boldsymbol{y})]. \quad (8)$$

Define the universal decoding metric $U(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ as the minimum among the following three expressions: $U_0(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_1-R_2$, $U_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_1$, and $U_2(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_2$. We argue that $U(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ competes favorably with the best m_θ in a sense analogous to that asserted in Theorem 1. This decoding metric is different from the universal decoding metrics used for the MAC, for example, in [13] and [7], which were based on the MMI decoder and the minimum empirical conditional entropy (minimum equivocation) rule, respectively. Similarly as before, suppose that U_0, U_1 and U_2 can be uniformly upper bounded by U_0', U_1' and U_2', respectively, and assume that:

$$\max_{\boldsymbol{y}}\sum_{\boldsymbol{x}_1,\boldsymbol{x}_2}Q_1(\boldsymbol{x}_1)Q_2(\boldsymbol{x}_2)2^{nU_0'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \overset{\cdot}{\le} 1 \quad (9)$$

$$\max_{\boldsymbol{x}_2,\boldsymbol{y}}\sum_{\boldsymbol{x}_1}Q_1(\boldsymbol{x}_1)2^{nU_1'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \overset{\cdot}{\le} 1 \quad (10)$$

$$\max_{\boldsymbol{x}_1,\boldsymbol{y}}\sum_{\boldsymbol{x}_2}Q_2(\boldsymbol{x}_1)2^{nU_2'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \overset{\cdot}{\le} 1. \quad (11)$$

Then, U_0', U_1' and U_2' can replace U_0, U_1 and U_2, respectively, in the universal decoding metric, denoted in turn by U', and the upper and lower bounds continue to hold with U' replacing U. The application of this to the LZ decoding metric is a straightforward extension to the one exercised above in the single–user case (see [11] for details).

C. Comments on the Continuous Alphabet Case

It is possible to extend Theorem 1 to the case of continuous alphabets, but this requires more caution. For one thing, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ should be redefined by allowing some small tolerance, i.e., the requirement $m_\theta(\boldsymbol{x},\boldsymbol{y})=m_\theta(\boldsymbol{x},\boldsymbol{y})$ should be replaced by $|m_\theta(\boldsymbol{x},\boldsymbol{y})-m_\theta(\boldsymbol{x},\boldsymbol{y})|\le\epsilon$, where $\epsilon>0$ tends to zero after $n\to\infty$. This is to guarantee that $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ captures a positive volume and that $K_n(\boldsymbol{y})$ (now, redefined as the number of $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ required to cover the set of channel input vectors, possibly obeying an input constraint) is finite. We will not delve into the technical details of this extension any further[2] Instead, we will merely demonstrate the universal decoding metric in a certain special case, where the class of decoding metrics depend on $\boldsymbol{x}$ and $\boldsymbol{y}$ only via second order empirical statistics extracted from these sequences.

Example 3. Let $\mathcal{X}=\mathcal{Y}=\mathrm{I\!R}$ and let Q be an i.i.d. zero–mean Gaussian density with variance σ^2. Let $\theta=(\theta_1,\theta_2)$ and let $\mathcal{M}$ be the class of decoding metrics $m_\theta(\boldsymbol{x},\boldsymbol{y})=\theta_1\sum_{i=1}^n x_iy_i+\theta_2\sum_{i=1}^n x_i^2$. Denoting $C(\boldsymbol{x},\boldsymbol{y})=\frac{1}{n}\sum_{i=1}^n x_iy_i$ and $S(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^n x_i^2$, then $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ should be redefined as the set of $\boldsymbol{x}'$, where $C(\boldsymbol{x}',\boldsymbol{y})$ and $S(\boldsymbol{x}')$ are within ϵ close to $C(\boldsymbol{x},\boldsymbol{y})$ and $S(\boldsymbol{x})$, respectively. Using the methods of [10], it is easy to show that $U(\boldsymbol{x},\boldsymbol{y})=\frac{S(\boldsymbol{x})}{2\sigma^2}-\frac{1}{2}\ln[S(\boldsymbol{x})(1-\rho_{\boldsymbol{xy}}^2)]$, where $\rho_{\boldsymbol{xy}}=C(\boldsymbol{x},\boldsymbol{y})/\sqrt{S(\boldsymbol{x})S(\boldsymbol{y})}$ is the empirical correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$, and where we have used natural logarithms instead of base 2 logarithms.

References

[1] I. Csiszár, "Linear codes for sources and source networks: error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 4, pp. 585–592, July 1982.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[3] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1726–1745, September 1998.

[4] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, vol. 48, no. 6, pp. 1504–1517, June 2002.

[5] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Cont. Information Theory*, vol. 4, pp. 97–102, 1975.

[6] A. Lapidoth and J. Ziv, "On the universality of the LZ–based noisy channels decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1746–1755, September 1998.

[7] Y.-S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 376–386, March 1996.

[8] Y. Lomnitz and M. Feder, "Communication over individual channels – a general framework," arXiv:1023.1406v1 [cs.IT] 7 Mar 2012.

[9] Y. Lomnitz and M. Feder, "Universal communication over modulo–additive channels with an individual noise sequence," arXiv:1012.2751v2 [cs.IT] 7 May 2012.

[10] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

[11] N. Merhav, "Universal decoding for arbitrary channels relative to a given family of decoding metrics," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5566–5576, September 2013.

[12] V. Misra and T. Weissman, "The porosity of additive noise sequences," arXiv:1025.6974v1 [cs.IT] 31 May 2012.

[13] J. Pokorny and H. M. Wallmeier, "Random coding bound and codes produced by permutations for the multiple access channel," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 6, pp. 741–750, November 1985.

[14] O. Shayevitz and M. Feder, "Communicating using feedback over a binary channel with arbitrary noise sequence," *Proc. ISIT 2005*, pp. 1516–1520, Adelaide, Australia, September 2005.

[15] N. Shulman, *Communication over an Unknown Channel via Common Broadcasting*, Ph.D. dissertation, Department of Electrical Engineering – Systems, Tel Aviv University, July 2003.
http://www.eng.tau.ac.il/∼shulman/papers/Nadav_PhD.pdf

[16] J. Ziv, "Universal decoding for finite–state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.

[2] See [10] where these details have been fully worked out in the context of universal decoding for the Gaussian channel with a deterministic interference.

Equivalent Formulations of Hypercontractivity Using Information Measures

Extended Abstract

Chandra Nair

A pair of random variables (X, Y) defined on some probability space $(\Omega, \mathcal{F}, \mu)$, is said to be (p, q)-hypercontractive for $1 \leq \mathsf{q} \leq \mathsf{p} < \infty$ if the inequality $\| \operatorname{E}[g(Y)|X]\|_{\mathsf{p}} \leq \|g(Y)\|_{\mathsf{q}}$ holds for every bounded measurable function $g(Y)$. For any $\mathsf{p} \geq 1$ one can define $\mathsf{q}_{\mathsf{p}}(X;Y) = \inf\{\mathsf{q} : (X, Y) \text{ is } (\mathsf{p}, \mathsf{q})\text{-hypercontractive}\}$. Define the ratio $r_{\mathsf{p}}(X;Y) = \frac{\mathsf{q}_{\mathsf{p}}(X;Y)}{\mathsf{p}}$. Estimating $r_{\mathsf{p}}(X;Y)$ leads to the best hypercontractive inequality for a given p.

Hypercontractive inequalities have found a variety of applications in quantum physics[3], theoretical computer science[4], analysis[5], and in information theory[1, 2]. In this talk we present the following alternate characterizations of $r_{\mathsf{p}}(X;Y)$ using information measures.

Theorem 1. *The hypercontractive ratio $r_{\mathsf{p}}(X;Y)$ is also given by any of the following expressions*

(a)

$$\sup_{\nu_{X,Y} \ll \mu_{X,Y}} \frac{D_{KL}(\nu_Y || \mu_Y)}{\mathsf{p} D_{KL}(\nu_{X,Y} || \mu_{X,Y}) - (\mathsf{p}-1) D_{KL}(\nu_X || \mu_X)}$$

(b)

$$\sup_U \frac{I(U;Y)}{\mathsf{p} I(U;X,Y) - (\mathsf{p}-1) I(U;X)}$$

(c)

$$\inf\{\lambda : H(Y) - \lambda \mathsf{p} H(X,Y) + \lambda(\mathsf{p}-1) H(X) = \mathfrak{K}[H(Y) - \lambda \mathsf{p} H(X,Y) + \lambda(\mathsf{p}-1)H(X)]_\mu\},$$

where $\mathfrak{K}[f(\cdot)]_\mu$ denotes the lower convex envelope of the function $f(\cdot)$ (over joint distributions) evaluated at the joint distribution $\mu(X, Y)$

Remark: The above result generalizes the result equivalence result in both [1] and [2] which deal with the limiting case $\mathsf{p} \to \infty$.

References

[1] Rudolf Ahlswede and Peter Gács, Spreading of sets in product spaces and hypercontraction of the markov operator, The Annals of Probability (1976), 925–939.

[2] Venkat Anantharam, Amin Aminzadeh Gohari, Sudeep Kamath, and Chandra Nair, On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover, CoRR **abs/1304.6133** (2013).

[3] E. Brian Davies, Leonard Gross, and Barry Simon, Hypercontractivity: a bibliographic review, Ideas and methods in quantum and statistical physics (Oslo, 1988), Cambridge Univ. Press, Cambridge, 1992, pp. 370–389. MR 1190534 (93g:47052)

[4] Jeff Kahn, G. Kalai, and Nathan Linial, The influence of variables on boolean functions, Foundations of Computer Science, 1988., 29th Annual Symposium on, 1988, pp. 68–80.

[5] Michel Talagrand, On russo's approximate zero-one law, The Annals of Probability **22** (1994), no. 3, pp. 1576–1587 (English).

International Zurich Seminar on Communications (IZS), February 26 – 28, 2014

Capacity of Binary Symmetric POST Channels

Haim H. Permuter
Ben Gurion University
haimp@bgu.ac.il

Himanshu Asnani
Stanford University
asnani@stanford.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—We consider finite state channels where the state of the channel is its previous output. We refer to these as POST (Previous Output is the STate) channels. We focus on POST(a,b) channels. These channels have binary inputs and outputs, where the state determines if the channel behaves as a binary with parameters (a, b) or (b, a). We show that the non feedback capacity of the POST(a,b) channel equals its feedback capacity, despite the memory of the channel. The proof of this surprising result is based on showing that the induced output distribution, when maximizing the directed information in the presence of feedback, can also be achieved by an input distribution that does not utilize of the feedback. We show that this is a sufficient condition for the feedback capacity to equal the non feedback capacity for any finite state channel.**

Keywords*—Causal conditioning, Convex optimization, Channels with memory, Directed information, Feedback capacity, Finite state channel, KKT conditions, POST channel.

I. Introduction

The capacity of a memoryless channel is very well understood. There are many simple memoryless channels for which we know the capacity analytically. These include the binary symmetric channel, the erasure channel, the additive Gaussian channel and the Z Channel. Furthermore, using convex optimization tools, such as the Blahut-Arimoto algorithm [1], [2], we can efficiently compute the capacity of any memoryless channel with a finite alphabet. However, in the case of channels with memory, the exact capacities are known for only a few channels, such as additive Gaussian channels (water filling solution) [3], [4] and discrete additive channels with memory [5]. In cases where feedback is allowed, there are only a few more cases where the exact capacity is known, such as the modulo-additive noise channel, the additive noise channel where the noise is a first-order autoregressive moving-average Gaussian process [6], the trapdoor channel [7], and the Ising Channel [8]. If the state is known at the decoder, then knowledge of the state at the encoder can be considered as partial feedback, as considered and solved in [9] and in [10].

In this paper we introduce and consider a new family of channels that we refer to as “POST channels”. These are simple Finite State Channels (FSCs) where the state of the channel is the previous output. In particular, we focus on a family of POST channels that have binary inputs $\{X_i\}_{i\geq 1}$ and binary outputs $\{Y_i\}_{i\geq 1}$ related as follows: Consider the POST channel depicted in Fig. 1 with the following behavior. When $y_{i-1} = 0$, then the channel behaves as a binary channel with transition matrix

$$\begin{bmatrix} a & \bar{b} \\ \bar{a} & b \end{bmatrix} \qquad (1)$$

and when $y_{i-1} = 1$ then it behaves as a binary channel with the transition matrix

$$\begin{bmatrix} b & \bar{a} \\ \bar{b} & a \end{bmatrix}. \qquad (2)$$

We refer to this channel as the POST(a,b) channel. The POST(α) which was considered in [11] is a special case of POST(a,b), where $a = 1$ and $b = \bar{\alpha}$. The results in this paper extends our previouse results in [11]. An extended version of this conference paper that includes all the proofs may be found in [12].

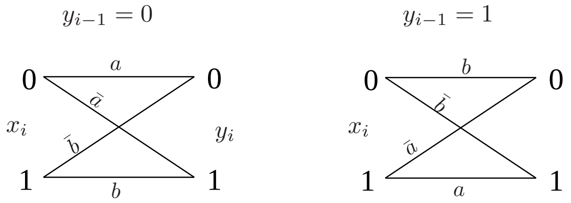


Fig. 1. POST(a,b) channel. If $y_{i-1} = 0$ then the channel behaves as DMC with parameters (a, b) and if $y_{i-1} = 1$ then the channel behaves as DMC with parameters (b, a).

Without loss of generality, we assume throughout that $a + b - 1 > 0$. It is easy to see that in the case where $a + b - 1 = 0$ or, equivalently, where $a = \bar{b}$, the capacity is simply 0. Additionally, if $a + b - 1 < 0$ then $\bar{a} + \bar{b} > 1$; hence by relabeling the inputs $(0 \leftrightarrow 1)$ we obtain a new channel (with parameter a', b' rather than a, b) where $a' = \bar{a}$ and $b' = \bar{b}$ and we have $a' + b' - 1 > 0$.

This channel arose in the investigation of controlled feedback in the setting of “to feed or not to feed back” [13]. The POST channel can also be useful in modeling memory affected by past channel outputs, as is the case in flash memory and other storage devices.

In order to prove that feedback does not increase the capacity of some families of POST channels, we look at two convex optimization problems: maximizing the directed information over regular input distributions (non feedback case), i.e., $P(x^n)$ and, secondly, over causal conditioning that is influenced by the feedback i.e., $P(x^n||y^{n-1})$. We show that a necessary and sufficient condition for the solutions of the two optimization problems to achieve the same value is that the induced output distributions $P(y^n)$ by the respective optimal values $P^*(x^n)$ and $P^*(x^n||y^{n-1})$ are the same. This necessary and sufficient condition that we establish, in the generality of any finite state channel, follows from the KKT conditions [14, Ch. 5] for convex optimization problems.

The remainder of the paper is organized as follows. In Section II, we briefly present the definitions of directed information and causal conditioning pmfs that we use throughout the paper. In Section III, we show that the optimization problem of maximizing the directed information over causal conditioning pmfs is convex. Additionally, using the KKT conditions, we show that if the output distribution induced by the conditional pmfs that achieve the maximum in the presence of feedback can also be induced by an input distribution that does not use feedback, then feedback does not increase the capacity. In Section IV we consider a binary POST(a,b) channel and we show that feedback does not increase capacity for this considerably larger class of channels. In Section V, we conclude and suggest some directions for further research on the family of POST channels.

II. Directed information, causal conditioning and notations

Throughout this paper, we denote random variables by capital letters such as X. The probability $\Pr\{X=x\}$ is denoted by $p(x)$. We denote the whole vector of probabilities by capital P, i.e., $P(x)$ is the probability vector of the random variable X.

We use the *causal conditioning* notation $(\cdot||\cdot)$ developed by Kramer [15]. We denote by $p(x^n||y^{n-d})$ the probability mass function of $X^n=(X_1,\ldots,X_n)$, *causally conditioned* on Y^{n-d} for some integer $d\geq 0$, which is defined as

$$p(x^n||y^{n-d}):=\prod_{i=1}^{n}p(x_i|x^{i-1},y^{i-d}). \tag{3}$$

By convention, if $i<d$, then y^{i-d} is set to null, i.e., if $i<d$ then $p(x_i|x^{i-1},y^{i-d})$ is just $p(x_i|x^{i-1})$. In particular, we use extensively the cases $d=0,1$:

$$p(x^n||y^{n}):=\prod_{i=1}^{n}p(x_i|x^{i-1},y^{i}), \tag{4}$$

$$p(x^n||y^{n-1}):=\prod_{i=1}^{n}p(x_i|x^{i-1},y^{i-1}). \tag{5}$$

The directed information was defined by Massey [16], inspired by Marko's work [17] on bidirectional communication, as

$$I(X^n\to Y^n):=\sum_{i=1}^{n}I(X^i;Y_i|Y^{i-1}). \tag{6}$$

The directed information can also be rewritten as

$$I(X^n\to Y^n)=\sum_{x^n,y^n}p(x^n||y^{n-1})p(y^n||x^n)\log\frac{p(y^n||x^n)}{\sum_{x^n}p(x^n||y^{n-1})p(y^n||x^n)} \tag{7}$$

This is due to the definition of causal conditioning and the chain rule

$$p(x^n,y^n)=p(x^n||y^{n-1})p(y^n||x^n). \tag{8}$$

We will make use the fact that directed information $I(X^n\to Y^n)$ is concave in $P(x^n||y^{n-1})$ for a fixed $P(y^n||x^n)$.

Directed information characterizes the capacity of point-to-point channels with feedback [10], [18]–[20]. For channels where the state is a function of the output, of which the POST channel is a special case, it was shown [7], [10] that the feedback capacity is given by

$$C_{fb}=\lim_{n\to\infty}\frac{1}{n}\max_{P(x^n||y^{n-1})}I(X^n\to Y^n). \tag{9}$$

On the other hand, without feedback the capacity is given by

$$C=\lim_{n\to\infty}\frac{1}{n}\max_{P(x^n)}I(X^n\to Y^n), \tag{10}$$

since the channel is indecomposable [21]. In the case where there is no feedback, namely, the Markov form $X_i-X^{i-1}-Y^{i-1}$ holds, $I(X^n\to Y^n)=I(X^n;Y^n)$, as shown in [16].

III. Maximization of the directed information as a convex optimization problem

In order to show that feedback does not increase the capacity of POST channels, we consider the two optimization problems:

$$\max_{P(x^n||y^{n-1})}I(X^n\to Y^n) \tag{11}$$

and

$$\max_{P(x^n)}I(X^n\to Y^n). \tag{12}$$

In this section, we claim that both problems are convex optimization problems, and use the KKT condition to state a necessary and sufficient condition for the two optimization problems to obtain the same value.

A convex optimization problem, as defined in [14, Ch. 4], is a problem of the form

$$\begin{aligned}&\text{minimize} && f_0(x)\\ &\text{subject to} && f_i(x)\leq b_i \quad i=1,\cdots,k\\ &&& g_j(x)=0 \quad j=1,\cdots,l\end{aligned} \tag{13}$$

where $f_0(x)$ and $\{f_i(x)\}_{i=1}^k$ are convex functions, and $\{g_j(x)\}_{j=1}^l$ are affine.

In order to convert the optimization problem in (11) into a convex optimization problem, as presented in (13), we need to show that the set of conditional pmfs $P(x^n||y^{n-1})$ can be expressed using inequalities that contains only convex functions and equalities that contains affine functions.

Lemma 1 (Causal conditioning is a polyhedron): The set of all causal conditioning distributions of the form $P(x^n||y^{n-1})$ is a polyhedron in $\mathbb{R}^{|\mathcal{X}|^n|\mathcal{Y}|^{n-1}}$ and is given by a set of linear equalities and inequalities of the form:

$$\begin{array}{ll}p(x^n||y^{n-1})\geq 0, & \forall x^n,y^{n-1},\\ \sum_{x_{i+1}^n}p(x^n||y^{n-1})=\gamma_{x^i,y^{i-1}}, & \forall x^i,y^{n-1},i\geq 1,\\ \sum_{x_1^n}p(x^n||y^{n-1})=1, & \forall y^{n-1}.\end{array} \tag{14}$$

Note that the two equalities in (14) may be unified into one if we add $i=0$ to the equality cases and we restrict the corresponding γ to be unity. Furthermore, for $n=1$ we obtain the regular vector probability, i.e., $p(x)\geq 0$, $\forall x$ and $\sum_x P(x)=1$.

Note that the optimization problem given in (11) is a convex optimization one since the set of causal conditioning pmfs is a polyhedron (Lemma 1) and the directed information is concave in $P(x^n||y^{n-1})$ for a fixed $P(y^n||x^n)$ [22, Lemma 2]. Therefore, the KKT conditions [14, Ch 5.5.3] are necessary and sufficient. The next theorem states these conditions explicitly for our setting.

Theorem 2: A set of necessary and sufficient conditions for an input probability $P(x^n||y^{n-1})$ to maximize the optimization problem in (10) is that for some numbers $\beta_{y^{n-1}}$

$$\begin{aligned}
&\sum_{y_n} p(y^n||x^n)\log\frac{p(y^n||x^n)}{ep(y^n)} = \beta_{y^{n-1}},\ \text{if } p(x^n||y^{n-1})>0,\\
&\sum_{y_n} p(y^n||x^n)\log\frac{p(y^n||x^n)}{ep(y^n)} \le \beta_{y^{n-1}},,\ \text{if } p(x^n||y^{n-1})=0,
\end{aligned} \tag{15}$$

where $p(y^n)=\sum_{x^n} p(y^n||x^n)p(x^n||y^{n-1})$. Furthermore, the solution of the optimization problem is

$$\max_{P(x^n||y^{n-1})} I(X^n\to Y^n) = \sum_{y^{n-1}} \beta_{y^{n-1}} + 1. \tag{16}$$

IV. Capacity of the POST(a,b) channel with and without feedback

Before considering the POST(a,b) let us first consider the binary DMC with parameters (a,b). The capacity of the binary DMC with parameters (a,b) was derived by Ash in [23, Ex 3.7] by applying [23, Theorem 3.3.3] and is given by

$$C = \log\left[2^{\frac{\bar{a}H_b(b)-bH_b(a)}{a+b-1}} + 2^{\frac{\bar{b}H_b(a)-aH_b(b)}{a+b-1}}\right]. \tag{17}$$

The capacity achieving input distribution is

$$\begin{aligned}
P(x=0) &= c_0\left(b2^{\frac{H(b)}{a+b-1}} - \bar{b}2^{\frac{H(a)}{a+b-1}}\right),\\
P(x=1) &= c_0\left(-\bar{a}2^{\frac{H(b)}{a+b-1}} + a2^{\frac{H(a)}{a+b-1}}\right),
\end{aligned} \tag{18}$$

where c_0 is a normalizing coefficient so that the sum $P(x=0)+P(x=1)$ is equal to 1. The induced output distribution is

$$P(y=0) = c_0(ab-\bar{a}\bar{b})2^{\frac{H(b)}{a+b-1}} \tag{19}$$

$$P(y=1) = c_0(ab-\bar{a}\bar{b})2^{\frac{H(a)}{a+b-1}}. \tag{20}$$

Lemma 3 (Feedback capacity of POST(a,b)): The feedback capacity of the POST(a,b) channel is the same as of the memoryless DMC with parameters (a,b), which is given in (17).

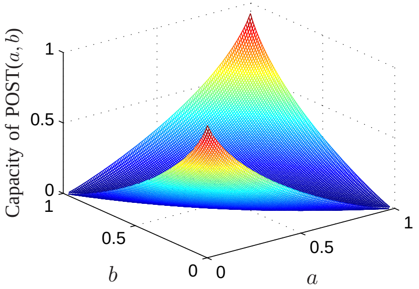


Fig. 2. The capacity of the POST(a,b) channel with and without feedback. This is also the capacity of the binary DMC with parameters (a,b)

We now present sufficient conditions on a,b implying that feedback does not increase the capacity of the POST(a,b) channel. That these conditions are indeed sufficient we establish in the next subsection. Define the following intervals:

$$\begin{aligned}
\mathcal{L}_1 &= \Bigg\{\max(\frac{\bar{a}}{\bar{b}}\gamma, \frac{\gamma(\bar{a}+b)-\sqrt{\gamma^2(\bar{a}+b)^2-4a\bar{b}}}{2\bar{b}})\\
&\qquad \le \beta \le \frac{\gamma(\bar{a}+b)+\sqrt{\gamma^2(\bar{a}+b)^2-4a\bar{b}}}{2\bar{b}}.\Bigg\}\\
\mathcal{L}_2 &= \left\{\frac{(a+\bar{b})+\sqrt{(a+\bar{b})^2-4\bar{a}b\gamma^2}}{2b\gamma} \le \beta \le \frac{\bar{a}}{\bar{b}}\gamma\right\}\\
\mathcal{L}_3 &= \left\{\beta \le \min(\frac{\bar{a}}{\bar{b}}\gamma, \frac{(a+\bar{b})-\sqrt{(a+\bar{b})^2-4\bar{a}b\gamma^2}}{2b\gamma})\right\}\\
\mathcal{L}_4 &= \left\{\beta \le \min(\frac{b\gamma}{a}, \frac{\gamma(\bar{a}+b)-\sqrt{\gamma^2(\bar{a}+b)^2-4a\bar{b}}}{2a})\right\}\\
\mathcal{L}_5 &= \left\{\frac{\gamma(\bar{a}+b)+\sqrt{\gamma^2(\bar{a}+b)^2-4a\bar{b}}}{2a} \le \beta \le \frac{b\gamma}{a}\right\}\\
\mathcal{L}_6 &= \Bigg\{\max(\frac{b\gamma}{a}, \frac{(a+\bar{b})-\sqrt{(a+\bar{b})^2-4\bar{a}b\gamma^2}}{2\bar{a}\gamma})\\
&\qquad \le \beta \le \frac{(a+\bar{b})+\sqrt{(a+\bar{b})^2-4\bar{a}b\gamma^2}}{2\bar{a}\gamma}\Bigg\},
\end{aligned} \tag{21}$$

where γ is defined as

$$\gamma = 2^{\frac{H(b)-H(a)}{a+b-1}}. \tag{22}$$

In addition, let

$$\mathcal{L}_0 = \left\{1 \le \beta \le \min(\frac{a}{\bar{a}\gamma}, \frac{b\gamma}{\bar{b}})\right\} \tag{23}$$

Lemma 4: If the intersections of the intervals $\mathcal{L}_1\cup\mathcal{L}_2\cup\mathcal{L}_3$ with $\mathcal{L}_4\cup\mathcal{L}_5\cup\mathcal{L}_6$ and $\mathcal{L}_0$ is nonempty then feedback does not increase the capacity of the POST(a,b) channel.

Lemma 5: The condition in Lemma 4 holds for all POST channel parameters (a,b). Thus, feedback does not increase capacity of POST(a,b).

A. Deriving the sufficient conditions of Lemma 4

Proof of Lemma 4: Let $P_{n,0}$ and $P_{n,1}$ be defined as

$$P_{n,0} = \begin{bmatrix} a\cdot P_{n-1,0} & \bar{b}\cdot P_{n-1,0}\\ \bar{a}\cdot P_{n-1,1} & b\cdot P_{n-1,1}\end{bmatrix} \tag{24}$$

and

$$P_{n,1} = \begin{bmatrix} b\cdot P_{n-1,0} & \bar{a}\cdot P_{n-1,0}\\ \bar{b}\cdot P_{n-1,1} & a\cdot P_{n-1,1}\end{bmatrix} \tag{25}$$

where $P_{0,0} = P_{0,1} = 1$. Inverting the matrices, we obtain

$$P_{n,0}^{-1} = \begin{bmatrix} \frac{b}{ba-\bar{a}\bar{b}} P_0^{-1} & -\frac{\bar{b}}{ba-\bar{a}\bar{b}} P_1^{-1} \\ -\frac{\bar{a}}{ba-\bar{a}\bar{b}} P_0^{-1} & \frac{a}{ba-\bar{a}\bar{a}} P_1^{-1} \end{bmatrix} \tag{26}$$

$$P_{n,1}^{-1} = \begin{bmatrix} \frac{a}{ba-\bar{a}\bar{b}} P_0^{-1} & -\frac{\bar{a}}{ba-\bar{a}\bar{b}} P_1^{-1} \\ -\frac{\bar{b}}{ba-\bar{a}\bar{b}} P_0^{-1} & \frac{b}{ba-\bar{a}\bar{b}} P_1^{-1} \end{bmatrix} \tag{27}$$

Now we compute $P_1(x^n)$ and $P_0(x^n)$

$$P_0(x^n) = P_{n,0}^{-1} P_0(y^n) \tag{28}$$

$$P_1(x^n) = P_{n,1}^{-1} P_1(y^n) \tag{29}$$

where $P_0(x^0) = P_1(x^0) = 1$. We can rewrite $P_0(x^n)$ and $P_1(x^n)$ follows:

$$P_0(x^n) = \frac{1}{(a+b-1)(\gamma+1)} \begin{bmatrix} b\gamma P_0(x^{n-1}) - \bar{b}P_1(x^{n-1}) \\ -\bar{a}\gamma P_0(x^{n-1}) + aP_1(x^{n-1}) \end{bmatrix} \tag{30}$$

$$P_1(x^n) = \frac{1}{(a+b-1)(\gamma+1)} \begin{bmatrix} aP_0(x^{n-1}) - \bar{a}\gamma P_1(x^{n-1}) \\ -\bar{b}P_0(x^{n-1}) + b\gamma P_1(x^{n-1}) \end{bmatrix} \tag{31}$$

We need to show that indeed the probability expressions are valid, namely nonnegative and sum to 1. Showing the non-negativity of each of the terms in the above expression is equivalent to showing $\forall n \geq 1$ and for all x^{n-1},

$$\begin{aligned} \min\{\frac{a}{\bar{a}\gamma}, \frac{b\gamma}{\bar{b}}\} P_0(x^{n-1}) &\geq P_1(x^{n-1}) \\ \min\{\frac{a}{\bar{a}\gamma}, \frac{b\gamma}{\bar{b}}\} P_1(x^{n-1}) &\geq P_0(x^{n-1}). \end{aligned} \tag{32}$$

For $n = 1$ this follows from the fact that $\min\{\frac{a}{\bar{a}\gamma}, \frac{b\gamma}{\bar{b}}\} \geq 1$. To prove for $n \geq 1$ we use the following lemma. ■

Lemma 6: If the condition in Lemma 4 holds then there exists, $1 \leq \beta \leq \min\{\frac{a}{\bar{a}\gamma}, \frac{b\gamma}{\bar{b}}\}$ such that $\forall n$, the inequalities

$$\begin{aligned} \beta P_1(x^{n-1}) \geq P_0(x^{n-1}),\ \forall x^{n-1}, \\ \beta P_0(x^{n-1}) \geq P_1(x^{n-1}),\ \forall x^{n-1}, \end{aligned} \tag{33}$$

imply

$$\begin{aligned} \beta P_1(x^n) \geq P_0(x^n),\ \forall x^n, \\ \beta P_0(x^n) \geq P_1(x^n),\ \forall x^n. \end{aligned} \tag{34}$$

V. Conclusion and further research

We have introduced and studied the family of POST channels and showed, somewhat surprisingly, that feedback does not increase the capacity of the general $POST(a, b)$ channel. The proof is based on finding the output probability that is induced by the input causal conditioning pmf which optimizes the directed information when feedback is allowed, and then proving that this output pmf can be also be induced by an input distribution without feedback. There may be a more direct way, that has thus far eluded us, for proving that feedback does not increase the capacity of the Simple POST channel. We hope that the POST channel introduced in this paper will enhance our understanding of capacity of finite state channels with and without feedback, and help us to find simple capacity-achieving codes.

Acknowledgement

The authors are grateful to Jiantao Jiao who suggested a proof of an inequity that we used in order to prove Lemma 6.

References

[1] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory*, 18:460–473, 1972.

[2] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory*, 18:14–20, 1972.

[3] C. E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, January 1949.

[4] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Izv. Akad. Nauk, Moskva, 1960. in Russian, translated by A. Feinstein in 1964.

[5] F. Alajaji and T. Fuja. Effect of feedback on the capacity of discrete additive channels with memory. In *Proceedings ISIT94*, Norway, 1994. IEEE.

[6] Y.-H. Kim. Feedback capacity of stationary Gaussian channels. *IEEE Trans. Inf. Theory.*, 56(1):57–85, 2010.

[7] H. H. Permuter, P. Cuff, B. Van Roy, and T. Weissman. Capacity of the trapdoor channel with feedback. *IEEE Trans. Inf. Theory*, 54(7):3150–3165, 2009.

[8] O. Elishco and H. H. Permuter. Capacity and coding for the Ising channel with feedback. submitted to *IEEE Trans. Inf. Theory.* Available at arxiv.org/abs/1205.4674, 2012.

[9] A.J. Goldsmith and P.P. Varaiya. Capacity of fading channels with channel side information. *IEEE Trans. Inf. Theory*, 43(6):1986 –1992, 1997.

[10] J. Chen and T. Berger. The capacity of finite-state Markov channels with feedback. *IEEE Trans. Inf. Theory*, 51:780–789, 2005.

[11] H. Asnani, H. Permuter, and T. Weissman. Capacity of a post channel with and without feedback. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2013.

[12] H. Permuter, H. Asnani, and T. Weissman. Capacity of a post channel with and without feedback. submitted to *IEEE Trans. Inf. Theory.* Available at arxiv.org/abs/1309.5440, 2013.

[13] H. Asnani, H. H. Permuter, and T. Weissman. To feed or not to feed back. 2010. submitted to *IEEE Trans. Inf. Theory.* Available at arxiv.org/abs/1011.1607.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New-York, 2004.

[15] G. Kramer. *Directed information for channels with feedback.* Ph.D. dissertation, Swiss Federal Institute of Technology (ETH) Zurich, 1998.

[16] J. Massey. Causality, feedback and directed information. *Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*, pages 303–305, Nov. 1990.

[17] H. Marko. The bidirectional communication theory- a generalization of information theory. *IEEE Trans. on communication*, COM-21:1335–1351, 1973.

[18] Y.-H. Kim. A coding theorem for a class of stationary channels with feedback. *IEEE Trans. Inf. Theory.*, 25:1488–1499, April, 2008.

[19] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *IEEE Trans. Inf. Theory*, 55:323–349, 2009.

[20] H. H. Permuter, T. Weissman, and A. J. Goldsmith. Finite state channels with time-invariant deterministic feedback. *IEEE Trans. Inf. Theory*, 55(2):644–662, 2009.

[21] R. G. Gallager. *Information theory and reliable communication.* Wiley, New York, 1968.

[22] I. Naiss and H. H. Permuter. Extension of the Blahut-Arimoto algorithm for maximizing directed information. *IEEE Trans. Inf. Theory*, 59:204–222, 2013.

[23] R. Ash. *Information Theory*. Wiley, New York, 1965.

Unshared Secret Key Cryptography

(Invited Paper)

Shuiyin Liu, Yi Hong, and Emanuele Viterbo
ECSE Department, Monash University
Melbourne, VIC 3800, Australia
Email: shuiyin.liu, yi.hong, emanuele.viterbo@monash.edu

***Abstract*—Inspired by the *artificial noise* technique by Goel *et al.*, we propose an unshared secret key (USK) cryptosystem, where the artificial noise is redesigned as a one-time pad secret key aligned within the null space between transmitter and legitimate receiver. Unlike previously studied artificial noise techniques, rather than ensuring non-zero secrecy capacity, the USK cryptosystem guarantees Shannon's *perfect secrecy* without the need of secret key exchange.**

I. Introduction

Wireless communications provide flexibility and mobility for users, but equally the ease of access features undermines user privacy. Research on secure communication falls into two categories: network layer cryptography and physical layer security (PLS). The former assumes that the physical layer provides error-free data links, in which security depends on a shared secret key. In the latter, the strategy is to use wiretap codes to protect the secret data from eavesdropping, while security comes from specific channel limitations for the eavesdropper. Both categories are rooted in Shannon's *perfect secrecy* [1], defined as the mutual information $I(\mathbf{u};\, \mathbf{y}) = 0$; that is, the secret message $\mathbf{u}$ and the eavesdropper's received message $\mathbf{y}$ are mutually independent. Perfect secrecy requires one-time pad secret key [1].

The PLS scheme, known as *artificial noise* (AN) [2], is the basis for our unshared secret key (USK) cryptosystem. In the AN scheme, the transmitter (Alice) aligns a jamming signal, called artificial noise, within the null space between itself and the legitimate receiver (Bob), thus AN only degrades the eavesdropper's (Eve's) channel. The strategy is to use Gaussian distributed AN to guarantee non-zero secrecy capacity [3]. Given such secrecy capacity, infinite-length wiretap codes can be used to achieve strong secrecy [4]. More recently, we proposed a variant of AN using a finite *M*-QAM alphabet, called *practical secrecy* (PS) scheme, where instead of increasing the secrecy rate with AN, the eavesdropper's error probability is maximized [5].

In this work, we show that the PS scheme is *de facto* an USK, where AN serves as an unshared one-time pad secret key. The result is a development of our understanding of the benefits of AN, embracing both coding and cryptographic dimensions. We show that the USK provides Shannon's perfect secrecy, with no secret key exchange, under Goel *et al.*'s assumptions on the physical channels that enable the use of the AN scheme.

Our work differs from previous studies of AN [2], because it puts forward four new aspects that were not previously accounted for:

1) *Perfect secrecy*: we aim to achieving Shannon's perfect secrecy directly, rather than ensuring non-zero secrecy capacity.
2) *Finite alphabet*: we use a finite alphabet (*M*-QAM) rather than infinite-length wiretap codes.
3) *Artificial noise*: we have no requirement of the distribution of AN; that is, not necessarily Gaussian.
4) *Lattice precoding*: we introduce lattice precoding to MIMO wiretap channels, which avoids the diversity loss caused by conventional *singular value decomposition* (SVD) precoding of [2].

Notation: Matrices and column vectors are denoted by upper and lowercase boldface letters, and the Hermitian transpose, inverse, pseudoinverse of a matrix $\mathbf{B}$ by $\mathbf{B}^H$, $\mathbf{B}^{-1}$, and $\mathbf{B}^\dagger$, respectively. Let $\{X_n, X\}$ be defined on the same probability space. We write $X_n \overset{a.s.}{\to} X$ if X_n converges to X almost surely or with probability one. $\mathbf{I}_n$ denotes the identity matrix of size n. We write $\triangleq$ for equality in definition. A circularly symmetric complex Gaussian random variable x with variance σ^2 is denoted as $x \backsim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. The real, complex, integer and complex integer numbers are denoted by $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}$ and $\mathbb{Z}[i]$, respectively. $H(X)$, $H(X|Y)$ and $I(X;Y)$ represent entropy, conditional entropy and mutual information, respectively. We use the standard asymptotic notation $f(x) = O(g(x))$ when $\lim\limits_{x\to\infty} \sup |f(x)/g(x)| < \infty$. vol$(S)$ denotes the Euclidean volume of S.

II. System Model

We consider a MIMO wiretap system, including a transmitter (Alice), an intended receiver (Bob), and a passive eavesdropper (Eve), with N_{A}, N_{B}, and N_{E} antennas, respectively. The signals received by Bob and Eve are given, respectively, by

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{n}_{\text{B}}, \tag{1}$$

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{n}_{\text{E}}, \tag{2}$$

where the entries of $\mathbf{n}_{\text{B}}$ and $\mathbf{n}_{\text{E}}$ are i.i.d. complex random variables $\sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{B}}^2)$ and $\mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{E}}^2)$, respectively. We assume that the matrices $\mathbf{H}$ and $\mathbf{G}$, representing the channels from

This work is supported by ARC under Grant Discovery Project No. DP130100336.

Alice to Bob and Alice to Eve, respectively, are mutually independent, i.e., Bob and Eve are not co-located. The entries of $\mathbf{H}$ and $\mathbf{G}$ are i.i.d. complex random variables $\sim \mathcal{N}_{\mathbb{C}}(0,\, 1)$.

A. Artificial Noise Scheme

We first introduce the AN scheme [2]. Assuming $N_{\mathrm{B}} < N_{\mathrm{A}}$, $\mathbf{H}$ has a non-trivial null space $\mathbf{Z} = \mathrm{null}(\mathbf{H})$. Alice transmits

$$\mathbf{x} = \mathbf{P}\mathbf{u} + \mathbf{Z}\mathbf{v} \tag{3}$$

where $\mathbf{u}$ is the secret data vector and $\mathbf{P}$ is the precoding matrix. The AN $\mathbf{v}$ is generated by Alice and is unknown to Eve. In order to estimate the secrecy rate, both $\mathbf{u}$ and $\mathbf{v}$ are assumed to be Gaussian circularly symmetric random vectors.

The AN scheme is based on the channel assumptions below:

1) Alice only knows the realization of $\mathbf{H}$.
2) Eve knows the realizations of $\mathbf{H}$, $\mathbf{G}$, $\mathbf{Z}$ and $\mathbf{P}$.
3) $N_{\mathrm{A}} > N_{\mathrm{B}}$, $N_{\mathrm{A}} > N_{\mathrm{E}}$ and $N_{\mathrm{E}} \geq N_{\mathrm{B}}$.

Equations (1) and (2) can then be rewritten as

$$\mathbf{z} = \mathbf{H}\mathbf{P}\mathbf{u} + \mathbf{n}_{\mathrm{B}}, \tag{4}$$

$$\mathbf{y} = \mathbf{G}\mathbf{P}\mathbf{u} + \mathbf{G}\mathbf{Z}\mathbf{v} + \mathbf{n}_{\mathrm{E}}. \tag{5}$$

Thus, $\mathbf{v}$ only degrades Eve's reception, but not Bob's.

In (3), the transmitted signal $\mathbf{x}$ depends on the precoding matrix $\mathbf{P}$. The AN scheme uses *SVD precoding*, given by

$$\mathbf{x}_{\mathrm{SVD}}= \mathbf{V}_1\mathbf{u} + \mathbf{Z}\mathbf{v}_{\mathrm{SVD}}, \tag{6}$$

where $\mathbf{P} = \mathbf{V}_1$ and the columns of $\mathbf{V} = [\mathbf{V}_1,\, \mathbf{Z}]$ are the right-singular vectors of $\mathbf{H}$, i.e., $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$.

For the AN scheme, given a positive secrecy rate, infinite-length wiretap codes can be used to achieve strong secrecy [4], i.e.,

$$\lim_{n\to\infty} I(\mathbf{u};\mathbf{y}) =0, \tag{7}$$

where n represents the codeword length.

B. Practical Secrecy Scheme

Based on the AN scheme, we proposed the PS scheme, where the security measure in AN, secrecy capacity, is replaced by Eve's error probability [5]. Although the transmission model is the same as that given in (4) and (5), $\mathbf{u}$ and $\mathbf{v}$ are not required to be Gaussian distributed. The settings of the PS scheme are given below.

1) Uniform $M-$QAM signalling, i.e., $\Re(\mathbf{u})$ and $\Im(\mathbf{u}) \in \mathcal{C}^{N_{\mathrm{B}}}$, where $\mathcal{C} =\{-\sqrt{M}+1,\, -\sqrt{M}+3,\, ...,\, \sqrt{M}-1\}$, is used.
2) There is no requirement on the distribution of $\mathbf{v}$.

The PS scheme can use either SVD precoding or *lattice precoding* [6], in which

$$\mathbf{x}_{\mathrm{LP}}= \mathbf{H}^{\dagger}(\mathbf{u}{-}A\hat{\mathbf{w}}) + \mathbf{Z}\mathbf{v}_{\mathrm{LP}}, \tag{8}$$

where $A = 2\sqrt{M}$, $\mathbf{P} = \mathbf{H}^{\dagger}$ and

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathbb{Z}[i]^{N_{\mathrm{B}}}} \|\mathbf{H}^{\dagger}(\mathbf{u} - A\mathbf{w})\|^2. \tag{9}$$

Compared with the AN scheme, where the achievability of security is based on an infinite length code, the PS scheme is designed for practical communication systems, which make use of a finite alphabet. However, a security scheme based on error probability may be not safe in the sense of information-theoretic security.

In this work, we analyze and enhance the security of the PS scheme under the same channel assumptions as AN. To simplify our analysis, we unify the notation of $\mathbf{u}$ by defining

$$\tilde{\mathbf{u}} \triangleq \begin{cases} \mathbf{u}{-}A\hat{\mathbf{w}} & \text{lattice precoding} \\ \mathbf{u} & \text{SVD precoding} \end{cases} \tag{10}$$

We define the noise-plus-interference term at Eve as

$$\tilde{\mathbf{n}}_{\mathrm{v}} \triangleq \mathbf{G}\mathbf{Z}\mathbf{v} + \mathbf{n}_{\mathrm{E}}. \tag{11}$$

III. Unshared Secret Key Cryptosystem

In this section, we first interpret the PS scheme from a cryptographic perspective, and then prove its security in terms of perfect secrecy.

A. Encryption

The AN $\mathbf{v}$ used in the PS scheme can be treated as a one-time pad secret key. Alice randomly (without any predefined distribution) chooses $\mathbf{v}$ from the set S defined by

$$S \triangleq \left\{\mathbf{v} \in \mathbb{R}^{N_{\mathrm{A}}-N_{\mathrm{B}}}\colon \|\mathbf{v}\|^2 \leq P\right\}, \tag{12}$$

where P represents the transmission power constraint on $\mathbf{v}$.

The message $\tilde{\mathbf{u}}$ is received by Eve as a lattice point in: $\Lambda_{\mathbb{C}} = \{\mathbf{G}\mathbf{P}\tilde{\mathbf{u}},\, \tilde{\mathbf{u}} \in \mathbb{Z}\,[i]^{N_{\mathrm{B}}}\}$ (see Fig. 1). The set S can be further partitioned into D subsets S_1, ..., S_D, i.e.,

$$S = \bigcup_{k=1}^{D} S_k, \tag{13}$$

where

$$S_k \triangleq \left\{\mathbf{v}\colon \mathbf{G}\mathbf{P}\tilde{\mathbf{u}} \in \Lambda_{\mathbb{C}} \text{ is the } k^{\text{th}} \text{ closest lattice point to } \mathbf{y}\right\}.$$

Later, we will show that the value of D can be uniquely characterized by P.

Assuming $\mathbf{v} \in S_k$, $1 \leq k \leq D$, the PS scheme thus can be viewed as a cryptosystem that encrypts $\tilde{\mathbf{u}}$ to $\mathbf{y}$ using a secret key $\mathbf{v}$, such that $\mathbf{G}\mathbf{P}\tilde{\mathbf{u}}$ is the k^{th} closest lattice point to $\mathbf{y}$ (see Fig. 1).

From Eve's perspective, we assume that she knows P and the above encryption process. Since Eve cannot know the secret key $\mathbf{v}$, she cannot know the distribution of k either. It means that Eve only knows that $\mathbf{G}\mathbf{P}\tilde{\mathbf{u}}$ is hidden inside the D closest lattice points to $\mathbf{y}$, but cannot locate it. Moreover, Eve cannot distinguish which lattice point has the highest probability of being $\mathbf{G}\mathbf{P}\tilde{\mathbf{u}}$, thus the probability that Eve obtains $\mathbf{G}\mathbf{P}\tilde{\mathbf{u}}$ is uniform over all D lattice points. By taking the codebook size of $\mathbf{u}$ into account, for a given $\mathbf{G}\mathbf{P}$, we have

$$\Pr\{\mathbf{G}\mathbf{P}\tilde{\mathbf{u}}|\mathbf{y}\} = \Pr\{\mathbf{u}|\mathbf{y}\} = \frac{1}{\min\left\{D,\, M^{N_{\mathrm{B}}}\right\}}, \tag{14}$$

or equivalently

$$H(\mathbf{u}|\mathbf{y}) = \log\min\left\{D,\, M^{N_{\mathrm{B}}}\right\}. \tag{15}$$

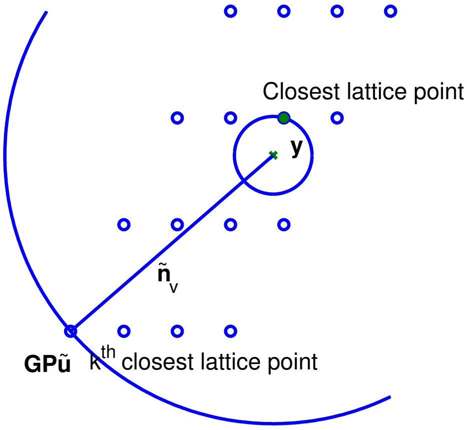


Fig. 1. Achieving perfect secrecy.

Different from Shannon's one-time pad cryptosystem, the one-time pad secret key $\mathbf{v}$ is not shared between Alice and Bob. In particular, it is independently generated by Alice, but not needed by Bob to decipher, while it is fully affecting Eve's ability to decipher the original message. We name this kind cryptosystem as Unshared Secret Key (USK) cryptosystem.

B. Decryption

From (4), Bob can simply run maximum likelihood decoding to estimate $\mathbf{u}$. We then show how to increase Eve's uncertainty of $\mathbf{u}$, i.e., $H(\mathbf{u}|\mathbf{y})$.

Based on (15), increasing $H(\mathbf{u}|\mathbf{y})$ is equivalent to increasing D. The value of D depends on the channel matrices $\mathbf{G}$ and $\mathbf{H}$. In this work, we assume $\mathbf{G}$ and $\mathbf{H}$ are not fixed, but are Gaussian random matrices. In this sense, for a given $\mathbf{v}$ and a positive integer c, $\Pr\{D > c|\mathbf{G},\mathbf{H}\}$ is also a random variable depending on $\mathbf{G}$ and $\mathbf{H}$, and $\tilde{\mathbf{n}}_{\mathrm{v}}$ is a Gaussian random vector with i.i.d. entries $\mathcal{N}_{\mathbb{C}}(0,\, \tilde{\sigma}_{\mathrm{v}}^2)$ where

$$\tilde{\sigma}_{\mathrm{v}}^2 = ||\mathbf{v}||^2 + \sigma_{\mathrm{E}}^2. \tag{16}$$

We recall that the realizations of G of $\mathbf{G}$ and H of $\mathbf{H}$ are known at Eve. Note that if we can ensure $\Pr\{D > c|\mathbf{G},\mathbf{H}\} \overset{a.s.}{\to} 1$, then $D > c$ for almost any realizations G and H (see [7, Def. 1.3]).

In fact, the idea behind the original PS scheme was to ensure $\Pr\{D > 1|\mathbf{G},\mathbf{H}\} \overset{a.s.}{\to} 1$, which is a special case of the USK with $c = 1$.

C. Achieving Perfect Secrecy

We now show how large P should be to guarantee perfect secrecy, i.e.,

$$I(\mathbf{u};\mathbf{y}) = 0. \tag{17}$$

From [1, Th.6], the necessary and sufficient condition to achieve perfect secrecy is

$$\Pr\{\mathbf{u}\} = \Pr\{\mathbf{u}|\mathbf{y}\}. \tag{18}$$

Since $\Pr\{\mathbf{u}\} = 1/M^{N_{\mathrm{B}}}$, based on (14), a sufficient condition to achieve perfect secrecy is $D \geq M^{N_{\mathrm{B}}}$.

In what follows, we evaluate the value of D by choosing $||\mathbf{v}||^2 = P$, i.e., on the surface of S in (12).

Lemma 1: Let vol$(\Lambda_{\mathbb{C}})$ be the volume of the *Voronoi* cell of $\Lambda_{\mathbb{C}}$.

$$\Pr\left\{D \leq M^{N_{\mathrm{B}}}|\mathbf{G},\mathbf{H}\right\} \leq \frac{M^{N_{\mathrm{B}}}\mathrm{vol}(\Lambda_{\mathbb{C}})}{\pi^{N_{\mathrm{E}}}P^{N_{\mathrm{E}}}} \triangleq \Delta. \tag{19}$$

Proof: See Appendix A. ■

Note that Δ is a random variable depending on $\Lambda_{\mathbb{C}}$ defined by the random matrices $\mathbf{G}$ and $\mathbf{H}$. From Lemma 1, by sending Δ to zero, $\Pr\{D \leq M^{N_{\mathrm{B}}}|\mathbf{G},\mathbf{H}\}$ is forced to zero as well, i.e., achieving perfect secrecy. In the following theorem, we show how to ensure $\Delta \overset{a.s.}{\to} 0$.

Lemma 2: Let

$$\kappa \triangleq M^{N_{\mathrm{B}}/(2N_{\mathrm{E}})}/\sqrt{\pi}. \tag{20}$$

If $P = \rho^2/\Phi^{2N_{\mathrm{B}}/N_{\mathrm{E}}}$ and $\rho > \kappa$, then $\Delta \overset{a.s.}{\to} 0$ as $N_{\mathrm{B}} \to \infty$, or equivalently,

$$\Pr\left\{\Delta > \left(\frac{\rho}{\kappa}\right)^{-N_{\mathrm{B}}}\right\} < O\left(\left(\frac{\rho}{\kappa}\right)^{-N_{\mathrm{B}}}\right) \tag{21}$$

where Φ depends on the precoder, i.e.,

$$\Phi_{\mathrm{LP}} = \left[\frac{(N_{\mathrm{E}}-N_{\mathrm{B}})!}{(N_{\mathrm{A}}-N_{\mathrm{B}})!}\cdot\frac{N_{\mathrm{A}}!}{N_{\mathrm{E}}!}\right]^{\frac{1}{2N_{\mathrm{B}}}} \quad \text{for lattice precoding} \tag{22}$$

$$\Phi_{\mathrm{SVD}} = \left[\frac{(N_{\mathrm{E}}-N_{\mathrm{B}})!}{N_{\mathrm{E}}!}\right]^{\frac{1}{2N_{\mathrm{B}}}} \quad \text{for SVD precoding} \tag{23}$$

Proof: Available in the journal version. ■

Lemmas 1 and 2 allow us to deduce our main theorem.

Theorem 1: If $P > \kappa^2/\Phi^{2N_{\mathrm{B}}/N_{\mathrm{E}}}$, perfect secrecy is achieved almost surely as $N_{\mathrm{B}} \to \infty$, where κ is given in (20) and Φ is given in (22) or (23).

IV. Conclusions

We have revisited the role that artificial noise plays in cryptography, showing that it can be used as unshared one-time pad secret keys. The proposed unshared secret key cryptosystem provides Shannon's perfect secrecy, and enjoys exemption from secret key exchange. Our work has highlighted that USK is valid for a finite alphabet such as M-QAM and a arbitrarily distributed artificial noise. Both lattice and SVD precoding are applicable to USK, significantly enhancing the utility of the cryptosystem. The basis is now established for future advances on generalizing USK to other channel.

Appendix

A. Proof of Lemma 1

Let $\mathcal{S}_{\mathrm{p}}$ be a sphere of radius R centered at $\mathbf{y}$, where vol$(\mathcal{S}_{\mathrm{p}}) = M^{N_{\mathrm{B}}}vol(\Lambda_{\mathbb{C}})$. Let K be the number of the points in $\mathcal{S}_{\mathrm{p}} \cap \Lambda_{\mathbb{C}}$. We have

$$\begin{aligned} K &\approx \frac{\mathrm{vol}(\mathcal{S}_{\mathrm{p}})}{\mathrm{vol}(\Lambda_{\mathbb{C}})} \\ &= M^{N_{\mathrm{B}}}. \end{aligned} \tag{24}$$

We recall that $\mathbf{GP\tilde{u}}$ is the k^{th} closest lattice point to $\mathbf{y}$ and $D \geq k$. Thus, if $\mathbf{GP\tilde{u}} \notin \mathcal{S}_{\mathrm{p}}$, we have $D > M^{N_{\mathrm{B}}}$.

Let $\mathcal{S}'_{\rm p}$ be a sphere with the same radius R centered at $\mathbf{GP\tilde{u}}$. If $\mathbf{GP\tilde{u}} \notin \mathcal{S}_{\rm p}$, then $\mathbf{y} \notin \mathcal{S}'_{\rm p}$, and vice versa. Therefore, we have

$$\begin{aligned} & \Pr\left\{D \leq M^{N_{\rm B}} | \mathbf{G}, \mathbf{H}\right\} \\ = \;& \Pr\left\{\mathbf{GP\tilde{u}} \in \mathcal{S}_{\rm p}\right\} \\ = \;& \Pr\left\{\mathbf{y} \in \mathcal{S}'_{\rm p}\right\} \\ = \;& \int_{\mathcal{S}'_{\rm p}} f(\tilde{\mathbf{n}}_{\rm v}) d\tilde{\mathbf{n}}_{\rm v} \\ \leq \;& \frac{M^{N_{\rm B}} \mathrm{vol}(\Lambda_{\mathbb{C}})}{\pi^{N_{\rm E}} \tilde{\sigma}_{\rm v}^{2N_{\rm E}}} \\ \leq \;& \frac{M^{N_{\rm B}} \mathrm{vol}(\Lambda_{\mathbb{C}})}{\pi^{N_{\rm E}} P^{N_{\rm E}}}, \end{aligned} \tag{25}$$

where $f(\tilde{\mathbf{n}}_{\rm v})$ is the probability density function (pdf) of $\tilde{\mathbf{n}}_{\rm v}$. The last inequalities hold since

$$\begin{aligned} f(\tilde{\mathbf{n}}_{\rm v}) &= \frac{1}{\pi^{N_{\rm E}} \tilde{\sigma}_{\rm v}^{2N_{\rm E}}} \exp\left(-\frac{||\tilde{\mathbf{n}}_{\rm v}||^2}{\tilde{\sigma}_{\rm v}^2}\right) \\ &\leq \frac{1}{\pi^{N_{\rm E}} \tilde{\sigma}_{\rm v}^{2N_{\rm E}}} \\ &= \frac{1}{\pi^{N_{\rm E}} \left(P + \sigma_{\rm E}^2\right)^{N_{\rm E}}} \\ &\leq \frac{1}{\pi^{N_{\rm E}} P^{N_{\rm E}}}. \end{aligned} \tag{26}$$

■

References

[1] C. E. Shannon, "Communication theory of secrecy systems," *Confidential report*, 1946.

[2] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2180–2189, Jun. 2008.

[3] S. K. Leung-Yan-Cheong and M. E. Hellman, "The Gaussian wire-tap channel," *IEEE Trans. Inf. Theory*, vol. 24, no. 4, pp. 451–456, Jul. 1978.

[4] I. Csiszár, "Almost independence and secrecy capacity," *Problems of Information Transmission*, vol. 32, pp. 40–47, 1996.

[5] S. Liu, Y. Hong, and E. Viterbo, "Practical secrecy using artificial noise," *IEEE Communications Letters*, vol. 17, no. 7, pp. 1483–1486, 2013.

[6] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector perturbation technique for near-capacity multiantenna multiuser communications-Part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, pp. 537–544, Mar. 2005.

[7] A. DasGupta, "Asymptotic theory of statistics and probability," in *Springer Texts in Statistics*. Springer-Verlag, 2008.

Some Constructions of Storage Codes from Grassmann Graphs

Frédérique Oggier
Division of Mathematical Sciences
Nanyang Technological University
Singapore
Email: frederique@ntu.edu.sg

***Abstract*—Codes for distributed storage systems may be seen as families of m-dimensional subspaces of the vector space $\mathbb{F}_q^n$, where $\mathbb{F}_q$ is the finite field with q elements, q a prime power. These subspaces need to intersect, to allow (collaborative) repair. We consider the Grassmann graph $\mathcal{G}_q(n,m)$ which has for vertex set the collection of m-dimensional subspaces of $\mathbb{F}_q^n$, and two vertices are adjacent whenever they intersect in a hyperplane. To obtain subspaces with regular intersection pattern, we look for cliques in the Grassmann graph, and obtain preliminary examples of storage codes, whose parameters we study, in terms of storage overhead, and repairability.**

I. Introduction

When data is stored across a network of nodes, it is usually replicated several times and the copies are stored on distinct nodes, to prevent data loss in case of node failures. From a coding point of view, this means that the data is encoded using a repetition code. It is thus natural to replace this code by a more efficient code, such as a maximal distance separable (MDS) code, which ensures the maximum reliability, given a storage overhead (or amount of data stored, versus amount of actual data). There is however a major difference between classical coding theory, and the design of codes for distributed storage systems, that of repairability. When some coefficients of a codeword are missing, it is desirable to recover these missing coefficients by downloading data from live nodes, *without* having to (necessarily) decode the codeword.

There has been an intense research activity around the notion of repairability over the past few years, and there is no complete consensus as of now, of what "good" repairability is. In [1], the authors propose adaptations of Reed-Solomon codes, where extra bits of parity are added to allow easy degraded reads, that is, to allow the data to be read, even though some coefficients of the codeword are missing. In [2], [3], repairs are done in a collaborative manner, that is once several coefficients of a codeword are missing, several nodes try to reconstruct the missing coefficients, by possibly exchanging data among each others. The authors focus on minimizing the amount of data downloaded per repair, called repair bandwidth. In [4] instead, repairs are optimized by contacting as little live nodes as possible. A survey of different design criteria for good repairability, and corresponding code constructions, is available in [5].

In this paper, we consider a different view point. We do not try to a priori design codes with respect to one of the known design criteria - repair bandwidth, degraded reads, or local repairs. Instead, we abstract codes for distributed storage systems as families of m-dimensional subspaces of the vector space $\mathbb{F}_q^n$ (or subset of the Grassmannian $G(m,n)$), try to design these subspaces with regular intersections, and analyze the preliminary examples obtained in terms of the relevance of their parameters to storage applications.

A similar formalization of storage codes in terms of linear subspaces (not in the context of collaborative repair) has been presented in [6]. The design of subsets of $G(m,n)$ with particular intersection has also been studied in the context of constant dimension codes for network coding [7].

We start by describing codes for distributed storage systems and abstract them in terms of subspaces and their intersection in Section II. To obtain subspaces whose pairwise intersection is of a given dimension, we look for cliques in the Grassmann graph $\mathcal{G}_2(m,n)$. The graph $\mathcal{G}_2(n-1,n)$ is considered in Subsection III-A. It is the simplest to understand, gives codes with minimum repair bandwidth, but unreasonable storage overhead. We then compute some other examples from other Grassmann graphs. A clique from the graph $\mathcal{G}_2(5,3)$ is computed in Subsection III-B, yielding a storage code with a slightly better overhead than the previous examples. A clique from the graph $\mathcal{G}_2(6,3)$ is reported in Subsection III-C, which offers different (collaborative) repair options.

II. System Model

We consider a storage network, composed of N storage nodes. Let $\mathbf{o} \in \mathbb{F}_q^B$ be a data object, represented as a row vector of length B with coefficients in the finite field $\mathbb{F}_q$, to be stored over this network. The object is stored using a linear erasure code, that is $\mathbf{o}$ is mapped to a codeword whose coefficients are stored over the storage network. Since the erasure code used for storage is linear, we will represent it as a family of vectors $\{v_j \in \mathbb{F}_q^B,\ j=1,\dots,n\}$, $n \geq B$. Every storage node then contains some codeword coefficients of the form $\mathbf{o}v_j^T$, for some $j \in I \subset \{1,\dots,n\}$. Since every node is enabled of computational power, it can compute linear combinations of the stored data, that is $\mathbf{o}\sum_{j\in I} a_j v_j^T$, $a_j \in \mathbb{F}_q$. This means that we can model the data stored at each node by a vector subspace

$$W_i = \langle v_j,\ j \in I_i \rangle \subset \mathbb{F}_q^n.$$

We assume that $\dim_{\mathbb{F}_q}(W_i) = \alpha$, $i=1,\dots,N$.

A. Collaborative Repair

Suppose that a repair process is triggered after t failures, thus t live nodes will start downloading coefficients $\mathbf{o}v_j^T$, β of them, each from d live nodes. Thus every node participating in the repair process obtains d subspaces

$$W_{rl} \leq W_l,\ l \in D_r,\ |D_r| = d,\ r = 1, \ldots, t.$$

The second index l tells the provenance of the subspace (the live node W_l), while the first index r tells which node is being repaired (without loss of generality, and to simplify the notation, we have reordered the nodes so that $W_1, \ldots, W_t$ are repaired). We assume that $\dim_{\mathbb{F}_q}(W_{rl}) = \beta$, for all r. There is no point for a node to download redundant data, thus we may assume that every of the t nodes each gets a subspace V_r, $r = 1, \ldots, t$, where

$$V_r = \oplus_{l \in D_r} W_{rl},\ |D_r| = d,$$

thus $\dim_{\mathbb{F}_q}(V_r) = d\beta$. Finally these t nodes exchange some more data among each other, say each of them will receive some subspaces V_{rl} each of dimension β', where $l \in T_r$ indicates the provenance of the data, and $|T_r| = T$ how many subspaces are received. Note that T may vary from 1 to $t-1$, but is the same among the nodes performing the repair. The case $t = 1$ corresponds to the repair of one node failure, when there is no collaboration, while $|T| = t - 1$ is the scenario studied in [2], [3], where every repair node exchanges data with the others. For a repair of t faults to be successful, it is necessary that

$$\dim(V_r) + \sum_{l \in T_r} \dim(V_{rl}) = \alpha.$$

We may indeed assume that the V_{rl} at one node are not intersecting, since there is no need to transfer redundant data.

Finally, the information stored across the network must be preserved through the repair process. In the case of exact repair, every of the t subspaces lost has been reconstructed, while for functional repair, the t subspaces generated during the collaborative process might be different from those lost, but the overall amount of information about the stored object stays the same.

Consider the case of exact repair. Then we must have

$$W_i = \langle v_j,\ j \in I_i \rangle = \langle V_i, \oplus_{l \in T_i} V_{il} \rangle = \langle \oplus_{l \in D_i} W_{il}, \oplus_{l \in T_i} V_{il} \rangle,$$

which forces the subspaces W_i to intersect in a specific manner. For example, if $t = 2$ nodes cooperate, the node repairing node 1 will receive V_{12} from the node repairing node 2, and send V_{21}. Thus after the cooperation phase, both nodes will intersect on $\langle V_{12}, V_{21} \rangle$, a subspace of dimension $2\beta'$.

B. Object Recovery

If needed, the data object $\mathbf{o}$ should be retrievable, despite the presence of potential node failures. We may want the constraint that $\mathbf{o}$ can be computed by contacting any choice of k out of the N storage nodes that store $\mathbf{o}$ (as in [2], [3]). This is not a necessary condition, one may alternatively prefer that $\mathbf{o}$ can be recover out of many sets of k storage nodes (as in [4]).

III. Some Examples of Constructions

Let V be an n-dimensional vector space over $\mathbb{F}_q$, for q a prime power.

Definition 1: [8, 9.3] The *Grassmann graph* $\mathcal{G}_q(n, m)$ of the m-subspaces of V has for vertex set the collection of linear subspaces of V of dimension m. Two vertices W, W' are adjacent whenever $\dim(W \cap W') = m - 1$, that is, W and W' intersect in a hyperplane.

Let $\begin{bmatrix} n \\ m \end{bmatrix}$ be the q-ary Gaussian binomial coefficient

$$\begin{bmatrix} n \\ m \end{bmatrix} = \frac{(q^n - 1) \cdots (q^{n-m+1} - 1)}{(q^m - 1) \cdots (q - 1)}. \tag{1}$$

The number of vertices of $\mathcal{G}_q(n, m)$ is $\begin{bmatrix} n \\ m \end{bmatrix}$, and every vertex has degree

$$q \frac{(q^{n-m} - 1)(q^m - 1)}{(q - 1)^2}. \tag{2}$$

We recall some well-known formulas about the dimension of sums of vector subspaces.

Lemma 1: Let W_1, W_2, W_3, W_4 be any m-dimensional subspaces of $\mathbb{F}_q^n$. Then

$$\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2). \tag{3}$$

Similarly for 3 subspaces

$$\begin{gathered} \dim(W_1 + W_2 + W_3) = \\ \textstyle\sum_{i=1}^3 \dim(W_i) - \dim(W_2 \cap W_3) - \dim(W_1 \cap (W_2 + W_3)) \end{gathered} \tag{4}$$

and for 4 subspaces:

$$\begin{gathered} \dim(W_1 + W_2 + W_3 + W_4) \\ = \textstyle\sum_{i=1}^4 \dim(W_i) - \dim(W_1 \cap W_2) - \dim(W_3 \cap W_4) \\ - \dim((W_1 + W_2) \cap (W_3 + W_4)). \end{gathered} \tag{5}$$

Proof: The first formula (3) is well-known. The second is obtained by applying it recursively:

$$\begin{aligned} & \dim(W_1 + W_2 + W_3) \\ = \; & \dim(W_1) + \dim(W_2 + W_3) - \dim(W_1 \cap (W_2 + W_3)) \\ = \; & \sum_{i=1}^3 \dim(W_i) - \dim(W_2 \cap W_3) - \dim(W_1 \cap (W_2 + W_3)) \end{aligned}$$

and so is (5):

$$\begin{aligned} & \dim(W_1 + W_2 + W_3 + W_4) \\ = \; & \dim(W_1 + W_2) + \dim(W_3 + W_4) \\ & - \dim((W_1 + W_2) \cap (W_3 + W_4)) \\ = \; & \sum_{i=1}^4 \dim(W_i) - \dim(W_1 \cap W_2) \\ & - \dim(W_3 \cap W_4) - \dim((W_1 + W_2) \cap (W_3 + W_4)). \end{aligned}$$

■

A. The Graph $\mathcal{G}_2(n, n-1)$

If $m = n - 1$, then from (1) the number of vertices of $\mathcal{G}_2(n, n-1)$ is

$$\frac{(2^n - 1)}{(2-1)} = 2^n - 1$$

and from (2) the degree of each vertex is

$$2\frac{(2^{n-1} - 1)}{(2-1)} = 2^n - 2$$

showing that the graph is complete, and any two subspaces W, W' are intersecting in a subspace of dimension $m - 1 = n - 2$. Now from (3)

$$\dim(W+W') = 2m - \dim(W \cap W') = 2n-2-(n-2) = n.$$

This corresponds to the case $k = 2$, where an object may be recovered from any two nodes. The repair of one failure can (of course) be done by contacting 2 live nodes (and 2 live nodes are needed). Indeed, if say W_1 needs to be repaired, contacting any node W_l allows to get W_{1l} of dimension $n-2$, and only one subspace of dimension 1 is missing, which can obtained from another node W_i, $i \neq l$: every W_i intersects W_1 in a subspace of dimension $n-2$, thus either $W_i \cap W_1 = W_{1l}$, in which case adding W_i does not allow to recover W_1, or $W_1 \subset \langle W_i, W_l \rangle$. For the latter to fail, it is needed that all subspaces intersect in the same subspace W_{1l}, which is not possible.

Example 1: The smallest such graph is $\mathcal{G}_2(3, 2)$. It is a complete graph with 7 vertices, given by

$$\begin{array}{ll} W_1 = \langle 100, 010 \rangle, & W_5 = \langle 101, 110 \rangle, \\ W_2 = \langle 100, 001 \rangle, & W_6 = \langle 010, 101 \rangle, \\ W_3 = \langle 100, 011 \rangle, & W_7 = \langle 010, 001 \rangle. \\ W_4 = \langle 110, 001 \rangle, & \end{array}$$

If $W_1 = \{100, 010, 110\}$ fails, 100 may be repaired by contacting W_2 or W_3, 110 by contacting W_4 or W_5, and 010 by contacting W_6 or W_7. There are then $\frac{1}{2}\binom{6}{1}\binom{4}{1} = 12$ ways of doing this repair, while the maximum would be $15 = \binom{6}{2}$ (for $d = 2$). This is true for each of the 7 nodes. The repair bandwidth reaches the minimum: two symbols downloaded to repair two, however a huge amount of storage is used: 14 symbols are stored, for a length 3 data object. This gives a storage overhead of $14/3 > 9/3$ which is the cost of 3-way replication.

It is possible to get a lesser storage overhead by reducing the length of the code, and take only 4 nodes, giving $8/3 < 9/3$. However then, 2 failures only can be tolerated.

Example 2: The graph $\mathcal{G}_2(4, 3)$ is a complete graph with 15 vertices.

$$\begin{array}{llll} W_1 & = \langle 1000, 0100, 0010 \rangle, & W_9 = \langle 0110, 0101, 1010 \rangle, \\ W_2 & = \langle 1110, 0001, 1000 \rangle, & W_{10} = \langle 0011, 1010, 1101 \rangle, \\ W_3 & = \langle 1111, 1000, 1100 \rangle, & W_{11} = \langle 1001, 1101, 0110 \rangle, \\ W_4 & = \langle 0111, 1100, 1110 \rangle, & W_{12} = \langle 0100, 0110, 0011 \rangle, \\ W_5 & = \langle 1011, 1110, 1111 \rangle, & W_{13} = \langle 0010, 0011, 1001 \rangle, \\ W_6 & = \langle 0101, 1111, 0111 \rangle, & W_{14} = \langle 0001, 1001, 0100 \rangle, \\ W_7 & = \langle 1010, 0111, 1011 \rangle, & W_{15} = \langle 1000, 0100, 0010 \rangle, \\ W_8 & = \langle 1101, 1011, 0101 \rangle, & \end{array}$$

The storage overhead of 3-way replication is $12/4 = 3$, thus we should keep at most 4 nodes to equate the amount of storage overhead, and 3 nodes to get less. This makes the length of the code too short, only two, respectively one failure(s) can then be tolerated.

This family of graphs clearly suffers from a terrible storage overhead of

$$\frac{(2^n - 1)m}{n}$$

if all the nodes are used. To number of nodes used should be (strictly) less than $3n/m$ to get a reasonable overhead, which in turn reduces significantly the number of failures tolerated. This overall behavior is likely to be caused by the fact that these subspaces share too big an intersection, though this in turn results in a minimum repair bandwidth.

B. The Graph $\mathcal{G}_2(2m-1, m)$

Consider a clique of the graph $\mathcal{G}_2(2m-1, m)$, such that every pair of subspaces intersect in a subspace of dimension 1. Then

$$2m - \dim(W \cap W') = n = 2m - 1$$

which shows that the object may be recovered from any choice of $k = 2$ nodes. When $m = 2$, we get the graph $\mathcal{G}_2(3, 2)$ already considered above. When $m = 3$, this is the graph $\mathcal{G}_2(5, 3)$.

Example 3: Consider the following (non-maximal) clique of $\mathcal{G}_2(5, 3)$, computed using *cliquer* [9]:

$$\begin{array}{ll} W_1 = \langle 10001, 01101, 00010 \rangle, & W_5 = \langle 10001, 01001, 00101 \rangle, \\ W_2 = \langle 10000, 01001, 00010 \rangle, & W_6 = \langle 10010, 00110, 00001 \rangle, \\ W_3 = \langle 11001, 00100, 00011 \rangle, & W_7 = \langle 10101, 01101, 00011 \rangle, \\ W_4 = \langle 10000, 01010, 00110 \rangle, & W_8 = \langle 01010, 00100, 00001 \rangle. \end{array}$$

It has the property that every pair of subspaces intersects in a subspace of dimension 1, and that every triple of subspaces has trivial intersection. Since $k = 2$ (the object is retrievable from any choice of 2 live nodes), we consider the repair of one failure. Suppose for example that $W_1 = \{10001, 01101, 00010, 11100, 10011, 01111, 11110\}$ fails. These vectors are available across the network as shown in Table I. To repair W_1, any two nodes may be contacted. Since the intersection of any 3 nodes is trivial, this will give necessarily two distinct vectors, which generate a subspace

vector	nodes	vector	nodes
10001	5	10011	6
01101	7	01111	8
00010	2	11110	3
11100	4		

TABLE I
VECTORS STORED AT NODE W_1 AND THEIR AVAILABILITY ACROSS THE NETWORK.

vector	nodes W_i, W_j, W_k
001110	1,3,5
110110	1,4,6
001101	2,3,6
000101	2,4,5

TABLE II
THE TRIPLE W_i, W_j, W_k WHOSE INTERSECTION IS OF DIMENSION 1, TOGETHER WITH THE VECTOR IN THIS INTERSECTION, ARE COMPUTED.

of dimension 2. Now the third vector can be anything, as long as it does not belong to the span of the vectors already obtained. Thus the number of ways of repairing W_1 is $\frac{1}{6}\binom{7}{1}\binom{6}{1}\binom{4}{1} = 28 < 35 = \binom{7}{3}$ (for $d = 3$).

The storage overhead of 3-way replication is $15/5 = 5$ so we may keep up to 5 nodes, which is slightly better than the code construction of Example 2.

C. The Graph $\mathcal{G}_2(6, 3)$

Consider the graph $\mathcal{G}_2(6, 3)$, and the following (non-maximal) clique, computed using cliquer [9]:

$$\begin{aligned} W_1 &= \langle 100010, 010100, 001110\rangle, \\ W_2 &= \langle 010001, 001000, 000101\rangle, \\ W_3 &= \langle 100101, 001101, 000011\rangle, \\ W_4 &= \langle 100000, 010011, 000101\rangle, \\ W_5 &= \langle 001010, 000100, 000001\rangle, \\ W_6 &= \langle 110110, 001100, 000001\rangle, \end{aligned}$$

Every pair of subspaces intersects in a subspace of dimension 1. By (4), for any W, W', W''

$$\dim(W + W' + W'') = 9 - 1 - \dim(W \cap (W' + W''))$$

and for this particular clique

$$\dim(W + W' + W'') = 6$$

which shows that any choice of 3 subspaces allows a data collector to retrieve the object. Some triples have an intersection of dimension 1, as summarized in Table II. Suppose the node W_1 fails. There are 4 repair options: (2,3,4), (2,3,6), (2,4,5) and (2,5,6), since $3, 5$ cannot be in a triple together, and $4, 6$ cannot either.

If two nodes fail, say W_1, W_2, then the node that repairs W_1 may get 001110 from W_5, 010100 from W_2 and 110110 from either W_4 or W_6. Then the node that repairs W_2 may get 001101 from W_2, 100101 from W_4 and either 001110 from W_5 or 001101 from W_6. So each has two repair options. A collaborative repair could also be done: once one node gets 001110, it may give it directly to the other repair node. The storage overhead is $18/6$ which is the same as 3-way replication.

IV. CONCLUSION

In this paper, we abstracted codes for distributed storage systems in terms of subspaces and their intersection. This suggested the design of subspaces with regular intersection, and we started with pairwise intersection. To find such subspaces, we computed cliques from Grassmannian graphs, to obtain families of subspaces whose pairwise distance has a given dimension, and studied the obtained parameters in terms of storage codes.

The choice of pairwise intersection is also natural, since it is related to the design of constant dimension codes for network coding [7]. However, though the examples that we found have some potential for storage applications, the requirement of pairwise intersection seems less critical than for network coding. There are obvious continuations of this preliminary study:

1) Find a theoretical characterization of (collaborative) repair in terms of subspace intersection.
2) Find more systematic constructions of such codes, to get instances with interesting parameters for storage applications.
3) Move from pairwise intersection to other types of intersection patterns.

ACKNOWLEDGMENT

This work is supported in part by a grant from the company Xvid, and by the MoE Tier-2 grant "eCODE: Erasure Codes for Datacenter Environments".

REFERENCES

[1] C. Huang, M. Chen, and J. Li, "Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems, in *6th IEEE International Symposium on Network Computing and Applications, NCA 2007*.

[2] A.-M. Kermarrec, N. Le Scouarnec, and G. Straub, "Repairing multiple failures with coordinated and adaptive regenerating codes, in *International Symposium on Network Coding (NetCod) 2011*.

[3] K. Shum, "Cooperative regenerating codes for distributed storage systems, in *2011 IEEE International Conference on Communications (ICC)*.

[4] F. Oggier, A. Datta, "Self-repairing Homomorphic Codes for Distributed Storage Systems", in the proceedings of *The 30th IEEE International Conference on Computer Communications (INFOCOM) 2011*.

[5] F. Oggier, A. Datta. "Coding Techniques for Repairability in Networked Distributed Storage Systems", in the series *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 4, June 2013.

[6] H. D. L. Hollmann, "Storage codes – coding rate and repair locality", in the proceedings of *the International Conference on Computing, Networking and Communications (ICNC) 2013*.

[7] R. Koetter, F. R. Kschischang, "Coding for Errors and Erasures in Random Network Coding", *IEEE Transactions on Information Theory*, 54 (8), (2008).

[8] A. E. Brouwer, A. M. Cohen, A. Neumaier, "Distance-Regular Graphs", *Springer*, 1983.

[9] S. Niskanen, P. R. J. Östergård, "Cliquer User's Guide, Version 1.0," Communications Laboratory, Helsinki University of Technology, Espoo, Finland, Tech. Rep. T48, 2003.

Lattice Gaussian Coding for Capacity and Secrecy: Two Sides of One Coin

(Invited Paper)

Cong Ling
Department of Electrical and Electronic Engineering
Imperial College London
London, UK
Email: c.ling@imperial.ac.uk

Jean-Claude Belfiore
Department of Communications and Electronics
Telecom ParisTech
Paris, France
Email: belfiore@telecom-paristech.fr

***Abstract*—Based on the lattice Gaussian distribution and the associated flatness factor, we present a unified view of lattice coding for achieving the Shannon capacity of the additive white Gaussian noise (AWGN) channel and for approaching the secrecy capacity of the Gaussian wiretap channel. In the former scenario, we apply Gaussian shaping to an AWGN-good lattice; in the latter scenario, we use a secrecy-good lattice nested with an AWGN-good lattice. We show that they represent different aspects of the lattice Gaussian distribution.**

I. Introduction

The lattice Gaussian distribution is emerging as a common theme in diverse areas. In mathematics, Banaszczyk [1] firstly used it to prove the transference theorems of lattices. In cryptography, Micciancio and Regev used it to propose lattice-based cryptosystems based on the worst-case hardness assumptions [2], and recently, it has underpinned the fully-homomorphic encryption for cloud computing [3]. In communications, Forney applied the lattice Gaussian distribution to shaping of lattice codes [4] (see also [5]), and studied lattice-aliased Gaussian noise in [6].

More recently, we defined the *flatness factor* associated with the lattice Gaussian distribution and derived its many properties [7, 8]. With this new tool, we are now able to answer/address several major open questions in lattice coding. For example, Erez and Zamir [9] proposed nested lattice codes achieving the capacity of the power-constrained additive white Gaussian noise (AWGN) channel, where a quantization-good lattice serves as the shaping lattice while the AWGN-good lattice serves as the coding lattice (dithering is also required). In [8], we proposed *lattice Gaussian coding*, where the codebook has a discrete Gaussian distribution over an AWGN-good lattice. As another example, in [7] we used the lattice Gaussian distribution to achieve *semantic security* over the Gaussian wiretap channel, which led to the notion of *secrecy-good lattices*. In both cases, we do not need a shaping lattice or a dither.

In this review paper, we aim to present a unified view of lattice Gaussian coding for capacity and secrecy. In Section II, we review lattice Gaussian distributions and the flatness factor. Section III describes the lattice Gaussian coding scheme for the AWGN channel. Section IV gives the scheme for the Gaussian wiretap channel, where the fine code is a Gaussian-shaped AWGN-good lattice achieving the capacity of the legitimate channel, and the coarse code is a secrecy-good lattice which ensures the information leakage on the eavesdropper's channel is negligible. We try to shed light on the commonality of the schemes for capacity and for secrecy [7, 8].

Throughout this paper, we use the natural logarithm, denoted by log, and information is measured in nats.

II. Lattice Gaussian Distribution and Flatness Factor

An n-dimensional lattice Λ in the Euclidean space $\mathbb{R}^n$ is a set defined by

$$\Lambda = \mathcal{L}(\mathbf{B}) = \{\mathbf{B}\mathbf{x} : \mathbf{x} \in \mathbb{Z}^n\}$$

where $\mathbf{B}$ is the n-by-n generator matrix. The dual lattice Λ^* of a lattice Λ is defined as the set of vectors $\mathbf{v} \in \mathbb{R}^n$ such that $\langle \mathbf{v}, \boldsymbol{\lambda} \rangle \in \mathbb{Z}$, for all $\boldsymbol{\lambda} \in \Lambda$.

For $\sigma > 0$ and $\mathbf{c} \in \mathbb{R}^n$, the usual Gaussian distribution of variance σ^2 centered at $\mathbf{c} \in \mathbb{R}^n$ is given by

$$f_{\sigma,\mathbf{c}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\|\mathbf{x}-\mathbf{c}\|^2}{2\sigma^2}},$$

for all $\mathbf{x} \in \mathbb{R}^n$. For convenience, we write $f_\sigma(\mathbf{x}) = f_{\sigma,\mathbf{0}}(\mathbf{x})$.

Consider the Λ-periodic function (see Fig. 1(a))

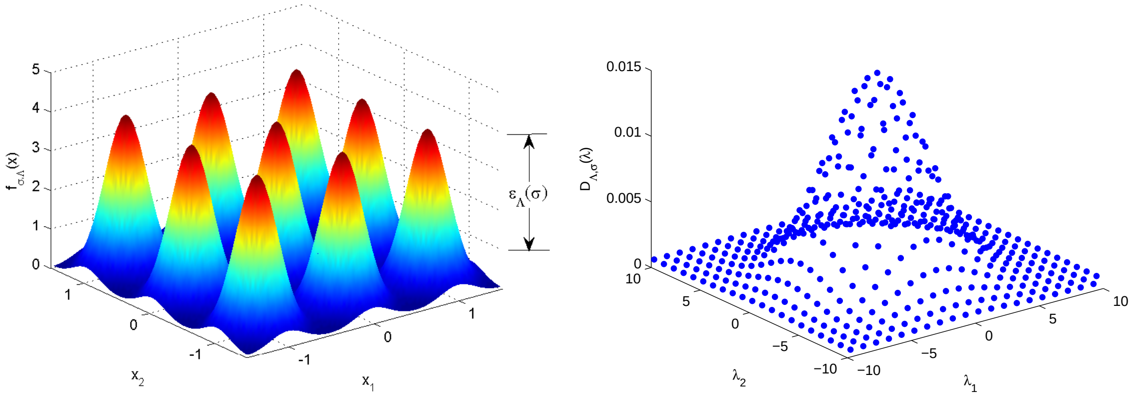
$$f_{\sigma,\Lambda}(\mathbf{x}) = \sum_{\boldsymbol{\lambda}\in\Lambda} f_{\sigma,\boldsymbol{\lambda}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \sum_{\boldsymbol{\lambda}\in\Lambda} e^{-\frac{\|\mathbf{x}-\boldsymbol{\lambda}\|^2}{2\sigma^2}}, \quad (1)$$

for all $\mathbf{x} \in \mathbb{R}^n$. Observe that $f_{\sigma,\Lambda}$ restricted to the fundamental region $\mathcal{R}(\Lambda)$ is a probability density.

We define the *discrete Gaussian distribution* over Λ centered at $\mathbf{c} \in \mathbb{R}^n$ as the following discrete distribution taking values in $\boldsymbol{\lambda} \in \Lambda$:

$$D_{\Lambda,\sigma,\mathbf{c}}(\boldsymbol{\lambda}) = \frac{f_{\sigma,\mathbf{c}}(\boldsymbol{\lambda})}{f_{\sigma,\mathbf{c}}(\Lambda)}, \quad \forall \boldsymbol{\lambda} \in \Lambda,$$

where $f_{\sigma,\mathbf{c}}(\Lambda) \triangleq \sum_{\boldsymbol{\lambda}\in\Lambda} f_{\sigma,\mathbf{c}}(\boldsymbol{\lambda}) = f_{\sigma,\Lambda}(\mathbf{c})$. Again for convenience, we write $D_{\Lambda,\sigma} = D_{\Lambda,\sigma,\mathbf{0}}$. Fig. 1(b) illustrates the discrete Gaussian distribution over $\mathbb{Z}^2$. As can be seen, it resembles a continuous Gaussian distribution, but is only defined over a lattice.



(a) Continuous periodic distribution $f_{\sigma,\Lambda}(\mathbf{x})$.

(b) Discrete Gaussian distribution $D_{\Lambda,\sigma}(\boldsymbol{\lambda})$.

Fig. 1. Lattice Gaussian distributions.

In some sense, the continuous distribution $f_{\sigma,\Lambda}$ and the discrete distribution $D_{\Lambda,\sigma}$ are the Fourier dual of each other. To see this, note that since $f_{\sigma,\Lambda}(\mathbf{x})$ is Λ-periodic, it has the Fourier expansion on the dual lattice Λ^*

$$f_{\sigma,\Lambda}(\mathbf{x}) = \frac{1}{V(\Lambda)} \sum_{\boldsymbol{\lambda}^* \in \Lambda^*} \hat{f}_\sigma(\boldsymbol{\lambda}^*) e^{j2\pi\langle \boldsymbol{\lambda}^*, \mathbf{x}\rangle}$$

where

$$\hat{f}_\sigma(\mathbf{y}) = \int f_\sigma(\mathbf{x}) e^{-j2\pi\langle \mathbf{x},\mathbf{y}\rangle} d\mathbf{x} = e^{-2\pi^2\sigma^2\|\mathbf{y}\|^2} \tag{2}$$

is the Fourier transform. Thus, the Fourier coefficients $\hat{f}_\sigma(\boldsymbol{\lambda}^*)$ have a discrete Gaussian distribution over the dual lattice Λ^* (upon normalization).

The flatness factor of a lattice Λ quantifies the maximum variation of $f_{\sigma,\Lambda}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$.

Definition 1 (Flatness factor [7])**.** *For a lattice Λ and for a parameter σ, the flatness factor is defined by:*

$$\epsilon_\Lambda(\sigma) \triangleq \max_{\mathbf{x}\in\mathcal{R}(\Lambda)} |V(\Lambda) f_{\sigma,\Lambda}(\mathbf{x}) - 1|.$$

In other words, $\frac{f_{\sigma,\Lambda}(\mathbf{x})}{1/V(\Lambda)}$, the ratio between $f_{\sigma,\Lambda}(\mathbf{x})$ and the uniform distribution over $\mathcal{R}(\Lambda)$, is within the range $[1-\epsilon_\Lambda(\sigma), 1+\epsilon_\Lambda(\sigma)]$.

Proposition 1 (Expression of $\epsilon_\Lambda(\sigma)$ [7])**.** *We have:*

$$\epsilon_\Lambda(\sigma) = \left(\frac{\gamma_\Lambda(\sigma)}{2\pi}\right)^{\frac{n}{2}} \Theta_\Lambda\left(\frac{1}{2\pi\sigma^2}\right) - 1$$

where $\gamma_\Lambda(\sigma) = \frac{V(\Lambda)^{\frac{2}{n}}}{\sigma^2}$ is the volume-to-noise ratio (VNR).

The following result guarantees the existence of sequences of mod-p lattices whose flatness factors can vanish as $n \to \infty$.

Theorem 1 ([7])**.** *$\forall \sigma > 0$ and $\forall \delta > 0$, there exists a sequence of mod-p lattices $\Lambda^{(n)}$ such that*

$$\epsilon_{\Lambda^{(n)}}(\sigma) \le (1+\delta)\cdot\left(\frac{\gamma_{\Lambda^{(n)}}(\sigma)}{2\pi}\right)^{\frac{n}{2}}, \tag{3}$$

i.e., the flatness factor can go to zero exponentially for any fixed VNR $\gamma_{\Lambda^{(n)}}(\sigma) < 2\pi$.

The importance of a small flatness factor is two-fold. Firstly, it assures the “folded" distribution $f_{\sigma,\Lambda}(\mathbf{x})$ is flat; secondly, it implies the discrete Gaussian distribution $D_{\Lambda,\sigma,\mathbf{c}}$ is “smooth". In the following, we collect properties of lattice Gaussian distributions.

Lemma 1 ([7])**.** *Let $\Lambda' \subset \Lambda$ be a pair of nested lattices such that $\epsilon_{\Lambda'}(\sigma) < \frac{1}{2}$. If $\mathbf{a}$ is uniformly distributed in Λ/Λ' and $\mathbf{b}$ is sampled from $D_{\Lambda',\sigma,\mathbf{c}-\mathbf{a}}$, then the distribution $D_{\mathbf{a}+\mathbf{b}}$ of $\mathbf{a}+\mathbf{b}$ satisfies*

$$\mathbb{V}(D_{\mathbf{a}+\mathbf{b}}, D_{\Lambda,\sigma,\mathbf{c}}) \le \frac{2\epsilon_{\Lambda'}(\sigma)}{1-\epsilon_{\Lambda'}(\sigma)}.$$

Lemma 2 (Variance of lattice Gaussian [7])**.** *Let $\mathbf{x} \sim D_{\Lambda,\sigma,\mathbf{c}}$. If $\varepsilon = \epsilon_\Lambda\left(\sigma/\sqrt{\frac{\pi}{\pi-t}}\right) < 1$ for $0 < t < \pi$, then*

$$\left|\mathbb{E}\left[\|\mathbf{x}-\mathbf{c}\|^2\right] - n\sigma^2\right| \le \frac{2\pi\varepsilon_t}{1-\varepsilon}\sigma^2$$

where

$$\varepsilon_t \triangleq \begin{cases} \varepsilon, & t \ge 1/e; \\ (t^{-4}+1)\varepsilon, & 0 < t < 1/e. \end{cases}$$

Lemma 3 (Entropy of lattice Gaussian [7])**.** *Let $\mathbf{x} \sim D_{\Lambda,\sigma,\mathbf{c}}$. If $\varepsilon = \epsilon_\Lambda\left(\sigma/\sqrt{\frac{\pi}{\pi-t}}\right) < 1$ for $0 < t < \pi$, then the entropy rate of $\mathbf{x}$ satisfies*

$$\left|\frac{1}{n}\mathbb{H}(\mathbf{x}) - \left[\log(\sqrt{2\pi e}\sigma) - \frac{1}{n}\log V(\Lambda)\right]\right| \le \varepsilon',$$

where $\varepsilon' = -\frac{\log(1-\varepsilon)}{n} + \frac{\pi\varepsilon_t}{n(1-\varepsilon)}$.

Lemma 4 ([10])**.** *Given any vector $\mathbf{c} \in \mathbb{R}^n$, and $\sigma_s, \sigma > 0$. Let $\tilde{\sigma} \triangleq \frac{\sigma_s\sigma}{\sqrt{\sigma_s^2+\sigma^2}}$ and let $\sigma_s' = \sqrt{\sigma_s^2+\sigma^2}$. Consider the continuous distribution g on $\mathbb{R}^n$ obtained by adding a continuous Gaussian of variance σ^2 to a discrete Gaussian $D_{\Lambda-\mathbf{c},\sigma_s}$:*

$$g(\mathbf{x}) = \frac{1}{f_{\sigma_s}(\Lambda-\mathbf{c})} \sum_{\mathbf{t}\in\Lambda-\mathbf{c}} f_{\sigma_s}(\mathbf{t}) f_\sigma(\mathbf{x}-\mathbf{t}), \quad \mathbf{x} \in \mathbb{R}^n.$$

If $\varepsilon = \epsilon_\Lambda\left(\tilde{\sigma}\right) < \frac{1}{2}$, *then* $\frac{g(\mathbf{x})}{f_{\sigma'_s}(\mathbf{x})}$ *is uniformly close to* 1*:*

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \left| \frac{g(\mathbf{x})}{f_{\sigma'_s}(\mathbf{x})} - 1 \right| \leq 4\varepsilon. \tag{4}$$

Regev's lemma leads to an important property, namely, the discrete Gaussian distribution over a lattice is almost capacity-achieving if the flatness factor is small [8].

III. Achieving Channel Capacity

Consider the classic AWGN channel

$$\mathsf{Y}^n = \mathsf{X}^n + \mathsf{W}^n$$

where W^n is an n-dimensional Gaussian noise vector with zero mean and variance σ_w^2.

In [8], we proposed a new coding scheme based on the lattice Gaussian distribution with power constraint P. The SNR is defined by $\mathsf{SNR} = P/\sigma_w^2$. Let Λ be an AWGN-good lattice of dimension n. The encoder maps the information bits to points in Λ, which obey the lattice Gaussian distribution (cf. Fig. 1(b))

$$\mathbf{x} \sim D_{\Lambda,\sigma_s}.$$

Since the continuous Gaussian distribution is capacity-achieving, we want the lattice Gaussian distribution to behave like the continuous Gaussian distribution (in particular $P \approx \sigma_s^2$). This can be assured by a small flatness factor $\epsilon_\Lambda\left(\sigma_s/\sqrt{\frac{\pi}{\pi-t}}\right)$ for $0 < t < \pi$. For $t \to 0$, this condition is essentially $\epsilon_\Lambda\left(\sigma_s\right) \to 0$. Thus, while we are concerned with the discrete distribution D_{Λ,σ_s}, we in fact require the associated periodic distribution $f_{\sigma_s,\Lambda}$ to be flat.

Since the lattice points are not equally probable a priori in the lattice Gaussian coding, we will use maximum-a-posteriori (MAP) decoding. In [7], it was shown that MAP decoding is equivalent to Euclidean lattice decoding of Λ using a scaling coefficient $\alpha = \frac{\sigma_s^2}{\sigma_s^2+\sigma_w^2}$, which is asymptotically equal to the MMSE coefficient $\frac{P}{P+\sigma_w^2}$. In fact, the error probability of the proposed scheme under MMSE lattice decoding admits almost the same expression as that of Poltyrev [11], with σ_w replaced by $\tilde{\sigma}_w = \frac{\sigma_s\sigma_w}{\sqrt{\sigma_s^2+\sigma_w^2}}$. To satisfy the sphere bound, we choose the fundamental volume $V(\Lambda)$ such that

$$V(\Lambda)^{2/n} > 2\pi e \tilde{\sigma}_w^2. \tag{5}$$

Meanwhile, the rate of the scheme is given by the entropy of the lattice Gaussian distribution. By Lemma 3, we have

$$\begin{aligned} R &\to \log(\sqrt{2\pi e}\sigma_s) - \frac{1}{n}\log V(\Lambda) \\ &< \log(\sqrt{2\pi e}\sigma_s) - \frac{1}{2}\log\left(2\pi e\frac{\sigma_s^2\sigma_w^2}{\sigma_s^2+\sigma_w^2}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{\sigma_s^2}{\sigma_w^2}\right) \\ &\to \frac{1}{2}\log\left(1 + \mathsf{SNR}\right). \end{aligned}$$

In fact, the rate can be arbitrarily close to the channel capacity. A more careful analysis also shows that the condition $\mathsf{SNR} > e$ is needed.

Theorem 2 (Coding theorem for lattice Gaussian coding [8])**.** *Consider a lattice code whose codewords are drawn from the discrete Gaussian distribution* D_{Λ,σ_s} *for an AWGN-good lattice* Λ*. If* $\mathsf{SNR} > e$*, then any rate up to the channel capacity* $\frac{1}{2}\log\left(1 + \mathsf{SNR}\right)$ *is achievable, while the error probability of MMSE lattice decoding vanishes exponentially fast.*

IV. Approaching Secrecy Capacity

Now consider the Gaussian wiretap channel where Alice and Bob are the legitimate users, while Eve is an eavesdropper. The outputs Y^n and Z^n at Bob and Eve's ends respectively are given by

$$\begin{cases} \mathsf{Y}^n = \mathsf{X}^n + \mathsf{W}_b^n, \\ \mathsf{Z}^n = \mathsf{X}^n + \mathsf{W}_e^n, \end{cases} \tag{6}$$

where W_b^n, W_e^n are n-dimensional Gaussian noise vectors with zero mean and variance σ_b^2, σ_e^2 respectively.

For secrecy rate R, we use coset coding induced by a lattice partition $\Lambda_e \subset \Lambda_b$ such that

$$\frac{1}{n}\log\left|\Lambda_b/\Lambda_e\right| = R.$$

The fine lattice Λ_b is the usual coding lattice for Bob, i.e., it is an AWGN-good lattice. The coarse lattice Λ_e is new, and turns out to be a secrecy-good lattice. To encode, Alice uses the secret bits to select one coset of Λ_e and transmits a random point inside this coset.

Let us discuss intuitively why this scheme is secure. Informally, given message m, Alice samples a lattice point uniformly at random from a coset $\Lambda_e + \boldsymbol{\lambda}_m$ (this corresponds to Poltyrev's setting of infinite lattice coding [11]). Due to the channel noise, Eve observes the periodic distribution

$$\frac{1}{(\sqrt{2\pi}\sigma_e)^n} \sum_{\boldsymbol{\lambda} \in \Lambda + \boldsymbol{\lambda}_m} e^{-\frac{\|\mathbf{z}-\boldsymbol{\lambda}\|^2}{2\sigma_e^2}}.$$

If the flatness factor $\epsilon_{\Lambda_e}(\sigma_e)$ is small, it will be close to a uniform distribution, regardless of message m. Then Eve would not be able to distinguish which message Alice has sent. With a careful design of Λ_e, this is possible, because Eve's channel is noisier. Of course, the technical difficulty here is that one cannot really sample a lattice point uniformly from a lattice or its coset.

Now we describe the wiretap coding scheme more formally. Consider a message set $\mathcal{M}_n = \{1, \ldots, e^{nR}\}$, and a one-to-one function $\phi : \mathcal{M}_n \to \Lambda_b/\Lambda_e$ which associates each message $m \in \mathcal{M}_n$ to a coset $\widetilde{\boldsymbol{\lambda}}_m \in \Lambda_b/\Lambda_e$. One could choose the coset representative $\boldsymbol{\lambda}_m \in \Lambda_b \cap \mathcal{R}(\Lambda_e)$ for any fundamental region $\mathcal{R}(\Lambda_e)$. In order to encode the message $m \in \mathcal{M}_n$, Alice actually samples X_m^n from lattice Gaussian distribution

$$\mathsf{X}_m^n \sim D_{\Lambda_e + \boldsymbol{\lambda}_m, \sigma_s}.$$

equivalently, Alice transmits $\boldsymbol{\lambda}+\boldsymbol{\lambda}_m$ where $\boldsymbol{\lambda} \sim D_{\Lambda_e,\sigma_s,-\boldsymbol{\lambda}_m}$. Let $\tilde{\sigma}_e = \frac{\sigma_s\sigma_e}{\sqrt{\sigma_s^2+\sigma_e^2}}$ and $\sigma_s' = \sqrt{\sigma_s^2+\sigma_e^2}$. Regev's Lemma 4 implies that if $\epsilon_{\Lambda_e}(\tilde{\sigma}_e) < \frac{1}{2}$, then:

$$\mathbb{V}\left(p_{\mathsf{Z}^n|\mathsf{M}}(\cdot|m), f_{\sigma_s'}\right) \leq 4\epsilon_{\Lambda_e}(\tilde{\sigma}_e).$$

We see that the received signals converge to the same Gaussian distribution $f_{\sigma_s'}$. This already gives *distinguishing security*, which means that, asymptotically, the channel outputs are indistinguishable for different input messages.

An upper bound on the amount of leaked information then follows.

Theorem 3 (Information leakage [7]). *Suppose that the wiretap coding scheme described above is employed on the Gaussian wiretap channel (6), and let $\varepsilon_n = \epsilon_{\Lambda_e}(\tilde{\sigma}_e)$. Assume that $\varepsilon_n < \frac{1}{2}$ for all n. Then the mutual information between the confidential message and the eavesdropper's signal is bounded as follows:*

$$\mathbb{I}(\mathsf{M};\mathsf{Z}^n) \leq 8\varepsilon_n nR - 8\varepsilon_n \log 8\varepsilon_n. \tag{7}$$

A wiretap coding scheme is secure in the sense of *strong secrecy* if $\lim_{n\to\infty}\mathbb{I}(\mathsf{M};\mathsf{Z}^n) = 0$. From (7), a flatness factor $\varepsilon_n = o(\frac{1}{n})$ would be enough. In practice, an exponential decay of the information leakage is desired, and this motivates the notion of secrecy-good lattices:

Definition 2 (Secrecy-good lattices). *A sequence of lattices $\Lambda^{(n)}$ is* secrecy-good *if*

$$\epsilon_{\Lambda^{(n)}}(\sigma) = e^{-\Omega(n)}, \quad \forall \gamma_{\Lambda^{(n)}}(\sigma) < 2\pi. \tag{8}$$

In the notion of strong secrecy, plaintext messages are often assumed to be random and uniformly distributed in $\mathcal{M}$. This assumption is deemed problematic from the cryptographic perspective, since in many setups plaintext messages are not random. This issue can be resolved by using the standard notion of *semantic security* [12] which means that, asymptotically, it is impossible to estimate any function of the message better than to guess it without considering Z^n at all. The relation between strong secrecy and semantic security was recently revealed in [7, 13], namely, achieving strong secrecy for all distributions of the plaintext messages is equivalent to achieving semantic security. Since in our scheme we make no *a priori* assumption on the distribution of m, it achieves semantic security.

It was shown in [7] that, under mild conditions, the secrecy rate

$$R < \frac{1}{2}\log(1+\mathsf{SNR}_b) - \frac{1}{2}\log(1+\mathsf{SNR}_e) - \frac{1}{2} \tag{9}$$

is achievable, which is within a half nat from the secrecy capacity.

Lastly, let us scrutinize the distribution of Alice's constellation. For this purpose only, we assume the confidential message $\boldsymbol{\lambda}_m \in [\Lambda_b/\Lambda_e]$ is uniformly distributed (or the secrecy rate will be smaller). By Lemma 1, if $\epsilon_{\Lambda_e}(\sigma_s) \leq \varepsilon$ (which we trivially have, since even $\epsilon_{\Lambda_e}(\tilde{\sigma}_e) \to 0$), then

$$\mathbb{V}(p_{\mathsf{X}^n}, D_{\Lambda_b,\sigma_s}) \leq \frac{2\varepsilon}{1-\varepsilon}.$$

Namely, the density p_{X^n} is close to the discrete Gaussian distribution over Λ_b. This shows that in fact, the fine code is capacity-achieving for Bob's channel. In contrast, from (9), we know that the coarse code has a rate $> \frac{1}{2}\log(1+\mathsf{SNR}_e)+\frac{1}{2}$, i.e., above the capacity of Eve's channel.

V. Discussion

In this paper, we have demonstrated the applications of the lattice Gaussian distribution to coding problems for the AWGN channel and the Gaussian wiretap channel. For capacity it is desired that the discrete Gaussian distribution of the lattice codebook behaves like the continuous Gaussian distribution, while for secrecy it is required that the aliased Gaussian distribution of the noise becomes flat. Both scenarios demand a vanishing flatness factor and thus can be viewed as two sides of one coin.

Acknowledgment

This work was supported in part by FP7 project PHYLAWS (EU FP7-ICT 317562).

References

[1] W. Banaszczyk, "New bounds in some transference theorems in the geometry of numbers," *Math. Ann.*, vol. 296, pp. 625–635, 1993.

[2] D. Micciancio and O. Regev, "Worst-case to average-case reductions based on Gaussian measures," in *Proc. Ann. Symp. Found. Computer Science*, Rome, Italy, Oct. 2004, pp. 372–381.

[3] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. STOC*, 2009.

[4] G. Forney and L.-F. Wei, "Multidimensional constellations–Part II: Voronoi constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 941–958, Aug 1989.

[5] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 39, pp. 913–929, May 1993.

[6] G. Forney, M. Trott, and S.-Y. Chung, "Sphere-bound-achieving coset codes and multilevel coset codes," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 820–850, May 2000.

[7] C. Ling, L. Luzzi, J.-C. Belfiore, and D. Stehlé, "Semantically secure lattice codes for the Gaussian wiretap channel," submitted to IEEE Trans. Inform. Theory, Oct. 2012, revised, Oct. 2013. [Online]. Available: http://arxiv.org/abs/1210.6673

[8] C. Ling and J.-C. Belfiore, "Achieiving the AWGN channel capacity with lattice Gaussian coding," submitted to IEEE Trans. Inform. Theory, Mar. 2012, revised, Nov. 2013. [Online]. Available: http://arxiv.org/abs/1302.5906

[9] U. Erez and R. Zamir, "Achieving $\frac{1}{2}$ log(1+SNR) on the AWGN channel with lattice encoding and decoding," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2293–2314, Oct. 2004.

[10] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," *J. ACM*, vol. 56, no. 6, pp. 34:1–34:40, 2009.

[11] G. Poltyrev, "On coding without restrictions for the AWGN channel," *IEEE Trans. Inform. Theory*, vol. 40, pp. 409–417, Mar. 1994.

[12] S. Goldwasser and S. Micali, "Probabilistic encryption," *J. Comput. Syst. Sci.*, vol. 28, no. 2, pp. 270–299, 1984.

[13] M. Bellare, S. Tessaro, and A. Vardy, "Semantic security for the wiretap channel," in *Proc. CRYPTO 2012*, ser. Lecture Notes in Computer Science, vol. 7417. Springer-Verlag, pp. 294–311.

Measuring the growth of inverse determinants sums of a family of quasi-orthogonal codes

Roope Vehkalahti
Department of mathematics, University of Turku
Finland
Email: roiive@utu.fi

Laura Luzzi
Laboratoire ETIS, CNRS - ENSEA - UCP
Cergy-Pontoise, France
laura.luzzi@ensea.fr

Abstract—**Inverse determinant sums appear naturally as a tool for analyzing performance of space-time codes in Rayleigh fading channels. This work will analyze the growth of inverse determinant sums of a family of quasi-orthogonal codes and will show that the growths are in logarithmic class. This is considerably lower than that of comparable number field codes.**

I. Introduction

In [5] inverse determinant sums were proposed as a tool to analyze the performance of algebraic space-time codes for MIMO fading channels. These sums can be seen as a generalization of the theta series for the Gaussian channel. They arise naturally from the union bound on the pairwise error probability for spherical constellations, but also in the analysis of fading wiretap channels [4].

In [5] the authors analyzed the growth of the inverse determinant sums of diagonal number field codes and of most well known division algebra codes. In this work we are going to extend the analysis to a large class of quasi-orthogonal codes. Our work will reveal that the growth of inverse determinant sums of the analyzed codes is considerably smaller than that of the corresponding diagonal number field codes. This difference suggest that asymptotically, with growing constellation, quasi-orthogonal codes are considerably better than number field codes. This difference can not be captured in the framework of diversity-multiplexing gain tradeoff.

For related work, we refer the reader to [1] and [3].

II. Inverse determinant sum

We begin by providing basic definitions concerning matrix lattices and spherical constellations, that are needed in the sequel.

A. Matrix lattices and spherically shaped coding schemes

Definition 2.1: A *space-time lattice code* $C \subseteq M_n(\mathbb{C})$ has the form

$$\mathbb{Z}B_1 \oplus \mathbb{Z}B_2 \oplus \cdots \oplus \mathbb{Z}B_k,$$

where the matrices $B_1, \ldots, B_k$ are linearly independent over $\mathbb{R}$, *i.e.*, form a lattice basis, and k is called the *rank* or the *dimension* of the lattice.

Definition 2.2: If the minimum determinant of the lattice $L \subseteq M_n(\mathbb{C})$ is non-zero, i.e. it satisfies

$$\inf_{\mathbf{0} \neq X \in L} |\det(X)| > 0,$$

we say that the code has a *non-vanishing determinant* (NVD).

We now consider a coding scheme based on a k-dimensional lattice L inside $M_n(\mathbb{C})$. For a given positive real number M we define the finite code

$$L(M) = \{a \mid a \in L, \|a\|_F \leq M\},$$

where $\|a\|_F$ refers to the Frobenius norm. In the following we will also use the notation

$$B(M) = \{a \mid a \in M_n(\mathbb{C}),\ \|a\|_F \leq M\},$$

for the sphere with radius M.

Let $L \subseteq M_n(\mathbb{C})$ be a k-dimensional lattice. For any fixed $m \in \mathbb{Z}^+$ we define

$$S_L^m(M) := \sum_{X \in L(M) \setminus \{\mathbf{0}\}} \frac{1}{|\det(X)|^m}.$$

Our main goal is to study the growth of this sum as M increases. Note, however, that in order to have a fair comparison between two different space-time codes, these should be normalized to have the same average energy. Namely, the volume $\mathrm{Vol}(L)$ of the fundamental parallelotope of each lattice L should be normalized to 1. The normalized version of the inverse determinant sums problem is then to consider the growth of the sum

$$\tilde{S}_L^m(M) = \mathrm{Vol}(L)^{mn/k} S_L^m(M), \qquad (1)$$

where k is the real dimension of the lattice Λ.

B. Inverse determinant sums and error performance of space-time lattice codes

Let us now consider the slow Rayleigh fading MIMO channel with n transmit and n_r receive antennas. The channel equation can then be written as

$$Y \ = \ HX + N,$$

where H and N are respectively the channel and noise matrices. We suppose that the transmitted codeword X belongs to a finite code $L(M) \subset M_n(\mathbb{C})$ carved from a k-dimensional

NVD lattice L as defined previously. In terms of pairwise error probability, we have for $X \neq X'$

$$P(X \to X') \leq \frac{1}{|\det(X - X')|^{2n_r}},$$

and the corresponding upper bound on overall error probability

$$P_e \ \leq \sum_{X \in L,\, 0 < \|X\|_F \leq 2M} \frac{1}{|\det(X)|^{2n_r}}.$$

Our main goal is now to study the growth of the sum $S_L^m(M)$ as M increases. In particular, we want to find, if possible, a function $f(M)$ such that

$$S_L^m(M) \sim f(M).$$

III. Inverse determinant sums of algebraic number field codes

In this section we will give and review some results concerning inverse determinant sums of diagonal number field codes. These results will play an important role in our analysis of quasi-orthogonal codes. The proofs are analogous to those given in [5] and we will skip them. Unlike in the rest of the paper we will state the results in the normalized from $\tilde{S}_L^m(M)$ following the general normalization given in [**?**].

A. Inverse determinant sums of real diagonal number field codes

Let K be a totally real number field of degree n and let $\{\sigma_1, \cdots, \sigma_n\}$ be the $\mathbb{Q}$-embeddings from K to $\mathbb{R}$. We then have the canonical embedding $\psi : K \mapsto M_n(\mathbb{R})$ defined by

$$\psi(x) = \mathrm{diag}(\sigma_1(x), \ldots, \sigma_n(x)).$$

It is a well known result that $\psi(\mathcal{O}_K)$ is an n-dimensional NVD lattice in $M_n(\mathbb{R})$. Let us now consider the corresponding inverse determinant sum. The main role in the analysis is played by the following unit group density result.

Theorem 3.1 ([6]): Let us suppose that $[K : \mathbb{Q}] = n$, we then have that

$$|\psi(\mathcal{O}_K^*) \cap B(M)| = N(\log M)^{n-1} + O((\log M)^{n-2}),$$

where $N = \frac{\omega n^{n-1}}{i_K R(n-1)!}$.
Here R is the regulator of the number field K, ω the number of roots of unity in K and i_K the index of norm 1 units in $\mathcal{O}_K^*$.

Proposition 3.2: Let us suppose that K is a totally real number field with $[K : \mathbb{Q}] = n$ and that $m > 1$. Then

$$\tilde{S}^m_{\psi(\mathcal{O}_K)}(M) \leq \tilde{N}\zeta_K(m) \, (\log M)^{n-1} + O((\log M)^{n-2})$$

and

$$\tilde{N} \, (\log M)^{n-1} + O((\log M)^{n-2}) \leq \tilde{S}^m_{\psi(\mathcal{O}_K)}(M),$$

where $\tilde{N} = \frac{\omega(n)^{n-1}}{i_K R(n-1)!}(\sqrt{|d(K/\mathbb{Q})|})^m$.
Here $d(K/\mathbb{Q})$ is the discriminant of the field K and $\zeta_K(m)$ is the value of the Dedekind zeta function of the field K at point m

B. Inverse determinant sums of complex diagonal number field codes

Let us suppose that we a complex quadratic field F and degree n field extension K/F. We then have n F-embeddings $\{\sigma_1, \cdots, \sigma_n\}$ from K to $\mathbb{C}$. We can define a *relative canonical embedding* from K into $M_n(\mathbb{C})$ by

$$\psi(x) = \mathrm{diag}(\sigma_1(x), \ldots, \sigma_n(x)),$$

where x is an element in K.

Proposition 3.3: Let K be an algebraic number field with $[K : F] = n$. If $n_r > 1$, we have that

$$\tilde{S}^{2n_r}_{\psi(\mathcal{O}_K)}(M) \leq \tilde{N}\zeta_K(n_r) \, (\log M)^{n-1} + O((\log M)^{n-2})$$

and

$$\tilde{N} \, (\log M)^{n-1} + O((\log M)^{n-2}) \leq \tilde{S}^{2n_r}_{\psi(\mathcal{O}_K)}(M),$$

where $\tilde{N} = \frac{\omega(n)^{n-1}}{R(n-1)!}(2^{-n}\sqrt{|d(K/\mathbb{Q})|})^{n_r}$.
Here R is the regulator, ω is the number of roots of unity in K and $d(K/\mathbb{Q})$ is the discriminant.

IV. Quasi-orthogonal codes from division algebras

In the following we are considering the Alamouti-like multiblock codes from [2]. With respect to their complexity and other properties, all of the codes of this type are quasi-orthogonal. It is even possible to prove that many of the fully diverse quasi-orthogonal codes in the literature are unitarily equivalent to these multi-block codes. In the following we will use several results and concepts from the theory of central simple algebras. We refer the reader to [7] for an introduction to this theory.

Let us consider the field $E = KF$ that is a compositum of a complex quadratic field F and a totally real Galois extension $K/\mathbb{Q}$ of degree k. We suppose that $K \cap F = \mathbb{Q}$, $Gal(F/\mathbb{Q}) =< \sigma >$ and $Gal(K/\mathbb{Q}) = \{\tau_1, \tau_2, \ldots, \tau_k\}$. Here σ is simply the complex conjugation. We can then write that $Gal(FK/\mathbb{Q}) = Gal(K/\mathbb{Q}) \otimes < \sigma >$.

Let us now consider a cyclic division algebra

$$\mathcal{D} = (E/K, \sigma, \gamma) = E \oplus uE,$$

where $u \in \mathcal{D}$ is an auxiliary generating element subject to the relations $xu = ux^*$ for all $x \in E$ and $u^2 = \gamma \in \mathcal{O}_K$, where $()^*$ is the complex conjugation. We can consider $\mathcal{D}$ as a right vector space over E and every element $a = x_1 + ux_2 \in \mathcal{D}$ maps to

$$\phi(a) = \begin{pmatrix} x_1 & x_2 \\ \gamma x_2^* & x_1^* \end{pmatrix}.$$

This mapping can then be extended into a multi-block representation $\phi : \mathcal{D} \mapsto M_{2k}(\mathbb{C})$.

$$\psi(a) = \mathrm{diag}(\tau_1(\phi(a)), \tau_2(\phi(a)) \ldots, \tau_k(\phi(a))). \qquad (2)$$

Example 4.1: In the case where $k = 2$ each element $a \in \mathcal{D}$ gets mapped as

$$\psi(a) = \begin{pmatrix} x_1 & x_2 & 0 & 0 \\ \gamma x_2^* & x_1^* & 0 & 0 \\ 0 & 0 & \tau(x_1) & \tau(x_2) \\ 0 & 0 & \tau(\gamma x_2)^* & \tau(x_1)* \end{pmatrix}.$$

In order to build a space-time lattice code from the division algebra $\mathcal{D}$ we will need the following definition.

Definition 4.1: Let $\mathcal{O}_K$ be the ring of integers of K. An *$\mathcal{O}_K$-order* Λ in $\mathcal{D}$ is a subring of $\mathcal{D}$, having the same identity element as $\mathcal{D}$, and such that Λ is a finitely generated module over $\mathcal{O}_K$ and generates $\mathcal{D}$ as a linear space over K.

Let us suppose that Λ is an $\mathcal{O}_K$-order in $\mathcal{D}$. We call $\phi(\Lambda)$ an *order code*. In the rest of this paper, we suppose that the division algebras under consideration are of the previous type.

Lemma 4.1: If Λ is an $\mathcal{O}_K$-order in $\mathcal{D}$

$$|\det(\phi(x))| = \sqrt{[\Lambda : x\Lambda]}, \tag{3}$$

where x is a non-zero element of Λ.

Lemma 4.2: Let us suppose that Λ is a $\mathcal{O}_K$-order of a division algebra $\mathcal{D}$ with center K of degree k and that ϕ is a multi-block representation. Then the order code $\phi(\Lambda)$ is a $4k$-dimensional lattice in the space $M_{2k}(\mathbb{C})$ and

$$\det_{min}(\psi(\Lambda)) = 1.$$

Let $\mathcal{D}$ be an index-n K-central division algebra and Λ a $\mathcal{O}_K$-order in $\mathcal{D}$. The (right) *Hey zeta function* of the order Λ is

$$\zeta_\Lambda(s) = \sum_{I \in \mathbf{I}_\Lambda} \frac{1}{[\Lambda : I]^s},$$

where $\Re(s) > 1$ and $\mathbf{I}_\Lambda$ is the set of right ideals of Λ. When $\Re(s) > 1$, this series is converging.

The unit group Λ^* of an order Λ consists of elements $x \in \Lambda$ such that there exists a $y \in \Lambda$ with $xy = 1_{\mathcal{A}}$. Another way to define this set is $\Lambda^* = \{x \in \Lambda \,|\, |\det \psi(x)| = 1\}$.

A. Inverse determinant sums of quasi-orthogonal codes

Let us suppose that K, $\mathcal{D}$ and Λ are as in the previous section and that $[K : \mathbb{Q}] = k$. We then have that $\phi(\Lambda)$ is a $4k$-dimensional NVD lattice in $M_{2k}(\mathbb{C})$ and we can consider the growth of the sum

$$\sum_{\psi(x) \in \psi(\Lambda)(M)} \frac{1}{|\det \psi(x)|^{2n_r}} = S^{2n_r}_{\psi(\Lambda)}(M).$$

Just as in [5] the previous sum can be analyzed further into

$$S^{2n_r}_{\psi(\Lambda)}(M) = \sum_{x \in X(M)} \frac{|\psi(x\Lambda^*) \cap B(M)|}{|\det(\psi(x))|^{2n_r}}, \tag{4}$$

where $X(M)$ is some collection of elements $x \in \Lambda$ such that $\|\psi(x)\|_F \le M$, each generating a different right ideal.

B. Uniform upper and lower bounds for $|\psi(x\Lambda^*) \cap B(M)|$

The key element in the analysis of $|\psi(x\Lambda^*) \cap B(M)|$ is the following.

Lemma 4.3 (Siegel): The unit group Λ^* has a subgroup

$$\mathcal{O}_K^* = \{x \,|\, x \in \Lambda^*, x \in \mathcal{O}_K\},$$

and we have $[\Lambda^* : \mathcal{O}_K^*] < \infty$.

Let $j = [\Lambda^* : \mathcal{O}_K]$. By choosing a set $\{a_1, \ldots, a_j\}$ of coset leaders of $\mathcal{O}_K^*$ in Λ^*, we have that

$$|\psi(x\Lambda^*) \cap B(M)| \le \sum_{i=1}^{j} |\psi(xa_i\mathcal{O}_K^*) \cap B(M)|. \tag{5}$$

In order to give an uniform upper bound for $|\psi(x\Lambda^*) \cap B(M)|$, it is now enough to give a uniform upper bound for $|\psi(xa_i\mathcal{O}_K^*) \cap B(M)|$. Before stating our main results we need few lemmas. We will skip the proofs of some of them.

Lemma 4.4: Let us suppose that A is a diagonal matrix in $M_n(\mathbb{C})$ with $|\det A| \ge 1$. We then have that

$$|A\psi(\mathcal{O}_K^*) \cap B(M)| \le |\psi(\mathcal{O}_K^*) \cap B(cM)|,$$

where c is a fixed constant, independent of A and M.

Lemma 4.5: Let us suppose that x and y are elements in $\mathcal{O}_{KF}$, we then have that

$$|\psi(x)\psi(\mathcal{O}_K^*) \cap B(M)| \le |\psi(\mathcal{O}_K^*) \cap B(cM)|,$$

and

$$|\psi(uy)\psi(\mathcal{O}_K^*) \cap B(M)| \le |\psi(\mathcal{O}_K^*) \cap B(cM)|,$$

where c is a real constant independent of x, y and M.

Proof: The first result is simply Lemma 4.4 and the second follows as $\psi(u)$ is a fixed matrix. □

Lemma 4.6: Let us suppose that x and y are elements in E. We then have that

$$||\psi(x) + \psi(uy)||_F^2 = ||\psi(x)||_F^2 + ||\psi(uy)||_F^2.$$

Proof: By an elementary calculation we see that $<\psi(x), \psi(uy) >= 0$ and the claim follows. □

Proposition 4.7: Let us suppose that $x \in \Lambda$, we then have that

$$|\psi(x)\psi(\mathcal{O}_K^*) \cap B(M)| \le |\psi(\mathcal{O}_K^*) \cap B(cM)|,$$

where c is a constant independent of M and x.

Proof: Let us suppose first that $x = x_1 + ux_2$, where $x_i \in \mathcal{O}_E$ and where $u^2 \in \mathcal{O}_K$. According to Lemma 4.6, we have that

$$||\psi(x)\psi(y)||^2 = ||\psi(x_1)\psi(y)||^2 + ||\psi(ux_2)\psi(y)||^2,$$

for any $y \in \mathcal{O}_E$.

Therefore if $\psi(x)\psi(y) \in B(M)$, then also

$$\psi(x_1)\psi(y) \in B(M) \text{ and } \psi(ux_2)\psi(y) \in B(M).$$

It follows that we can upper bound $|\psi(x)\psi(\mathcal{O}_K^*) \cap B(M)|$ with

$$\max\{|\psi(x_1)\psi(\mathcal{O}_K^*) \cap B(M)|, |\psi(ux_2)\psi(\mathcal{O}_K^*) \cap B(M)|\}.$$

According to lemma 4.5 we then have that

$$|\psi(x)\psi(\mathcal{O}_K^*) \cap B(M)| \leq |\psi(\mathcal{O}_K^*) \cap B(cM)|.$$

Let us now suppose that Λ is a general order in $\mathcal{D}$. As $\psi(\Lambda)$ is finitely generated as an additive group in $M_n(E)$, we can choose an integer d such that $d\psi(\Lambda) \subseteq \psi(\mathcal{O}_E) + \psi(u\mathcal{O}_E)$. The result now follows from the previous consideration. □

Proposition 4.8: Using the previous notation we have

$$|\psi(x\Lambda^*) \cap B(M)| \leq [\Lambda^* : \mathcal{O}_K] \cdot$$
$$\log M^{k-1} \frac{\omega(k)^{k-1}}{Ri_K(k-1)!} + O(\log M^{k-2}).$$

Proof: Let $j = [\Lambda^* : \mathcal{O}_K]$. By choosing a set $\{a_1, \dots, a_j\}$ of coset leaders of $\mathcal{O}_K^*$ in Λ^*, we then have that

$$|\psi(x\Lambda^*) \cap B(M)| \leq \sum_{i=1}^{j} |\psi(xa_i\mathcal{O}_K^*) \cap B(M)|.$$

According to Proposition 4.7 we then have that

$$|\psi(x\Lambda^*) \cap B(M)| \leq [\Lambda^* : \mathcal{O}_K]|\psi(\mathcal{O}_K^*) \cap B(cM)|.$$

Applying Theorem 3.1 to this equation, we get the final result. □

C. Upper and lower bounds for inverse determinant sums of quasi-orthogonal codes

Proposition 4.9: Let us suppose that $[K : \mathbb{Q}] = k$ and set $n = 2k$. We then have that $\psi(\Lambda)$ is a $2n$-dimensional lattice in $M_n(\mathbb{C})$ and

$$\log M^{n/2-1} \frac{\omega(\frac{n}{2})^{n/2-1}}{Ri_K(n/2-1)!} + O(\log M^{n/2-2}) \leq S^{2n_r}_{\psi(\Lambda)}(M)$$

$$\leq \zeta_\Lambda(n_r)[\Lambda^* : \mathcal{O}_K] \log M^{\frac{n-2}{2}} \frac{\omega(\frac{n}{2})^{\frac{n-2}{2}}}{Ri_K(\frac{n-2}{2})!} + O(\log M^{n/2-2}),$$

where $n_r > 1$ and R and ω are the regulator and the number of roots of unity in the center K and i_K the index of norm 1 units.

Proof: As previously mentioned, we can imitate [5] to get

$$S^{2n_r}_{\psi(\Lambda)}(M) = \sum_{x \in X(M)} \frac{|\psi(x\Lambda^*) \cap B(M)|}{|\det(\psi(x))|^{2n_r}}. \tag{6}$$

According to Lemma 4.1 we have that $|\det(\psi(x))|^{2n_r} = [\Lambda : x\Lambda]^{n_r}$. Now

$$\sum_{x \in X(M)} \frac{1}{|\det(\psi(x))|^{2n_r}} \leq \sum_{x \in X(M)} \frac{1}{[\Lambda : x\Lambda]^{n_r}} \leq$$
$$\leq \zeta_\Lambda(n_r).$$

Applying this inequality with Proposition 4.8 to (6) now gives us the final result. □

V. Quasi-orthogonal codes are better than diagonal number field codes

Let us now suppose we have an $n \times n_r$-MIMO channel, (for simplicity we assume $n_r > 1$). For the existence of quasi-orthogonal code we also have to assume that $2 \mid n$. Let us now compare the growth of determinant sums of quasi-orthogonal and comparable diagonal number field codes in this $n \times n_r$-MIMO channel.

In order to build a quasi-orthogonal code $\psi(\Lambda)$ in $M_n(\mathbb{C})$ the center K of the algebra $\mathcal{D}$ must be an $n/2$-dimensional totally real number field. For a number field code $\psi(\mathcal{O}_L) \subseteq M_n(\mathbb{C})$, the field L must be an n-dimensional extension of some complex quadratic field F.

As we earlier saw, we have that

$$\sum_{X \in \psi(\Lambda)(M)} \frac{1}{|\det(X)|^{2n_r}} = \theta(|\psi(\Lambda^*) \cap B(M)|)$$

and

$$|\psi(\Lambda^*) \cap B(M)| = \theta(|\psi(\mathcal{O}_K^*) \cap B(M)|) = \theta(\log M^{n/2-1}).$$

Therefore

$$\sum_{X \in \psi(\Lambda)(M)} \frac{1}{|\det(X)|^{2n_r}} = \theta(\log M^{n/2-1}).$$

On the other hand for the number field code we have that

$$\sum_{X \in \psi(\mathcal{O}_L)(M)} \frac{1}{|\det(X)|^{2n_r}} = \theta(|\psi(\mathcal{O}_L^*) \cap B(M)|)$$
$$= \theta(\log M^{n-1}).$$

Here the last result follows from [6, Theorem 2].

We can now see that the growth of the inverse determinant sum for the quasi-orthogonal code is considerably lower than that of the number field code. This is due to the fact that the unit group of the order Λ is essentially that of a low degree real number field. We note that this difference can not be captured in the context of DMT as both of these codes have the same DMT curve.

Acknowledgement

The research of R. Vehkalahti is supported by the Academy of Finland grant #252457.

References

[1] A.-Z. Wong and J.-K. Zhang, "Novel Rotated Quasi-Orthogonal Space-Time Block Codes with the Fixed Nearest Neighbor Number", vol. 17, *IEEE Signal Proc. Lett.*, pp. 965–968, Nov 2010.

[2] P. E. Elia and P. V. Kumar, "Approximately-Universal Space-Time Codes for the Parallel, Multi-Block and Cooperative-Dynamic-Decode-and-Forward Channels", preprint available at: http://arxiv.org/abs/0706.3502

[3] Hiltunen, C. Hollanti, J. Lahtonen, "Four Antenna Space-Time Lattice Constellations from Division Algebras", Proc. IEEE Int. Symp. on Inf. Theory, Chicago, USA, 2004.

[4] F. Oggier, J.-C. Belfiore, "An Error Probability Approach to MIMO Wiretap Channels", preprint available at http://arxiv.org/abs/1109.6437

[5] R.Vehkalahti, H.-f. Lu, L.Luzzi, "Inverse Determinant Sums and Connections Between Fading Channel Information Theory and Algebra", *IEEE Trans. Inf. Theory*, vol 59, pp. 6060–6082, September 2013.

[6] G. Everest and J.H. Loxton "Counting algebraic units with bounded height", *J. Number Theory*, vol 44., pp. 222–227, June 1993.

[7] I. Reiner, *Maximal Orders*, Academic Press, New York 1975.

Constrained Colluding Eavesdroppers: An Information-Theoretic Model

Mahtab Mirmohseni and Panagiotis Papadimitratos
KTH Royal Institute of Technology, Stockholm, Sweden
Email: {mahtabmi,papadim}@kth.se

***Abstract*—We study the secrecy capacity in the vicinity of colluding eavesdroppers. Contrary to the *perfect collusion* assumption in previous works, our new information-theoretic model considers *constraints* in collusion. We derive the achievable secure rates (lower bounds on the perfect secrecy capacity), for the discrete memoryless channel and the Gaussian channel. We also compare the proposed rates to the non-colluding and perfect colluding cases.**

I. Introduction

Wyner [1] introduced the information-theoretic model for confidentiality in noisy communications, called *wiretap channel*, where a legitimate transmitter wishes to transmit a confidential message to a legitimate receiver while keeping it hidden from an eavesdropper (wiretapper). The eavesdropper is assumed to have unlimited computation power, to know the coding scheme of the legitimate user, and to only listen to the channel. When the channel to the eavesdropper is a degraded version of the channel to the legitimate receiver, Wyner [1] proposed the secrecy capacity achieving scheme, known also as *Wyner's wiretap channel coding*, which comprises multicoding and randomized encoding [2, Section 22.1.1]. This result is extended to the broadcast channel with confidential message and to the general wiretap channel (not necessarily degraded) by Csiszár and Körner [3].

Recently, different legitimate-wiretapper user combinations were studied [4]–[8]. In this line of works, scenarios with multiple eavesdroppers considered only *non-colluding* ones. This implies that information leakage of a certain message to all eavesdroppers is computed as the maximum of the leakages (to each one). In some applications, this assumption may underestimate the eavesdroppers' power: they can collude, i.e., share their channel outputs (observations), and render the attack more effective [9]. Hence, combating colluding eavesdroppers, especially in wireless networks, has been a significant challenge [9]–[14]. To the best of our knowledge, all previous works modeled k colluding eavesdroppers as one eavesdropper with k antennas; we term these *perfect colluding* eavesdroppers. Using the equivalent Single-Input Multiple-Output (SIMO) Gaussian wiretap channel, the information leakage is determined by the aggregate Signal to Noise Ratio (SNR) of all eavesdroppers; compared to the maximum SNR in the non-colluding case [9]. This assumption significantly overestimates the eavesdropping capability, forcing a legitimate user to increase its power linearly with the number of eavesdroppers to achieve a positive secure rate. However, collusion (esp. in the wireless networks) necessitates communication resources and power consumption. This, in fact, restricts the collusion channel capacity and thus improves the achievable secure rate by the legitimate user. Hence, the problem at hand is to find an appropriate model and to analyze the effect of these constraints on the secrecy capacity.

In this paper, we model *constrained collusion* with an equivalent wiretap channel, called *Wiretap Channel with Constrained Colluding Eavesdroppers* (WTC-CCE). For our *general* WTC-CCE, we assume that colluding eavesdroppers communicate (by defining their channel inputs) over a virtual *collusion channel*, in addition to the main channel. The higher the collusion channel capacity, the more leaked information can be exchanged. Our model captures previously studied models as special cases: non-colluding eavesdroppers with zero collusion rates and perfect colluding ones with infinite collusion rates. We also propose a special case, the *orthogonal* WTC-CCE: the collusion channel is orthogonal to the main one (unlike the general WTC-CCE where the eavesdroppers share the same channel with the legitimate transmitter). First, we derive an achievable secure rate (a lower bound on the perfect secrecy capacity) for the general discrete memoryless WTC-CCE. The idea is to let the eavesdroppers do their best in colluding. Hence, the information leakage rate is derived by considering the outer bound on the capacity region of the collusion channel; this resembles the cut-set upper bound for the relay channel [2]. Next, we extend our result to the general Gaussian WTC-CCE and its orthogonal version. The main difference is that, in the general model, the eavesdroppers may use jamming techniques to confuse the legitimate receiver; but this way they could be exposed (to the legitimate user). In the orthogonal model, beyond the increased required resources, the eavesdroppers may loose some information leakage rate because they cannot send jamming signals. However, the orthogonality may serve eavesdroppers in hiding themselves. We provide numerical examples to analyze the achievable secure rate and evaluate the overestimation amount (by comparing to perfect colluding case) in different scenarios.

II. Channel Model and Preliminaries

Upper-case letters (e.g., X) denote Random Variables (RVs) and lower-case letters (e.g., x) their realizations. The probability mass function (p.m.f) of a RV X with alphabet set $\mathcal{X}$ is denoted by $p_X(x)$; occasionally, the subscript X is omitted. X_i^j indicates a sequence of RVs $(X_i, X_{i+1}, ..., X_j)$; we use

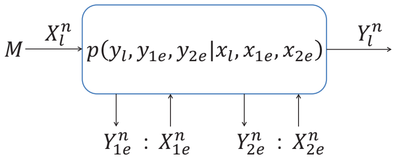


Fig. 1. General Wiretap Channel with Constrained Colluding Eavesdroppers (WTC-CCE).

X^j instead of X_1^j for brevity. $\mathcal{N}(0, \sigma^2)$ denotes a zero-mean Gaussian distribution with variance σ^2.

Consider the WTC-CCE in Fig. 1: a four terminal discrete channel (one transmitter, one legitimate receiver and two eavesdroppers), denoted by $(\mathcal{X}_l \times \mathcal{X}_{1e} \times \mathcal{X}_{2e}, p(y_l^n, y_{1e}^n, y_{2e}^n|x_l^n, x_{1e}^n, x_{2e}^n), \mathcal{Y}_l \times \mathcal{Y}_{1e} \times \mathcal{Y}_{1e})$. $X_l \in \mathcal{X}_l$ and $X_{je} \in \mathcal{X}_{je}$ are the channel inputs of the legitimate transmitter and eavesdropper j and $Y_l \in \mathcal{Y}_l$ and $Y_{je} \in \mathcal{Y}_{je}$ are the channel outputs at the legitimate receiver and eavesdropper j, for $j \in \{1, 2\}$. $p(y_l^n, y_{1e}^n, y_{2e}^n|x_l^n, x_{1e}^n, x_{2e}^n)$ is the channel transition probability distribution. We also assume that the channel is memoryless. In n channel uses, the legitimate transmitter desires to send the message M to the legitimate receiver using the following code.

Definition 1: A $(2^{nR}, n, P_e^{(n)})$ code for WTC-CCE consists of: (i) A message set $\mathcal{M} = [1 : 2^{nR}]$, where m is uniformly distributed over $\mathcal{M}$. (ii) A *randomized* encoding function, f_n, at the legitimate transmitter that maps a message m to a codeword $x_l^n \in \mathcal{X}_l^n$. (iii) Two sets of encoding functions at the eavesdroppers: $\{f_{je,t}\}_{t=1}^n : \mathbb{R}^{t-1} \longrightarrow \mathbb{R}$ such that $x_{je,t} = f_{je,t}(y_{je}^{t-1})$, for $j \in \{1, 2\}$ and $1 \leq t \leq n$. (iv) A decoding function at the legitimate receiver $g : \mathcal{Y}_l^n \mapsto \mathcal{M}$. (v) Probability of error for this code is defined as: $P_e^{(n)} = \frac{1}{2^{nR}} \sum_{m \in \mathcal{M}} Pr(g(y_l^n) \neq m|m \text{ sent})$. (vi) The information leakage rate at eavesdropper $j \in \{1, 2\}$ is defined as:

$$R_{L,j}^{(n)} = \frac{1}{n} I(M; Y_{je}^n). \tag{1}$$

All codewords are revealed to the eavesdroppers. However, the eavesdroppers' mappings are not known to the legitimate user.

Remark 1: The mutual information term in (1) is the same as in the non-colluding case, compared to $I(M; Y_{1e}^n, Y_{2e}^n)$ in the perfect colluding scenario. The difference here comes from the channel distribution and the fact that Y_{1e}^n and Y_{1e}^n given X_l are not independent (due to X_{1e} and X_{2e}).

Definition 2: A rate-leakage tuple $(R, R_{L,1}, R_{L,2})$ is achievable if there exists a sequence of $(2^{nR}, n, P_e^{(n)})$ codes such that $P_e^{(n)} \to 0$ as $n \to \infty$ and $\limsup_{n \to \infty} R_{L,j}^{(n)} \leq R_{L,j}$ for $j \in \{1, 2\}$. The secrecy capacity $\mathcal{C}_s$ is the supremum of all achievable rates R such that perfect secrecy is achieved, i.e., $R_{L,j} = 0$ for $j \in \{1, 2\}$.

Motivated by the fact that the eavesdroppers prefer to avoid exposure, we also consider a special case of the WTC-CCE. We assume that the collusion channel (used by the eavesdroppers) is decoupled from the main channel and we consider the orthogonal WTC-CCE in Fig. 2. Here, $Y_{je} = (Y_{je}^m, Y_{je}^c)$ for $j \in \{1, 2\}$ and $p(y_l, y_{1e}, y_{2e}|x_l, x_{1e}, x_{2e}) = p(y_l, y_{1e}^m, y_{2e}^m|x_l)p(y_{1e}^c, y_{2e}^c|x_{1e}, x_{2e})$, where the variables relating to the main and the collusion channels are indicated with the superscripts m and c respectively. Substituting $X_{1e} = X_{2e} = \emptyset$ results in the non-colluding case; $Y_{1e}^c = Y_{2e}^m, Y_{2e}^c = Y_{1e}^m$ results in the perfect colluding case. To simplify notation, let $\bar{j}$ be the complement of j in $\{1, 2\}$. Now, consider the general Gaussian WTC-CCE at time $t = 1, \ldots, n$ for $j \in \{1, 2\}$, modeled as:

$$\begin{aligned} Y_{l,t} &= h_l X_{l,t} + h_{1e}^l X_{1e,t} + h_{2e}^l X_{2e,t} + Z_{l,t} \\ Y_{je,t} &= h_l^{je} X_{l,t} + h_{\bar{j}e}^{je} X_{\bar{j}e,t} + Z_{je,t} \end{aligned} \tag{2}$$

where h_i^k is a known channel gain from transmitter i to receiver k. We assume perfect echo cancellation at eavesdroppers ($h_{1e}^{1e} = h_{2e}^{2e} = 0$). $X_{u,t}$ is an input signal with average power constraint $\frac{1}{n} \sum_{t=1}^{n} |x_{u,t}|^2 \leq P_u$ and $Z_{u,t}$ is an independent and identically distributed (i.i.d) zero-mean Gaussian noise component with power N_u, for $u \in \{l, 1e, 2e\}$. In practice, h_{1e}^l and h_{2e}^l may be small. The Gaussian counterpart of the orthogonal WTC-CCE for $j \in \{1, 2\}$ can be shown as:

$$\begin{aligned} Y_{l,t} &= h_l X_{l,t} + Z_{l,t} \\ Y_{je,t}^m &= h_{jm} X_{l,t} + Z_{je,t}^m \quad , \quad Y_{je,t}^c \; = \; h_{jc} X_{\bar{j}e,t} + Z_{je,t}^c \end{aligned} \tag{3}$$

where h_{jm} and h_{jc} are known channel gains received at eavesdropper j from the main channel and the collusion channel, respectively; power constraints of P_l, P_{1e}, P_{2e} apply for input signals; $Z_{je,t}^m$ and $Z_{je,t}^c$ are i.i.d zero-mean Gaussian noise components with powers N_{je}^m and N_{je}^c at eavesdropper j from the main channel and the collusion channel, respectively.

III. Discrete Memoryless Channel

Our first result establishes an achievable secure rate for the general discrete memoryless WTC-CCE.

Theorem 1: For the general discrete memoryless WTC-CCE, the secrecy capacity is lower-bounded by:

$$\begin{aligned} \mathcal{R}_s^{DM} = \sup \inf I(X_l; Y_l) - \min\{ & I(X_l; Y_{1e}, Y_{2e}|X_{1e}, X_{2e}), \\ & \max\{I(X_l, X_{1e}, X_{2e}; Y_{1e}), I(X_l, X_{1e}, X_{2e}; Y_{2e})\}\} \end{aligned} \tag{4}$$

where the supremum and infimum are taken over all joint p.m.fs of the form $p(x_l|x_{1e}, x_{2e})p(y_l, y_{1e}, y_{2e}|x_l, x_{1e}, x_{2e})$ and $p(x_{1e}, x_{2e})$ respectively.

Proof: The proof is based on the random coding scheme, which uses Wyner wiretap coding at the legitimate user. For the eavesdroppers, the idea is to let them do their best in colluding. Hence, the coding strategy of the eavesdroppers is

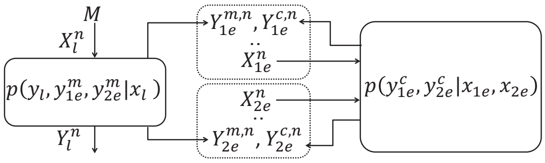

M
X_l^n
$p(y_l, y_{1e}^m, y_{2e}^m|x_l)$
Y_l^n
$Y_{1e}^{m,n}, Y_{1e}^{c,n}$
X_{1e}^n
X_{2e}^n
$Y_{2e}^{m,n}, Y_{2e}^{c,n}$
$p(y_{1e}^c, y_{2e}^c|x_{1e}, x_{2e})$

Fig. 2. Orthogonal WTC-CCE.

$$\mathcal{R}_s^{OG}{=}\theta(\frac{h_l^2P_l}{N_l})-\min\{\theta(P_l(\frac{h_{1m}^2}{N_{1e}^m}+\frac{h_{2m}^2}{N_{2e}^m})),\max\{\theta(\frac{h_{1m}^2P_l}{N_{1e}^m}+\frac{h_{1c}^2P_{2e}}{N_{1e}^c}+\frac{h_{1m}^2h_{1c}^2P_lP_{2e}}{N_{1e}^cN_{1e}^m}),\theta(\frac{h_{2m}^2P_l}{N_{2e}^m}+\frac{h_{2c}^2P_{1e}}{N_{2e}^c}+\frac{h_{2m}^2h_{2c}^2P_lP_{1e}}{N_{2e}^cN_{2e}^m})\}\}.\tag{7}$$

$$\begin{aligned}\mathcal{R}_s^{G}{=}\min_{\rho_1,\rho_2,\rho_{12}}&\theta(\frac{h_l^2P_l+\rho_1^2(h_{1e}^l)^2P_{1e}+\rho_2^2(h_{2e}^l)^2P_{2e}+2h_lh_{1e}^l\rho_1\sqrt{P_lP_{1e}}+2h_lh_{2e}^l\rho_2\sqrt{P_lP_{2e}}}{(h_{1e}^l)^2P_{1e}(1-\rho_1^2)+(h_{2e}^l)^2P_{2e}(1-\rho_2^2)+2h_{1e}^lh_{2e}^l\rho_{12}\sqrt{P_{1e}P_{2e}}+N_l})\\&-\min\{\max\{A(1),A(2)\},\theta(P_l(1-\frac{\rho_1^2P_{1e}^2+\rho_2^2P_{2e}^2+2\rho_1\rho_2\rho_{12}P_{1e}P_{2e}}{P_{1e}P_{2e}(1-\rho_{12}^2)})(\frac{(h_l^{1e})^2}{N_{1e}}+\frac{(h_l^{2e})^2}{N_{2e}}))\}.\end{aligned}\tag{8}$$

not determined in the scheme. As a result, the information leakage rate is derived by considering the outer bound on the capacity region of the collusion channel and it looks like the cut-set upper bound for the relay channel [2].

Codebook Generation: Generate $2^{n(R+R_s)}$ i.i.d x_l^n sequences, each with probability $\prod_{t=1}^{n}p(x_{l,t})$. Index them as $x_l^n(m,s)$, where $m\in[1:2^{nR}]$ and $s\in[1:2^{nR_s}]$.

Encoding: To send message $m\in[1:2^{nR}]$, the stochastic encoder at the legitimate transmitter uniformly randomly chooses s and transmits $x_l^n(m,s)$.

Decoding: The decoder at the legitimate receiver wants to correctly recover m,s and seeks a unique message $\tilde{m}$ and some $\tilde{s}$ such that $(x_l^n(\tilde{m},\tilde{s}),y_l^n)$ are jointly typical. Applying the packing lemma [2], with arbitrarily high probability $\tilde{m}=m$ if n is large enough and

$$R+R_s{\le}I(X_l;Y_l).\tag{5}$$

Analysis of the information leakage rate: To simplify the notation, let $X_e=(X_{1e},X_{2e})$ and $Y_e=(Y_{1e},Y_{2e})$. We derive two bounds for the randomness index rate, R_s. First, we obtain the second term of information leakage rates in the min term in (4), i.e., $R_{L2}=\max\{I(X_l,X_{1e},X_{2e};Y_{1e}),I(X_l,X_{1e},X_{2e};Y_{2e})\}$.

Now, consider the leaked information to Y_{1e}^n averaged over the random codebook $\mathcal{C}$.

$$\begin{aligned}&I(M;Y_{1e}^n|\mathcal{C})=H(M|\mathcal{C})-H(M|Y_{1e}^n,\mathcal{C})\\&=nR-H(M,Y_{1e}^n,X_l^n,X_e^n|\mathcal{C})+H(X_l^n,X_e^n|M,Y_{1e}^n,\mathcal{C})+H(Y_{1e}^n|\mathcal{C})\\&=nR-H(X_l^n,X_e^n|\mathcal{C})-H(M,Y_{1e}^n|X_l^n,X_e^n,\mathcal{C})\\&+H(X_l^n,X_e^n|M,Y_{1e}^n,\mathcal{C})+H(Y_{1e}^n|\mathcal{C})\\&{\le}nR-H(X_l^n|\mathcal{C})-H(Y_{1e}^n|X_l^n,X_e^n,\mathcal{C})\\&+H(X_l^n,X_e^n|M,Y_{1e}^n,\mathcal{C})+H(Y_{1e}^n|\mathcal{C})\\&=nR-n(R+R_s)+I(X_l^n,X_e^n;Y_{1e}^n|\mathcal{C})+H(X_l^n,X_e^n|M,Y_{1e}^n,\mathcal{C})\\&\overset{(a)}{\le}-nR_s+nI(X_l,X_e;Y_{1e})+H(X_l^n,X_e^n|M,Y_{1e}^n,\mathcal{C})\overset{(b)}{\le}n\delta_1\end{aligned}$$

(a) holds because the channel is memoryless; (b) follows from [2, Lemma 22.1]: if $R_s\ge I(X_l,X_{1e},X_{2e};Y_{1e})$, then $H(X_l^n,X_{1e}^n,X_{2e}^n|M,Y_{1e}^n,\mathcal{C})\le nR_s-nI(X_l,X_{1e},X_{2e};Y_{1e})+n\delta_1$. Following similar steps, one can show that if $R_s\ge I(X_l,X_{1e},X_{2e};Y_{2e})$, then $I(M;Y_{2e}^n|\mathcal{C})\le\delta_2$. Considering (1), combining (5) and these constraints on R_s gives $\mathcal{R}_s^{DM}$ with R_{L2}.

Now, to derive the first term of information leakage rates in min in (4), i.e., $R_{L1}=I(X_l;Y_{1e},Y_{2e}|X_{1e},X_{2e})$, we evaluate the leaked information to both Y_{1e}^n and Y_{2e}^n, averaged over the random codebook $\mathcal{C}$.

$$\begin{aligned}&I(M;Y_e^n|\mathcal{C})=H(M|\mathcal{C})-H(M|Y_e^n,\mathcal{C})\\&=nR-H(M,Y_e^n,X_l^n|\mathcal{C})+H(X_l^n|M,Y_e^n,\mathcal{C})+H(Y_e^n|\mathcal{C})\\&\overset{(a)}{=}nR-H(X_l^n|\mathcal{C})-H(M,Y_e^n|X_l^n,\mathcal{C})\\&+H(X_l^n|M,Y_e^n,X_e^n,\mathcal{C})+H(Y_e^n|\mathcal{C})\\&\overset{(b)}{\le}nR-n(R+R_s)+I(X_l^n;Y_e^n|\mathcal{C})+H(X_l^n|M,Y_e^n,X_e^n,\mathcal{C})\\&\overset{(c)}{=}-nR_s+\sum_{i=1}^{n}I(X_l^n;Y_{e,i}|Y_e^{i-1},X_{e,i},\mathcal{C})+H(X_l^n|M,Y_e^n,X_e^n,\mathcal{C})\\&\overset{(d)}{\le}-nR_s+nI(X_l;Y_e|X_e)+H(X_l^n|M,Y_e^n,X_e^n,\mathcal{C})\overset{(e)}{\le}n\delta_3\end{aligned}\tag{6}$$

(a) and (c) follow because $x_{je,t}=f_{je,t}(y_{je}^{t-1})$, for $j\in\{1,2\}$ and $1\le t\le n$; (b) is due to the fact that conditioning does not increase the entropy; (d) holds due to the memoryless property of the channel; (e) follows from [2, Lemma 22.1]: if $R_s\ge I(X_l;Y_{1e},Y_{2e}|X_{1e},X_{2e})$, then $H(X_l^n|M,Y_{1e}^n,Y_{2e}^n,X_{1e}^n,X_{2e}^n,\mathcal{C})\le nR_s-nI(X_l;Y_{1e},Y_{2e}|X_{1e},X_{2e})+n\delta_3$. Note that (6) implies $I(M;Y_{je}^n|\mathcal{C}){\le}n\delta_3$ for $j\in\{1,2\}$ (for the individual leakage rates). Now, combining (5) and this constraint on R_s gives $\mathcal{R}_s^{DM}$ with R_{L1}. This completes the proof. ■

Remark 2: Substituting $Y_{je}=(Y_{je}^m,Y_{je}^c)$ for $j\in\{1,2\}$ in (4) results in an achievable secure rate ($\mathcal{R}_s^{ODM}$) for the orthogonal discrete memoryless WTC-CCE, where the supremum is taken over all joint p.m.fs of the form $p(x_l|x_{1e},x_{2e})p(y_l,y_{1e}^m,y_{2e}^m|x_l)p(y_{1e}^c,y_{2e}^c|x_{1e},x_{2e})$.

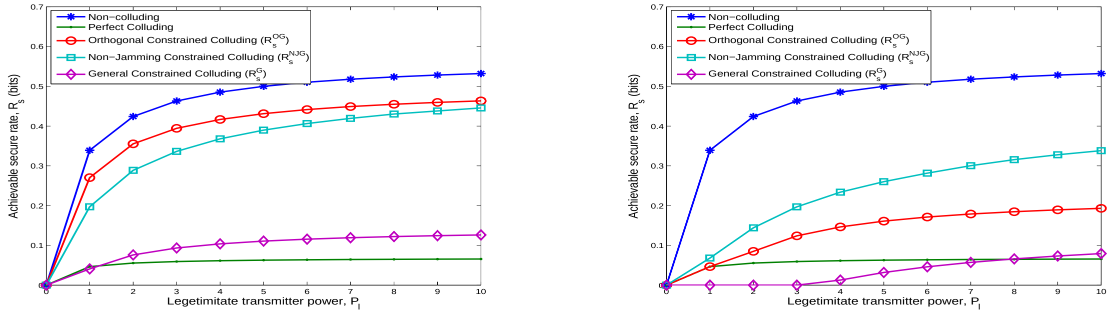
Remark 3: By setting $X_{1e}=X_{2e}=\emptyset$ in (4), $\mathcal{R}_s^{DM}$ reduces to $\sup I(X_l;Y_l)-\max\{I(X_l;Y_{1e}),I(X_l;Y_{2e})\}$ for the non-colluding case. Furthermore, redefining $Y_{1e}^c=Y_{2e}^m,Y_{2e}^c=Y_{1e}^m$ in $\mathcal{R}_s^{ODM}$ results in the achievable secure rate for the perfect colluding case, i.e., $\sup I(X_l;Y_l)-I(X_l;Y_{1e},Y_{2e})$.

IV. Gaussian Channel

We study the Gaussian WTC-CCE. First, we consider the orthogonal Gaussian WTC-CCE. Let $\theta(x)\doteq\frac{1}{2}\log(1+x)$.

Theorem 2: $\mathcal{R}_s^{OG}$ in (7), shown at the top of the page, is an achievable secure rate for orthogonal Gaussian WTC-CCE (defined in (3)).

Proof: We can extend the achievable secrecy rate in Theorem 1 (after applying Remark 2) to the Gaussian case with continuous alphabets using standard arguments [15]. As we do not know the optimal distribution $p(x_l|x_{1e},x_{2e})$ that maximizes $\mathcal{R}_s^{ODM}$, we use a Gaussian input distribution (at the legitimate transmitter) to achieve a lower bound. Let $X_l\sim\mathcal{N}(0,P_l)$. Note that the leakage rates in $\mathcal{R}_s^{ODM}$ (i.e., R_{L1} and R_{L2}) are Multiple Access Channel (MAC) type bounds. From the maximum-entropy theorem [15] (or [2, P. 21]), these bounds are largest (or equivalently $\mathcal{R}_s^{ODM}$ is minimized over $p(x_{1e},x_{2e})$) for Gaussian inputs at the eavesdroppers. Hence, set $X_{je}\sim\mathcal{N}(0,P_{je})$ for $j\in\{1,2\}$ and define $-1\le\rho_j\le1$ as the correlation coefficient between



(a) $h^{1e}_{2e} = h^{2e}_{1e} = h_{jc} = \sqrt{0.1}, j \in \{1, 2\}$.

(b) $h^{1e}_{2e} = h^{2e}_{1e} = h_{jc} = \sqrt{0.6}, j \in \{1, 2\}$.

Fig. 3. Achievable secure rates $\mathcal{R}_s$ for $P_{je} = 1$, $h^l_{je} = \sqrt{0.2}$, $h^{je}_l = h_{jm} = 1$, $N_l = N_{je} = N^m_{je} = N^c_{je} = 1, j \in \{1, 2\}$.

X_{je} and X_l, i.e., $E(X_{je}X_l) = \rho_j\sqrt{P_{je}P_l}$ for $j \in \{1, 2\}$ and $\rho_{12} = \frac{E(X_{1e}X_{2e})}{\sqrt{P_{1e}P_{2e}}}$. After calculating the mutual information terms in (4), one can easily show that the leakage rate is maximized (or secure rate in minimized) for $\rho_{12} = \rho_1 = \rho_2 = 0$. This means that in the orthogonal setup the best strategy for the eavesdroppers is to use independent codewords. This achieves $\mathcal{R}_s^{OG}$ in (7). ■

Remark 4: To achieve the non-colluding rate, i.e., $\theta(\frac{h_l^2 P_l}{N_l}) - \max\{\theta(\frac{h_{1m}^2 P_l}{N_{1e}^m}), \theta(\frac{h_{2m}^2 P_l}{N_{2e}^m}\}$, set $P_{1e} = P_{2e} = 0$ in $\mathcal{R}_s^{OG}$. Moreover, it is enough to set $P_{1e}, P_{2e} \to \infty$ in $\mathcal{R}_s^{OG}$ to derive the perfect colluding rate: $\theta(\frac{h_l^2 P_l}{N_l}) - \theta(P_l(\frac{h_{1m}^2}{N_{1e}^m} + \frac{h_{2m}^2}{N_{2e}^m}))$. Next, we obtain the secure rate for the general Gaussian WTC-CCE. The proof is similar to Theorem 2.

Theorem 3: $\mathcal{R}_s^G$ in (8), shown at top of the previous page, is an achievable secure rate for Gaussian WTC-CCE (in (2)). For $j \in \{1, 2\}$: $A(j) = \theta(\frac{(h_l^{je})^2 P_l + (h_{\bar{j}e}^{je})^2 P_{\bar{j}e} + 2h_l^{je} h_{\bar{j}e}^{je} \rho_2 \sqrt{P_l P_{\bar{j}e}}}{N_{je}})$.

Remark 5: Channel gains h_{1e}^l and h_{2e}^l make jamming possible for the eavesdroppers. However, they also increase the probability of exposure. In order to compare the two strategies (through numerical examples), we define the non-jamming rate $\mathcal{R}_s^{NJG}$ by setting $h_{1e}^l = h_{2e}^l = 0$ in $\mathcal{R}_s^G$. In addition, by setting $P_{1e}, P_{2e} \to \infty$ in $\mathcal{R}_s^G$, the secure rate is zero, which is less than (or equal to) the perfect colluding rate. This is due to jamming and it is achieved by $\rho_{12} = \rho_1 = \rho_2 = 0$.

Fig. 3 compares the secure rates for the Gaussian WTC-CCE, i.e., $\mathcal{R}_s^G, \mathcal{R}_s^{OG}, \mathcal{R}_s^{NJG}$, to the non-colluding and perfect colluding scenarios in two different collusion channel conditions. It can be seen that the perfect collusion assumption significantly overestimates the eavesdroppers. Recall that the WTC-CCE rates consider the best possible strategy for the eavesdroppers; which may not be achievable for them. In Fig. 3a (a weak collusion channel), using the orthogonal collusion channel for eavesdroppers is worse than using the non-orthogonal one (because $\mathcal{R}_s^{OG} \geq \mathcal{R}_s^{NJG}$). In fact, with weak direct collusion links, eavesdroppers may benefit from the main channel by relaying (transmitting correlated codewords). Hence, the optimal ρ_1, ρ_2 for $\mathcal{R}_s^{NJG}$ are not zero; but they are zero for $\mathcal{R}_s^{OG}$. However, for an improved collusion channel (in Fig. 3b), using an orthogonal collusion channel is better if one cannot use jamming (or does not want to use jamming, to avoid exposure). To evaluate $\mathcal{R}_s^G$, one should note the effect of jamming in addition to collusion, which enables the eavesdroppers (or, now, jammers) to make the secure rate zero.

V. Conclusion

We proposed WTC-CCE, a wiretap-based channel model to capture collusion constraints and derived the achievable secure rates. Our results showed that, indeed, the perfect collusion model overestimates the eavesdroppers if they choose to be unexposed. With no exposure constraint, they can jam to further reduce the secure rate in some cases.

References

[1] A. Wyner, “The wire-tap channel,” *Bell Syst. Tech. J.*, vol. 54, no. 8, Oct. 1975.

[2] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge Univ. Press, 2011.

[3] I. Csiszar and J. Korner, “Broadcast channels with confidential messages,” *IEEE Trans. Inf. Theory*, vol. 24, no. 3, May 1978.

[4] R. Liu, I. Maric, P. Spasojevic, and R. D. Yates, “Discrete memoryless interference and broadcast channels with confidential messages: secrecy rate regions,” *IEEE Trans. Inf. Theory*, vol. 54, no. 6, Jun. 2008.

[5] E. Ekrem and S. Ulukus, “Multi-receiver wiretap channel with public and confidential messages,” *IEEE Trans. Inf. Theory*, vol. 59, no. 4, April 2013.

[6] Y. K. Chia and A. El Gamal, “Three-receiver broadcast channels with common and confidential messages,” *IEEE Trans. Inf. Theory*, vol. 58, no. 5, May 2012.

[7] Y. Oohama, “Capacity theorems for relay channels with confidential messages,” *Proc. IEEE ISIT*, Nice, France, June 2007

[8] L. Lai and H. El Gamal, “The relay-eavesdropper channel: cooperation for secrecy,” *IEEE Trans. Inf. Theory*, vol. 54, no. 9, Sep. 2008.

[9] P. C. Pinto, J. Barros, and M. Z. Win, “Wireless physical-layer security: the case of colluding eavesdroppers,” *Proc. IEEE ISIT*, Jun. 2009

[10] O. O. Koyluoglu, C. E. Koksal, and H. A. El Gamal, “On Secrecy Capacity Scaling in Wireless Networks,” *IEEE Trans. Inf. Theory*, vol. 58, no. 5, May 2012.

[11] J. Zhang, L. Pu, X. Wang, “Impact of secrecy on capacity in large-scale wireless networks,” *Proc. IEEE INFOCOM, Mini-Conference*, 2012.

[12] P. C. Pinto, J. Barros, and M. Z. Win, “Secure communication in stochastic wireless networks part II: maximum rate and collusion,” *IEEE Trans. Inf. Forensics and Security*, vol. 7, no. 1, Feb. 2012.

[13] S. Goel and R. Negi, “Secret communication in presence of colluding eavesdroppers,” *Proc. IEEE MILCOM*, Oct. 2005.

[14] J. Wang, P. Huang, and X. Wang, “Cross-layer scheduling in multi-user system with delay and secrecy constraints,” http://arxiv.org/abs/1210.1139v2, Aug. 2013.

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

Broadcast Channel with Receiver Side Information: Achieving Individual Secrecy

O. Ozan Koyluoglu*, Yanling Chen†, Aydin Sezgin†
* Department of Electrical and Computer Engineering, The University of Arizona. Email: ozan@email.arizona.edu.
† Chair of Communication Systems, Ruhr University Bochum, Germany. Email: {yanling.chen-q5g, aydin.sezgin}@rub.de.

***Abstract*—In this paper, we study the problem of secure communication over the broadcast channel with receiver side information, under the lens of individual secrecy constraints (i.e., information leakage from each message to an eavesdropper is made vanishing). Several coding schemes are proposed by extending known results in broadcast channels to this secrecy setting. In particular, individual secrecy provided via one-time pad signal is utilized in the coding schemes. As a preliminary result, we obtain a general achievable region together with a characterization of the capacity region for the case of a degraded eavesdropper.**

I. Introduction

The broadcast channel is a fundamental communication model that involves transmission of independent messages to different users. In this paper, we consider the secure transmission of independent messages to two receivers which have, respectively, the desired message of the other receiver as side information. The model is shown in Fig. 1. The problem (without an eavesdropper) was originally motivated by the concept of the bidirectional relay channel, where two nodes exchange messages via a relay node. If the relay node decodes both messages, then it can broadcast a common codeword to both nodes each having their own message as side information. In [1], the broadcasting capacity region (without an eavesdropper) has been completely characterized.

The model of the broadcast channel with receiver side information (BC-RSI) with an external eavesdropper has been studied in [2]. The authors proposed achievable rate regions and outer bounds for a joint secrecy constraint, whereby the information leakage from *both* messages to the eavesdropper is made vanishing. Differently from [2], we review the problem under *individual* secrecy constraints that aim to minimize the information leakage from *each* message to the eavesdropper. Although individual secrecy constraints are by definition weaker than the joint one, they nevertheless provide an acceptable security strength that keeps each legitimate receiver away from an invasion of secrecy. In addition, a joint secrecy constraint can be difficult or even impossible to fulfill in certain cases. So, in this paper, our main concern is to characterize the fundamental limits of secure communications under the individual secrecy constraints for the BC-RSI model.

II. System model

Consider a discrete memoryless broadcast channel given by $p(y_1, y_2, z|x)$ with two legitimate receivers and one passive eavesdropper. The transmitter aims to send messages m_1, m_2 to receiver $1, 2$, respectively. Suppose x^n is the channel input, whilst y_1^n (at receiver 1), y_2^n (at receiver 2) and z^n (at eavesdropper), are the channel outputs. Besides, m_2 (available at receiver 1) and m_1 (available at receiver 2), serve also as side information that may help to decode the desired message. (Unless otherwise specified, we use capital letters for random variables and corresponding small cases for their realizations.)

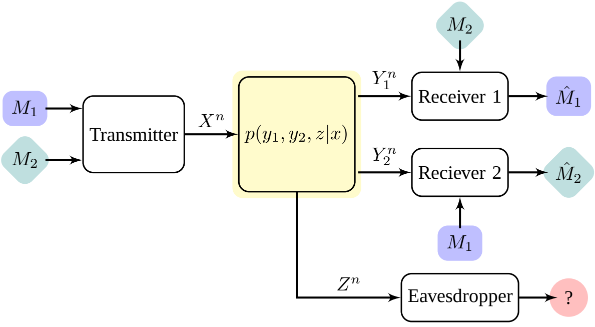


Fig. 1: Wiretap channel with receiver side information.

Denote the average probability of decoding error at receiver i to be $P_{e,i}$. The rate pair (R_1, R_2) is said to be *achievable*, if for any $\epsilon > 0$, there exists an encoder-decoder such that

$$\frac{1}{n}H(M_i) \geq R_i - \epsilon \tag{1}$$

$$P_{e,i} \leq \epsilon \tag{2}$$

$$\frac{1}{n}I(M_i; Z^n) \leq \epsilon, \tag{3}$$

for $i = 1, 2$ and for sufficiently large n. Equation (3) corresponds to *individual* secrecy constraints. If the coding scheme fulfills a stronger condition that

$$\frac{1}{n}I(M_1, M_2; Z^n) \leq \epsilon, \tag{4}$$

then it is said to satisfy the *joint* secrecy constraint.

We recall the capacity region of the discrete memoryless broadcast channel with receiver side information, when none of the secrecy constraints are taken into account.

Theorem 1. *([1, Theorem 1]) The capacity region of the discrete memoryless broadcast channel $p(y_1, y_2|x)$ with receiver side information is the set of the rate pairs (R_1, R_2) such that*

$$R_1 \leq I(X; Y_1) \quad \text{and} \quad R_2 \leq I(X; Y_2) \tag{5}$$

over all possible pmf $p(x)$.

III. Individual-Secrecy Rate Region

A. Secret key approach

Consider the symmetric secret rate region where $R_1 = R_2 = R$, i.e., M_1 and M_2 are of the same entropy. One can apply a one-time pad approach as proposed in [2]. With this scheme, the following rate region is achievable.

Proposition 2. *Any $(R_1, R_2) \in \mathbb{R}^+$ satisfying*

$$R_1 = R_2 \quad \leq \quad \min\{I(X;Y_1), I(X;Y_2)\} \tag{6}$$

for any $p(x)$ is achievable.

Proof: Randomly generate 2^{nR} codewords x^n according to $\prod_{i=1}^{n} p(x_i)$. Given (m_1, m_2), send $x^n(m_k)$ with $m_k = m_1 \oplus m_2$ to the channel. Both receivers can decode reliably by utilizing their side information to extract intended messages if $R_1 = R_2 \leq \min\{I(X;Y_1), I(X;Y_2)\}$.

For the secrecy constraint, we have for $i = 1, 2$,

$$I(M_i; Z^n) \leq I(M_i; Z^n, M_k) = I(M_i; M_k) = 0, \tag{7}$$

where the 1st equality is due to Markov chain $M_i \to M_k \to Z^n$; and the 2nd is since M_k is a one-time pad of M_i. ■

Note that the above achievable region is limited by the worse channel. In the following, we consider other coding schemes to enlarge the achievable region beyond the one stated above.

B. Secrecy coding approach

Consider those channel inputs $p(x)$ such that $I(X;Z) \leq \min\{I(X;Y_1), I(X;Y_2)\}$. Assume that $I(X;Y_2) \leq I(X;Y_1)$. For such cases, we split M_1 into two parts: one of entropy $n(I(X;Y_1) - I(X;Y_2))$ which is secured by using secrecy coding for classical wiretap channels; and the other of entropy $nI(X;Y_2)$ which is secured by capsuling with M_2 in a one-time pad (thus M_2 is also secured). We obtain the following.

Proposition 3. *Any $(R_1, R_2) \in \mathbb{R}^+$ satisfying*

$$I(X;Z) \leq R_1 \leq I(X;Y_1); \;\; I(X;Z) \leq R_2 \leq I(X;Y_2) \tag{8}$$

for $p(x)$ such that $I(X;Z) \leq \min\{I(X;Y_1), I(X;Y_2)\}$ is achievable.

Proof: Assume that $R_2 \leq R_1$. We split M_1 into two parts, i.e., $M_1 = (M_{1k}, M_{1s})$ with M_{1k} of entropy nR_2, the same as M_2; whilst M_{1s} of entropy $n(R_1 - R_2)$.

Randomly generate 2^{nR_1} codewords x^n according to $\prod_{i=1}^{n} p(x_i)$. Throw them into $2^{n(R_1-R_2)}$ bins [3] and index $x^n(i_k, i_{1s})$ with $(i_k, i_{1s}) \in [1 : 2^{nR_2}] \times [1 : 2^{n(R_1-R_2)}]$.

Given (m_1, m_2), send $x^n(m_k, m_{1s})$ with $m_k = m_{1k} \oplus m_2$ to the channel. Receiver 2 can decode m_k reliably using typical set decoding if $R_2 < I(X;Y_2)$ with the help of m_1, and thus extract m_2. Receiver 1 can decode both m_k and m_{1s} if $R_1 < I(X;Y_1)$, and extract m_{1k} from the former given m_2.

At the eavesdropper, for the secrecy of M_2, we have

$$I(M_2; Z^n) \leq I(M_2; Z^n, M_k, M_{1s}) = I(M_2; M_k, M_{1s}) = 0,$$

Further, the secrecy of M_1 is shown as follows. Since $R_2 \geq I(X;Z)$, for a fixed i_{1s}, one can further bin the codewords x^n and index them as $x^n(i_{kx}, i_{ks}, i_{1s})$ with $i_k = (i_{kx}, i_{ks}) \in [1 : 2^{n(I(X;Z)-\epsilon)}] \times [1 : 2^{n(R_2 - I(X;Z)+\epsilon)}]$. Correspondingly, split $M_k = (M_{kx}, M_{ks})$. We have

$$\begin{aligned}
&H(M_{1s}, M_{ks}|Z^n) \\
&\quad = H(M_{1s}, M_{ks}, X^n|Z^n) - H(X^n|M_{1s}, M_{ks}, Z^n) \\
&\quad \overset{(a)}{\geq} H(M_{1s}, M_{ks}, X^n, Z^n) - H(Z^n) - n\epsilon_1 \\
&\quad = H(X^n) + H(Z^n|X^n) - H(Z^n) - n\epsilon_1 \\
&\quad \overset{(b)}{\geq} nR_1 + nH(Z|X) - nH(Z) - n\epsilon_1 \\
&\quad \overset{(c)}{\geq} H(M_{1s}, M_{ks}) - n\delta(\epsilon),
\end{aligned}$$

where (a) follows as $H(X^n|M_{1s}, M_{ks}, Z^n) \leq n\epsilon_1$ due to Fano's inequality and that the eavesdropper can decode X^n reliably, given (M_{ks}, M_{1s}, Z^n); (b) is due to the fact that $H(X^n) = nR_1$; $H(Z^n|X^n) = nH(Z|X)$ since the channel is memoryless; and $H(Z^n) = \sum_{i=1}^{n} H(Z_i|Z_1^{i-1}) \leq \sum_{i=1}^{n} H(Z_i) = nH(Z)$; (c) is due to the fact that $H(M_{1s}, M_{ks}) = n(R_1 - R_2) + n(R_2 - I(X;Z) + \epsilon)$.

Above inequality implies $I(M_{1s}; Z^n) \leq n\delta(\epsilon)$. In addition, we bound $I(M_{1k}; Z^n|M_{1s}) \leq I(M_{1k}; Z^n, M_{1s}, M_k) = I(M_{1k}; M_k, M_{1s}) = 0$ due to Markov chain $M_{1k} \to (M_k, M_{1s}) \to Z^n$. Therefore, we obtain

$$I(M_1; Z^n) = I(M_{1s}; Z^n) + I(M_{1k}; Z^n|M_{1s}) \leq n\delta(\epsilon).$$

This concludes the individual secrecy proof. ■

Proposition 4. *If the channel to the eavesdropper is degraded with respect to the channels of both legitimate receivers, then the individual-secrecy capacity region is given by the union of $(R_1, R_2) \in \mathbb{R}^+$ pairs satisfying*

$$\begin{aligned}
R_1 &\leq \min\{I(X;Y_1) - I(X;Z) + R_2, I(X;Y_1)\}; \\
R_2 &\leq \min\{I(X;Y_2) - I(X;Z) + R_1, I(X;Y_2)\},
\end{aligned} \tag{9}$$

where the union is taken over $p(x)$.

Proof: With the *degraded* condition, we have $I(X;Z) \leq \min\{I(X;Y_1), I(X;Y_2)\}$ for any $p(x)$. Denote $\mathcal{R}_1$ to be the region achievable by Proposition 3, as defined in (8) . Further, denote $\mathcal{R}_2 = \{(R_1, R_2) : R_1 = 0, R_2 \leq I(X;Y_2) - I(X;Z)\}$ and $\mathcal{R}_3 = \{(R_1, R_2) : R_1 \leq I(X;Y_1) - I(X;Z), R_2 = 0\}$, which are achievable by employing Wyner's secrecy coding. The achievability of the region in (9) follows from the convex hull of $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3$. The converse follows directly from Theorem 1 together with Proposition 7 provided below. ■

C. Superposition coding

Consider a degraded broadcast channel where $X \to Y_1 \to Y_2$ forms a Markov chain. Then, one can utilize superposition coding to transmit a cloud center to the weak receiver and both the cloud center and satellite codewords to the strong receiver [3]. By utilizing the one-time pad message as the cloud center, one can readily achieve the following region.

Proposition 5. *The individual-secrecy rate region for BC-RSI is achievable for the set of the rate pairs (R_1, R_2) such that*

$$R_t = I(U;Y_t); \;\; R_{\bar{t}} \leq I(V;Y_{\bar{t}}|U) - I(V;Z|U) + R_t, \tag{10}$$

over all $p(u)p(v|u)p(x|v)$*, where* $t = \arg\min_{i\in\{1,2\}}\{I(U;Y_i)\}$ *and* $\bar{t} = \{1,2\}\backslash\{t\}$*.*

Proof: Assume that $R_2 \leq R_1$. (This corresponds to the case $t = 2$ in which $I(U;Y_2) \leq I(U;Y_1)$, since V can be always chosen such that $I(V;Y_{\bar{t}}|U) - I(V;Z|U)$ is non-negative). Represent M_1 by (M_{1k}, M_{1s}), with M_{1k} of entropy nR_2, the same as that of M_2 and M_{1s} of entropy $n(R_1 - R_2)$.

Codebook generation: Fix $p(u), p(v|u)$. First, randomly generate 2^{nR_2} i.i.d sequences $u^n(k)$, $k \in [1:2^{nR_2}]$, according to $\prod_{i=1}^n p(u_i)$. Secondly, for each $u^n(k)$, according to $\prod_{i=1}^n p(v_i|u_i)$, randomly generate i.i.d sequences $v^n(k,s,r)$ with $(s,r) \in [1:2^{n(R_1-R_2)}] \times [1:2^{n(I(V;Z|U)-\epsilon)}]$.

Encoding: To send messages (m_1, m_2), choose $u^n(k)$, where $k = m_k \triangleq m_{1k} \oplus m_2$. Given $u^n(k)$, randomly choose $r \in [1:2^{n(I(V;Z|U)-\epsilon)}]$ and find $v^n(k, m_{1s}, r)$. Generate x^n according to $\prod_{i=1}^n p(x_i|v_i)$, and transmit it to the channel.

Decoding: Receiver 2, upon receiving y_2^n, finds $u^n(\hat{k})$ such that $(u^n(\hat{k}), y_1^n)$ is jointly typical. (It is necessary that $R_2 < I(U;Y_2)$.) With the knowledge of m_1, decode $\hat{m}_2 = m_{1k} \oplus \hat{k}$.

Receiver 1, upon receiving y_1^n, finds $u^n(\hat{k})$ such that $(u^n(\hat{k}), y_1^n)$ is jointly typical. (This is possible since $R_2 < I(U;Y_2) \leq I(U;Y_1)$.) Corresponding to $u^n(\hat{k})$, further find $v^n(\hat{k}, \hat{m}_{1s}, \hat{r})$ which is jointly typical with y_1^n. With the knowledge of m_2, decode $\hat{m}_1 = (m_2 \oplus \hat{k}, \hat{m}_{1s})$.

Analysis of the probability error: Similar to the analysis of the superposition coding for general discrete memoryless broadcast channels, we have $P_{e,1}, P_{e,2} \to 0$ as $n \to \infty$ if $R_2 < I(U;Y_2) - \epsilon$ and $R_1 < I(V;Y_1|U) - I(V;Z|U) + R_2 - \epsilon$.

Analysis of individual secrecy: For the secrecy of M_2, due to the Markov chain $M_2 \to (M_k, M_{1s}) \to Z^n$, we have $I(M_2;Z^n) \leq I(M_2;Z^n, M_k, M_{1s}) = I(M_2;M_k, M_{1s}) = 0$, where the last equality is due to the fact that $M_k = M_2 \oplus M_{1k}$, is independent of M_2 as its one-time pad encryption.

For the secrecy of M_1, we have

$$
\begin{aligned}
I(M_1;Z^n) =& I(M_{1k}, M_{1s};Z^n) && (11)\\
=& I(M_{1k};Z^n) + I(M_{1s};Z^n|M_{1k}) && (12)\\
\overset{(a)}{=}& I(M_{1s};Z^n|M_{1k}) && (13)\\
\leq& I(M_{1s};Z^n, M_{1k}, M_k) && (14)\\
=& I(M_{1s};Z^n, M_k) + I(M_{1s};M_{1k}|Z^n, M_k) && (15)\\
\overset{(b)}{=}& I(M_{1s};Z^n, M_k) && (16)\\
=& H(M_{1s}) - H(M_{1s}|M_k, Z^n) && (17)\\
=& n(R_1 - R_2) - H(M_{1s}|M_k, Z^n), && (18)
\end{aligned}
$$

where (a) is due to the fact that $I(M_{1k};Z^n) = 0$ by following a similar proof of $I(M_2;Z^n) = 0$; (b) follows that $I(M_{1s};M_{1k}|Z^n, M_k) \geq 0$ and that $H(M_{1k}|Z^n, M_k, M_{1s}) = H(M_{1k}|M_k, M_{1s}) = H(M_{1k}) \geq H(M_{1k}|Z^n, M_k)$.

To complete the proof that $I(M_1;Z^n) \leq n\delta(\epsilon)$, we show in the following that $H(M_{1s}|M_k, Z^n) \geq n(R_1 - R_2) - n\delta(\epsilon)$.

$$
\begin{aligned}
H(M_{1s}|M_k, Z^n) \overset{(c)}{=}& H(M_{1s}|U^n, Z^n)\\
=& H(M_{1s}, Z^n|U^n) - H(Z^n|U^n)\\
=& H(M_{1s}, Z^n, V^n|U^n)\\
&- H(V^n|U^n, M_{1s}, Z^n) - H(Z^n|U^n)\\
=& H(V^n|U^n) + H(Z^n|U^n, V^n)\\
&- H(V^n|U^n, M_{1s}, Z^n) - H(Z^n|U^n)\\
\overset{(d)}{\geq}& n(R_1 - R_2) - n\delta(\epsilon),
\end{aligned}
$$

where (c) is due to the fact that U^n is uniquely determined by M_k; (d) follows as $H(V^n|U^n) = n(R_1 - R_2) + n(I(V;Z|U) - \epsilon)$ by codebook construction; $H(Z^n|U^n, V^n) = \sum_{i=1}^n H(Z_i|U_i, V_i) = nH(Z|U,V)$ since the channel is discrete memoryless; $H(V^n|U^n, M_{1s}, Z^n) \leq n\epsilon$ due to Fano's inequality and that the eavesdropper can decode V^n reliably, given (U^n, M_{1s}, Z^n); and $H(Z^n|U^n) = \sum_{i=1}^n H(Z_i|Z^{i-1}, U^n) \leq \sum_{i=1}^n H(Z_i|U_i) = nH(Z|U)$. ■

D. Marton's coding

A universal approach is to apply Marton's coding for the general broadcast channels, utilizing the one-time pad message as common message to transmit secure messages to both users.

Proposition 6. *The rate region is given by* $(R_1 = R_k + R_{1s}, R_2 = R_k + R_{2s})$ *pairs such that* (R_k, R_{1s}, R_{2s}) *belongs to the region given by the union of rate tuples*

$$
\begin{aligned}
R_k \leq& \min\{I(U;Y_1), I(U;Y_2)\}\\
R_{1s} \leq& \min\{I(V_1, V_2;Y_1|U) - R_0, I(V_1;Y_1, V_2|U)\}\\
R_{2s} \leq& \min\{I(V_1, V_2;Y_2|U) - R_0, I(V_2;Y_2, V_1|U)\}\\
R_{1s} + R_{2s} \leq& I(V_1;Y_1, V_2|U) + I(V_2;Y_2, V_1|U) - R_0
\end{aligned}
$$

over any pmf $p(u)p(v_1, v_2|u)p(x|v_1, v_2)$*, where* $R_0 = I(V_1;V_2|U) + I(V_1, V_2;Z|U)$*.*

Proof: Represent M_1, M_2 by $M_1 = (M_{1k}, M_{1s})$ and $M_2 = (M_{2k}, M_{2s})$ with M_{1k}, M_{2k} of entropy nR_k; whilst M_{1s} of entropy nR_{1s} and M_{2s} of entropy nR_{2s}.

Codebook generation: Fix $p(u), p(v_1|u), p(v_2|u)$ and $p(x|v_1, v_2)$. First, randomly generate 2^{nR_k} i.i.d sequences $u^n(k)$, $k \in [1:2^{nR_k}]$, according to $\prod_{i=1}^n p(u_i)$.

For each $u^n(k)$, randomly generate $2^{n(R_{1s}+R_{1c}+R_{1r})}$ i.i.d sequences $v_1^n(k, s_1, c_1, r_1)$ with $(s_1, c_1, r_1) \in [1:2^{nR_{1s}}] \times [1:2^{nR_{1c}}] \times [1:2^{nR_{1r}}]$, according to $\prod_{i=1}^n p(v_{1i}|u_i)$; and similarly generate $2^{n(R_{2s}+R_{2c}+R_{2r})}$ i.i.d sequences $v_2^n(k, s_2, c_2, r_2)$, $(s_2, c_2, r_2) \in [1:2^{nR_{2s}}] \times [1:2^{nR_{2c}}] \times [1:2^{nR_{2r}}]$, according to $\prod_{i=1}^n p(v_{2i}|u_i)$. For a fixed (k, s_1, s_2), we denote the product $V_1 \times V_2$ codebook to be $\mathcal{C}_{V_1,V_2|U}(k, s_1, s_2)$.

Encoding: To send messages (m_1, m_2), choose $u^n(k)$, where $k = m_k \triangleq m_{1k} \oplus m_{2k}$. Given $u^n(k)$, find in the product codebook $\mathcal{C}_{V_1,V_2|U}(k, m_{1s}, m_{2s})$ a jointly typical $(v_1^n(k, m_{1s}, c_1, r_1), v_2^n(k, m_{2s}, c_2, r_2))$ pair. (This is possible if $R_{1c} + R_{2c} > I(V_1;V_2|U)$). Generate and transmit $x^n(v_1^n, v_2^n)$ according to $\prod_{i=1}^n p(x_i|v_{1i}, v_{2i})$.

Decoding: Receiver 1, upon receiving y_1^n, finds $u^n(\hat{k})$ such that $(u^n(\hat{k}), y_1^n)$ is jointly typical. (It is necessary that $R_k < I(U;Y_1)$). With the knowledge of m_2 and $u^n(\hat{k})$, further find $(v_1^n(\hat{k}, \hat{m}_{1s}, \hat{c}_1, \hat{r}_1), v_2^n(\hat{k}, m_{2s}, \hat{c}_2, \hat{r}_2))$, which is jointly typical with y_1^n. Decode $\hat{m}_1 = (m_{2k} \oplus \hat{k}, \hat{m}_{1s})$.

Receiver 2, upon receiving y_2^n, finds $u^n(\hat{k})$ such that $(u^n(\hat{k}), y_1^n)$ is jointly typical. (It is necessary that $R_k < I(U;Y_2)$). With the knowledge of m_1 and $u^n(\hat{k})$, further find $(v_1^n(\hat{k}, m_{1s}, \hat{c}_1, \hat{r}_1)), v_2^n(\hat{k}, \hat{m}_{2s}, \hat{c}_2, \hat{r}_2))$, which is jointly typical with y_2^n. Decode $\hat{m}_2 = (m_{1k} \oplus \hat{k}, \hat{m}_{2s})$.

Analysis of decoding error: For $P_{e,1}$ (similar for $P_{e,2}$), a decoding error happens iff ≥ 1 of the following events occur:

$$\begin{aligned}
\mathcal{E}_{11} =& \{(u^n(k), y_1^n) \notin \mathcal{T}_\epsilon^{(n)}\},\\
\mathcal{E}_{12} =& \{(v_1^n(k, m_{1s}, c_1, r_1), v_2^n(k, m_{2s}, c_2, r_2)) \notin \mathcal{T}_\epsilon^{(n)}\},\\
\mathcal{E}_{13} =& \{(v_1^n(k, m_{1s}, c_1, r_1), v_2^n(k, m_{2s}, c_2, r_2), y_1^n) \notin \mathcal{T}_\epsilon^{(n)})\},\\
\mathcal{E}_{14} =& \{(v_1^n(k, m'_{1s}, c'_1, r'_1), v_2^n(k, m_{2s}, c'_2, r'_2), y_1^n) \in \mathcal{T}_\epsilon^{(n)},\\
& \quad m'_{1s} \neq m_{1s}\}.
\end{aligned}$$

The probability of error $P_{e,1}$ is upper bounded as $P_{e,1} \leq \Pr(\mathcal{E}_{11}) + \Pr(\mathcal{E}_{12}|\mathcal{E}_{11}^c) + \Pr(\mathcal{E}_{13}|\mathcal{E}_{11}^c, \mathcal{E}_{12}^c) + \Pr(\mathcal{E}_{14}|\mathcal{E}_{11}^c)$. By the LLN, $\Pr(\mathcal{E}_{11})$ and $\Pr(\mathcal{E}_{13}|\mathcal{E}_{11}^c, \mathcal{E}_{12}^c)$ tend to zero as $n \to \infty$; $\Pr(\mathcal{E}_{12}|\mathcal{E}_{11}^c)$, by the mutual covering lemma [3] , tends to zero as $n \to \infty$ since $R_{1c} + R_{2c} > I(V_1; V_2|U) + \epsilon$; The 4th term, $\Pr(\mathcal{E}_{14}|\mathcal{E}_{11}^c)$, by the packing lemma [3], tends to zero as $n \to \infty$ if $R_{1s} + R_{1c} + R_{2c} + R_{1r} + R_{2r} < I(V_1, V_2; Y_1|U) - \epsilon$, and $R_{1s} + R_{1c} + R_{1r} < I(V_1; Y_1, V_2|U) - \epsilon$.

Analysis of individual secrecy: For the secrecy of M_1 (similar for M_2), we follow the steps in (11)-(17) and obtain

$$I(M_1; Z^n) \leq nR_{1s} - H(M_{1s}|M_k, Z^n). \tag{19}$$

In the following, we show that $H(M_{1s}, M_{2s}|M_k, Z^n) \geq n(R_{1s} + R_{2s}) - n\delta'(\epsilon)$ holds if we take $R_{1r} + R_{2r} = I(V_1, V_2; Z|U) - \epsilon$. This implies that $H(M_{1s}|M_k, Z^n) \geq nR_{1s} - n\delta(\epsilon)$; and by (19) we obtain $I(M_1; Z^n) \leq n\delta(\epsilon)$.
$H(M_{1s}, M_{2s}|M_k, Z^n)$

$$\begin{aligned}
&= H(M_{1s}, M_{2s}, Z^n|U^n) - H(Z^n|U^n)\\
&\overset{(a)}{\geq} H(M_{1s}, M_{2s}, Z^n|W_{1c}, W_{2c}, U^n) - H(Z^n|U^n)\\
&= H(M_{1s}, M_{2s}, Z^n, V_1^n, V_2^n|W_{1c}, W_{2c}, U^n) - H(Z^n|U^n)\\
&\quad - H(V_1^n, V_2^n|W_{1c}, W_{2c}, U^n, M_{1s}, M_{2s}, Z^n)\\
&\overset{(b)}{\geq} H(M_{1s}, M_{2s}, Z^n, V_1^n, V_2^n|W_{1c}, W_{2c}, U^n)\\
&\quad - H(Z^n|U^n) - n\epsilon\\
&= H(M_{1s}, M_{2s}, V_1^n, V_2^n|W_{1c}, W_{2c}, U^n) - H(Z^n|U^n) - n\epsilon\\
&\quad + H(Z^n|W_{1c}, W_{2c}, U^n, M_{1s}, M_{2s}, V_1^n, V_2^n)\\
&= n(R_{1s} + R_{2s} + R_{1r} + R_{2r}) + H(Z^n|U^n, V_1^n, V_2^n)\\
&\quad - H(Z^n|U^n) - n\epsilon\\
&\overset{(c)}{\geq} n(R_{1s} + R_{2s}) - n\delta'(\epsilon)
\end{aligned}$$

where (a) follows by introducing random variable W_{1c}, W_{2c} for the covering indices c_1, c_2; (b) follows from the fact that the eavesdropper can decode V_1^n, V_2^n reliably given $(U^n, M_{1s}, M_{2s}, W_{1c}, W_{2c}, Z^n)$; (c) follows that $H(Z^n|U^n) \leq nH(Z|U)$ and $H(Z^n|U^n, V_1^n, V_2^n) = nH(Z|U, V_1, V_2)$ and additionally by the rate choice $R_{1r} + R_{2r} = I(V_1, V_2; Z|U) - \epsilon$.

Adding those conditions such that $P_{e,1}, P_{e,2} \to 0$ as $n \to \infty$ to the rate choice $R_{1r} + R_{2r} = I(V_1, V_2; Z|U) - \epsilon$, we have

$$\begin{aligned}
R_k &\leq \min\{I(U;Y_1), I(U;Y_2)\}\\
R_{1c} + R_{2c} &\geq I(V_1; V_2|U)\\
R_{is} + R_{1c} + R_{2c} + R_{1r} + R_{2r} &\leq I(V_1, V_2; Y_i|U) \quad \text{for } i = 1, 2\\
R_{1s} + R_{1c} + R_{1r} &\leq I(V_1; Y_1, V_2|U)\\
R_{2s} + R_{2c} + R_{2r} &\leq I(V_2; Y_2, V_1|U)
\end{aligned}$$

Eliminating $R_{1c}, R_{2c}, R_{1r}, R_{2r}$ by applying Fourier-Motzkin procedure [3], we get the desired region of (R_k, R_{1s}, R_{2s}). ■

Remark: Setting $U, Y_2, V_2 = \emptyset$, the region coincides with the secrecy capacity region of the wiretap channel [4]; If we let $U = \emptyset$, it reduces to an achievable region under the joint secrecy constraint (indicated by the above secrecy proof).

E. Upper bounds

For the individual secrecy capacity region of BC-RSI, an obvious upper bound is the capacity region of the BC-RSI without an eavesdropper as given in Theorem 1. Another upper bound follows directly the work of wiretap channel with shared key [5], as stated in the following proposition.

Proposition 7. *For any R_2 in the achievable region, R_1 is upper bounded by*

$$\max_{U \to V \to X \to (Y_1, Z)} \min\{I(V; Y_1|U) - I(V; Z|U) + R_2, I(V; Y_1)\}.$$

If the channel is degraded such that $X \to Y_1 \to Z$, then for any R_2 in the achievable region, R_1 is upper bounded by

$$\max_{X \to Y_1 \to Z} \min\{I(X; Y_1) - I(X; Z) + R_2, I(X; Y_1)\}.$$

Similar results hold for interchanging 1 and 2 above.

IV. Conclusion

In this paper, we studied the problem of secure communication over BC-RSI under the individual secrecy constraints. Compared to the joint secrecy constraint, this relaxed setting allows for higher secure communication rates at the expense of having a weaker notion of security. We provide some special case results together with several achievable schemes; whilst the characterization for the general case still remains as an open problem.

References

[1] G. Kramer and S. Shamai, "Capacity for classes of broadcast channels with receiver side information," in *Proc. 2007 IEEE Information Theory Workshop (ITW)*, pp. 313–318, Sep. 2007.

[2] R. Wyrembelski, A. Sezgin, and H. Boche, "Secrecy in broadcast channels with receiver side information," in *Proc. 45th Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 290–294, Nov. 2011.

[3] Abbas El Gamal and Young-Han Kim, *Network Information Theory*, Cambridge University Press, 2011.

[4] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 339–348, May 1978.

[5] W. Kang and N. Liu, "Wiretap channel with shared key," in *Proc. 2010 IEEE Information Theory Workshop (ITW)*, pp. 1–5, Sep. 2010.

Energy Limits of Message-Passing Error Control Decoders

Christian Schlegel
Dalhousie University
UMDCC
Halifax, Nova Scotia, Canada
Email: Christian.Schlegel@Dal.ca

Christopher Winstead
Department of Electrical and Computer Engineering
Utah State University
Logan, UT, United States of America
Email: Chris.Winstead@usu.edu

***Abstract*—Modern message-passing error control decoders are studied in regards to the processing energy required to extract digital information from a noisy received signal. It is shown that fundamental charge-based computational models, together with limits of error-free message passing along the processors' communications network, imply a lower limit to the energy efficiency achievable for such decoders in modern and future VLSI implementation technologies. The limiting energy of node processing is estimated for belief propagation decoding of LDPC codes, using estimates of nodes' internal processing activity. The limiting energy of message passing is estimated by using an energy-annotated density evolution procedure. For the class of decoders studied, the minimum energy is found to be on the order of 0.4 fJ per bit for node processing, and 9.86 aJ per bit for message passing.**

I. Introduction

As the miniaturization of very large scale integrated circuits continues to advance at an exponential pace, the power consumed by these circuits is becoming an ever more limiting problem, even as the computational resources — i.e. the density and number of switching devices — are becoming ever more available to accommodate even the most complex digital algorithms.

Integrated circuits (chips) with more than one billion transistors are now quite commonplace, and that growth is projected to continue at least for the next decade. The International Technology Roadmap for Semiconductors illustrates the exponential empirical law, known as *Moore's Law*, that appears to be underlying this progress.

The process is driven largely by the decreasing feature sizes of integrated circuits. Subsequent generations of fabrication technologies are known collectively as process nodes. Current leading-edge technology has arrived at the 22nm process node, and further miniaturization towards single digit nano-meter scales appears assured. Clearly there are immense challenges that are facing the industry in pushing this miniaturization forward, and we will not further concern ourselves with these, but accept this trend.

Instead, we will concentrate on the minimum power that is required to operate circuits of any size, in particular those at the most advanced nodes. In order to lead this discussion, we need to have a suitably general computational model of switching devices and the energy dissipated during an operation. The ubiquitous complementary metal-oxide semiconductor (CMOS) technology uses two complementary devices as shown in Fig. 1. This CMOS gate operates as a basic switch and thus is capable of processing binary information. The switch operates by charging and discharging a holding capacitance (usually the interconnect capacitance combined with the input capacitance of the next gate(s)).

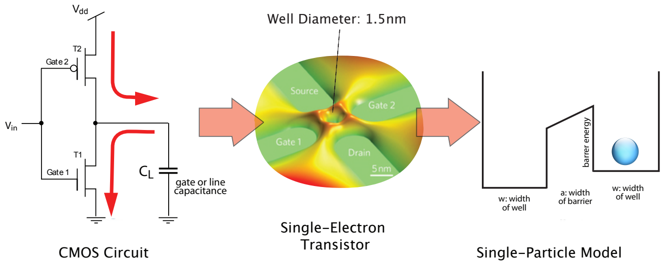


Fig. 1. Generic structure of a CMOS logic gate, and the evolution from MOS technologies toward single charge-based switching devices.

When the gate switches states it moves charge between the supply reservoir (V_{dd}) and ground. At each such transition, an amount of energy equal to $C_I V_{\mathrm{dd}}^2$ is dissipated into heat. When considering the future evolution of digital technologies to the quantum scale, as illustrated in Fig. 1, the energy of transition may be described as the work done when moving a particle across a potential barrier. The evolution toward single charge switches is inevitable, particularly since single-electron transistors and other quantum-scale devices have been demonstrated in the laboratory (e.g. [1]).

One can ask the question what the limits are of such a charge-based computational model. This is precisely the exercise that Zhirnov et al. [2] conducted, using the Landauer limit for irreversible computing [3], which postulates that the minimum amount of energy released in the (irreversible) processing of one bit of information is bounded by

$$E \geq E_L = k_B T \ln 2 = 0.017 \text{ eV}, \tag{1}$$

where $T = 300$ K and $1 \text{ eV} = 1.602 \times 10^{-19}$ V. Applying Heisenberg's uncertainty principle Zhirnov computed minimal size and switching times for such an $E = E_L$ minimum-energy switch as

$$x_{\min} = \frac{h/2\pi}{\sqrt{2m_e E}} = 1.5\text{nm}; \quad t_{\min} = \frac{h/2\pi}{E} = 0.04\text{ps} \tag{2}$$

These figures in turn imply a maximum integration density $\nu_{\max}$ of such minimum size switches, and a maximum power density $P_{\max}$ of

$$\nu_{\max} = 4.7 \times 10^{13} \text{ devices/cm}^2; \;\; P_{\max} = 3.7 \times 10^6 \text{ W/cm}^2 \tag{3}$$

The Landauer limit applies to a single charge storage model shown in Fig. 1, where a charged particle is in one of two wells, separated by an energy barrier. The minimum energy needed to move the charge is precisely E_L. Meindl also showed that the Landauer limit can be obtained when considering only the operation of an "ideal" MOSFET device operating in its subthreshold region [4]. Hence the Landauer model is quite applicable to the CMOS structure from Fig. 1.

Since the Landauer limit (1) is applicable to irreversible operations, one might speculate that it can be circumvented by using reversible or adiabatic computational circuits. Meindl and Davis [4] disposed of this possibility by showing that the Landauer limit can also be obtained by solving the Shannon capacity of an interconnect wire in the presence of thermal noise, and hence can be interpreted as a limit on signalling between gates and modules, one that cannot be circumvented by technology choices.

In the sequel we will study the two main aspects of energy consumption that affect an error control decoder, in particular a decoder for low-density parity check (LDPC) codes. LDPC codes are among a class of very powerful error control codes which can achieve the theoretical limits on communications, known as the Shannon limit. They have a low-complexity iterative decoding algorithm, which not only allows the construction of decoders for very large codes, but also has made these codes the de-facto standard in many communications applications. With error control coding now applied to very high speed applications, such as the 802.3 10Gbit/s modems, or optical communications systems at even larger rates, their power consumption has become a targeted concern that appears to be limiting future applications.

Fig. 2 shows the processing structure of an LDPC code. The two processes that need to be executed are local computation, summarized in the figure, and communication to connected nodes. These will be analyzed from a power consumption viewpoint.

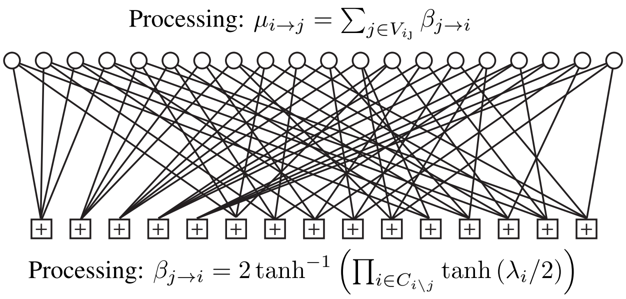

Processing: $\mu_{i \to j} = \sum_{j \in V_{i_j}} \beta_{j \to i}$
Processing: $\beta_{j \to i} = 2 \tanh^{-1} \left(\prod_{i \in C_{i \setminus j}} \tanh(\lambda_i/2) \right)$

Fig. 2. Network of an LDPC code. Their typical size is 1000–10,000 nodes.

II. Node Processing

Without going into the full implementation details of the local processing operations at the nodes, it can be shown that the computational complexity of each node is $\mathcal{O}(W)$, where W is the number of binary digits used in the number representation of a digital decoder. W is typically between 6–12 bits, and has a subtle impact on the performance of the code – see [8]. More specifically, activity simulations showed that there were on the average 2.7 digital transitions per bit and message line at the variable nodes (top nodes in Fig. 2), and 3.3 digital transitions per outgoing message in the check nodes (bottom nodes).

We approach the computational power limit in the following way: for each transition that occurs during decoding, a minimal amount of energy proportional to E_L is dissipated. The problem now is that if the barrier is set at the minimum energy E_L, there is a significant over-barrier probability $P = e^{-E_{\text{barrier}}/kT}$, which makes such a cell very unreliable. In fact, the over-barrier probability reaches 50% at E_L, making the cell quite useless for computation. To keep our nodes operating at acceptable levels of reliability, the barrier energy needs to exceed E_L. We somewhat arbitrarily assume that a factor $K = 10$, which leads to an over-barrier probability of 10^{-4}, can both be realized in the future, and is acceptable in the algorithm.

With these assumptions, the node processing per information bit is lower bounded by

$$E_b > K E_L I W (2.7 d_v + 3.3(d_c - 1)) \tag{4}$$

where I is the number of iterations in the code network, usually on the order of 5–20. Clearly, varying W, I, and the code parameters d_v, d_c can move this number by an order of magnitude or so, and the limit is to be seen as that for an average code. With $K = 20, I = 10, d_v = 3, d_c = 6, W = 8$, we obtain $E_b > 0.1\text{fJ}$.

Expressed in kT, the processing requirements are on the order of 25,000 kT per bit. In the next section we will argue that transporting the messages between the local nodes will incur an energy effort that is comparable to the energy expended in the computations themselves, thus arriving at a lower energy limit for a charge-based error control decoder.

III. Network Communication

In this section we address the second power-intensive portion of a message-passing error control decoder, which is the communication of the node messages along branches of the code network (see Fig. 2), by evaluating the minimum energy cost associated with transporting messages. The decoder is parameterized by the traditional degree distribution, by symbol and check node update algorithms, and by the *message representation* used for transmitting messages between check and symbol nodes. The message representation is defined as the mapping from possible messages to a set of corresponding physical signals. A similar approach has been taken previously to minimize activity in digital LDPC decoders (e.g. by Gaudet and Crowley [5], [6]).

The density evolution method is modified to examine the effects of internal signalling near the kT noise limit in electrical interconnects. By considering the physical signals passed between nodes, we obtain technology-independent conclusions about the minimum energy associated with error-free decoding. Low-energy signals are subject to upsets from kT noise, but a decoder is able to self-correct many such upsets.

In order to evaluate signalling energy limits in message

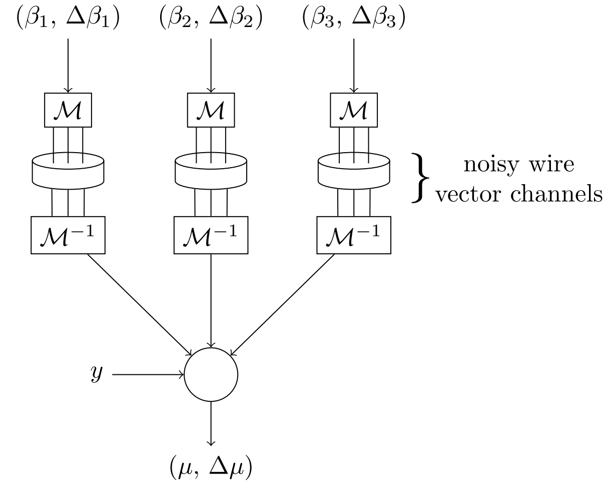


Fig. 3. Illustration of the density estimation procedure for a symbol node with $d_v = 3$. The incoming messages $(\beta_i,\,\Delta\beta_i)$ are generated from the joint distribution. The β messages are used to compute outgoing message μ. Then the $\beta + \Delta\beta$ messages are used to compute $\mu + \Delta\mu$. The energy is a function of $\Delta\beta$ and $\Delta\mu$.

passing, we consider that each signal wire is affected by Gaussian white thermal noise with energy kT Joules per sample. Physical signals are assumed to be voltages, specified in units of $(kT)^{1/2}$. The normalized energy of physical signals is estimated by assuming unit interconnect capacitance, so that $E_{\text{signal}} = \alpha V_{\text{dd}}^2$, where α is the *activity* defined as the wire's frequency of transitions. The unknown capacitance C_I is treated as a technology-dependent scale constant. The precise value of C_I has no influence on the energy calculations, since V_{dd} can be varied proportional to $C_I^{-1/2}$ (we could equivalently say that V_{dd} has units $(kT/C_I)^{1/2}$).

Based on this model, each signal wire is represented as an additive white Gaussian noise channel with zero mean and unit variance. For example, suppose a binary value $x \in \{0,\,1\}$ is transmitted as part of the message from a symbol node to a check node. The value is represented as a voltage $v_x \in \{0,\,V_{\text{dd}}\}$, and transmitted across the unit-variance AWGN channel, which adds a noise sample n. At the channel's output, the received signal $v_y = v_x + n$ is resolved to a binary value $y \in \{0,\,1\}$ by applying a threshold at $V_{\text{dd}}/2$.

In order to account for the energy per message in a fully-parallel LDPC decoder, it is necessary to track the joint distributions of messages and their transitions, as shown in Fig. 3. The detailed procedure was described by Gaudet [5], and is only briefly summarized here. The symbols $\beta\,(t)$ and $\Delta\beta\,(t+1)$ refer to messages passed from a check to symbol nodes, and the changes in those messages, respectively, during iteration t. Similarly, the symbols $\mu\,(t)$ and $\Delta\mu\,(t+1)$ refer to messages passed from symbol to check nodes and their changes during iteration t. Note that the dependence on t is dropped when there is no ambiguity.

To estimate the energy required for message passing, density evolution is performed as usual while tracking the joint distributions for $(\beta,\,\Delta\beta)$ and $(\mu,\,\Delta\mu)$. The symbol node estimation is performed as follows. Random samples are generated jointly for both β and $\Delta\beta$. The symbol node update equations are used to obtain μ and $\Delta\mu$. To compute the sample's transition energy, both μ and $\mu + \Delta\mu$ are mapped to their corresponding message representations. If a wire's signal value at iteration t is $v_x\,(t)$, then the wire's normalized transition energy during iteration t is equal to $(v_x\,(t+1) - v_x\,(t))^2$. The sample's total energy is the sum over all wire transition energies.

The same procedure is used to estimate the energy of sample check-to-symbol messages. The samples are accumulated to compute the mean energy per symbol message and check message, $\mathcal{E}_\mu$ and $\mathcal{E}_\beta$, respectively. Finally the average energy per message is

$$\mathcal{E}_m = \frac{d_v}{d_v + d_c}\mathcal{E}_\mu + \frac{d_c}{d_v + d_c}\mathcal{E}_\beta.$$

IV. Results

Simulations were performed across 30 iterations using the modified density evolution procedure described in Sec. III. The channel noise parameter σ was varied following the procedure described by Richardson [7]. Regular $(3, 6)$ code ensembles were considered. The message representation is a W-bit word with sign-magnitude encoding and uniform quantization in the interval $[-L_{\max},\,+L_{\max}]$. The maximum LLR magnitude for each simulation was calculated as $(d_v + 1) * 4.0/N_0$, where $N_0 = 2\sigma^2$ is the power spectral density of channel noise.

Fig. 4 shows the message error rate (MER) as a function of both σ and V_{dd} for the case $W = 8$. Fig. 5 shows the MER plotted against the corresponding $\mathcal{E}_m$ for $W = 8$. Figs. 6 and 7 show the MER vs V_{dd} and $\mathcal{E}_m$, respectively, for the case $W = 10$. The average $\mathcal{E}_m$ was calculated from the mean transition activity averaged over 30 iterations. The MER was calculated as the average rate of messages with erroneous sign during the final six iterations (time steps 26 to 30). Fig. 5 shows that the MER is a function of V_{dd} and decreases below our measurement threshold (i.e. MER $< 10^{-6}$) when $V_{\text{dd}} > 10$, which we consider to be successful convergence.

Counting the number of messages that need to be exchanged during the course of a decoding cycle normalized per bit, we obtain a bound for the minimum energy per symbol to drive the interconnect network as

$$E_n > 2Id_v\mathcal{E}_m \tag{5}$$

Using the thresholds evaluated in Figs. 5 and 7 of approximately 40 kT, we obtain $E_n > 2400kT$ for $I = 10$ and $d_v = 3$.

V. Conclusions

We have used fundamental physical and information theoretic limits to bound the minimal energy required to operate a (large) message passing error control decoder. Based on average switching activity within a simulated decoder design, the limiting energy per bit is on the order of 30,000kT, which equals 0.1 femto Joules per bit. This bound may even be an overestimate because the decoder may be able to tolerate some noise-induced errors if the barrier parameter K is lowered. Additionally, the average switching activity tends to diminish across iterations when the decoder converges. Adiabatic circuit techniques could conceivably be employed to reduce the irreversible information loss. In that case, the

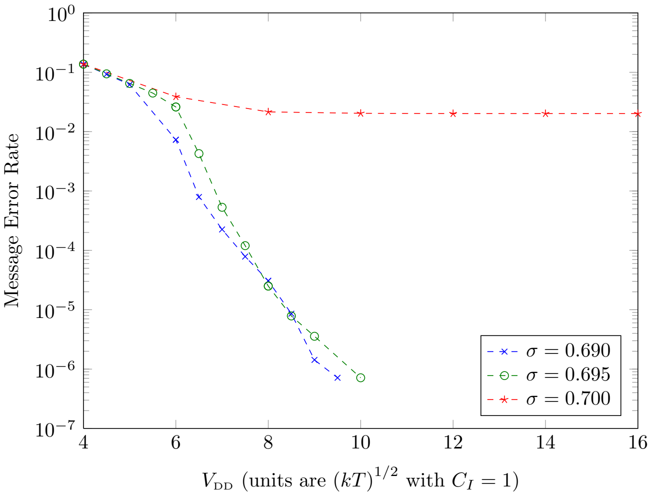


Fig. 4. MER for a regular $(3, 6)$ code ensemble with 8-bit linear-quantized message representation, varying the normalized supply voltage ($V_{\rm dd}$) and the channel noise standard deviation (σ). $\sigma = 0.7$ exceeds the ensemble's decoding threshold. Under ideal message passing, the threshold is $\sigma^* = 0.88$ [7].

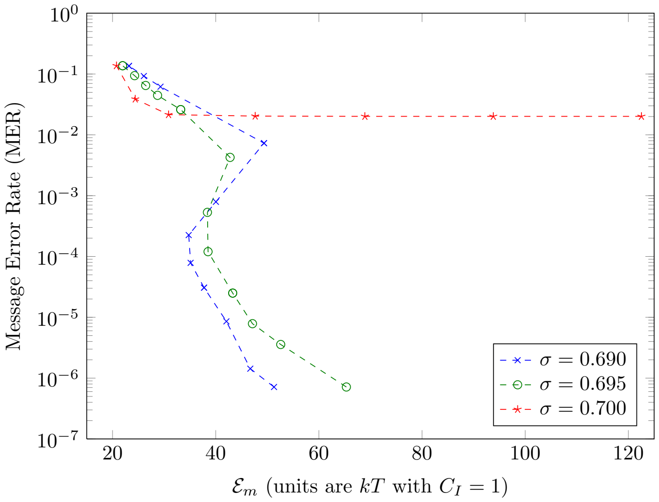


Fig. 5. MER for a regular $(3, 6)$ code ensemble with 8-bit linear-quantized message representation, plotted against the average energy per message.

energy in the communications network would be limiting and density evolution would provide an inescapable lower bound on the energy required for message passing, approximately $2400kT$ or 10 atto Joules per bit. The results in this analysis represent only a single class of codes and decoding algorithms. The methods presented here can conceivably be extended and applied to other codes and algorithms as a subject for future research.

Acknowledgment

C. Winstead's work was supported by the US National Science Foundation under award ECCS-0954747, and by the US Fulbright program. C. Schlegel's work was supported by the Canadian National Science and Engineering Research Council, under an Industrial Research Chair award.

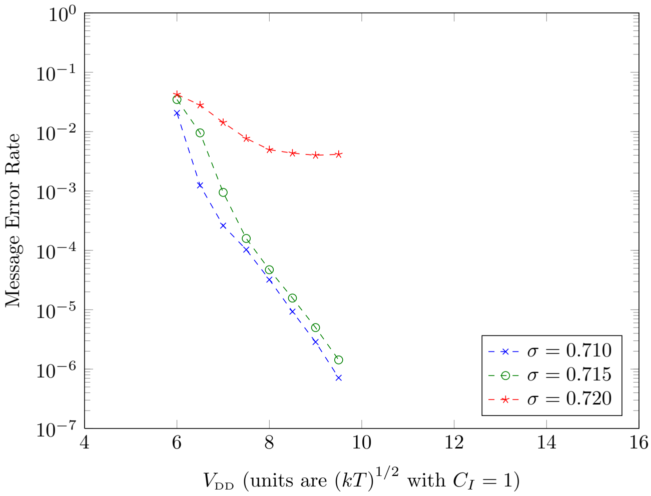


Fig. 6. MER for a regular $(3, 6)$ code ensemble with 10-bit linear-quantized message representation, while varying the normalized supply voltage ($V_{\rm dd}$) and the channel noise standard deviation (σ).

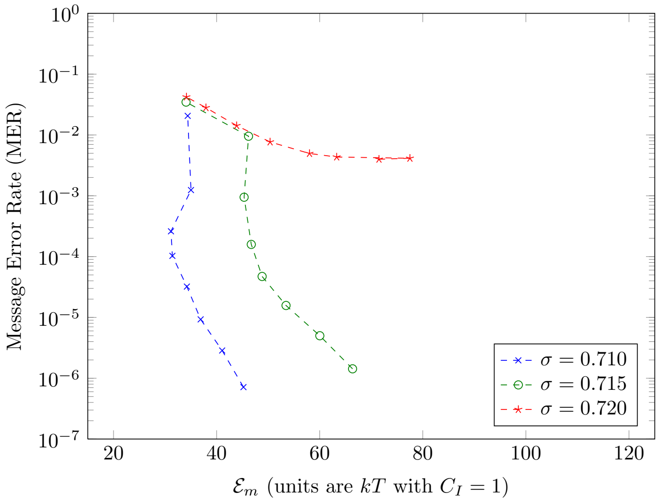


Fig. 7. MER for a regular $(3, 6)$ code ensemble with 10-bit linear-quantized message representation, plotted against the average energy per message.

References

[1] G. Cheng et al., "Sketched oxide single-electron transistor," *Nature Nanotechnology 6*, 343–347 (2011) doi:10.1038/nnano.2011.56.

[2] V. Zhirnov et al., "Limits to binary logic switch scaling – A gedanken model," *Proceedings of the IEEE*, Vol. 91, No. 11, November 2003.

[3] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Research and Development*, vol. 5, pp. 181–191, 1961.

[4] J. D. Meindl, J. A. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *Solid-State Circuits, IEEE Journal of* , vol. 35, no. 10, pp.1515,1516, October 2000.

[5] V. C. Gaudet, C. Schlegel, R. Dodd, "LDPC Decoder Message Formatting Based on Activity Factor Minimization Using Differential Density Evolution," *IEEE Inform. Theory Wrkshp (ITW)*, pp. 571,576, Sep. 2007.

[6] B. Crowley, V. C. Gaudet, "Switching Activity Minimization in Iterative LDPC Decoders," *Journal of Signal Processing Systems*, vol. 68, no. 1, pp. 63–73, July 2012.

[7] T. J. Richardson, R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp.599,618, February 2001.

[8] S. Zhang and C. Schlegel, "Controlling the Error Floor in LDPC Decoding," *IEEE Trans. Commun.*, Vol. 61, No. 9, pp. 3566–3575, 2013.

Low Complexity Decoding for Punctured Trellis-Coded Modulation Over Intersymbol Interference Channels

Fabian Schuh and Johannes B. Huber
Institute for Information Transmission, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
mail: {schuh, huber}@LNT.de

***Abstract*—Classical trellis-coded modulation (TCM) as introduced by Ungerboeck in 1976/1983 uses a signal constellation of twice the cardinality compared to an uncoded transmission with one bit of redundancy per PAM symbol, *i.e.*, application of codes with rates $\frac{n-1}{n}$ when 2^n denotes the cardinality of the signal constellation. The original approach therefore only comprises integer transmission rates, *i.e.*, $R = \{2, 3, 4 \ldots\}$, additionally, when transmitting over an intersymbol interference (ISI) channel an optimum decoding scheme would perform equalization and decoding of the channel code jointly.**

In this paper, we allow rate adjustment for TCM by means of puncturing the convolutional code (CC) on which a TCM scheme is based on. In this case a nontrivial mapping of the output symbols of the CC to signal points results in a time-variant trellis. We propose an efficient technique to integrate an ISI-channel into this trellis and show that the computational complexity can be significantly reduced by means of a reduced state sequence estimation (RSSE) algorithm for time-variant trellises.

***Index Terms*—trellis-coded modulation (TCM); punctured convolutional codes; Viterbi-Algorithm (VA); ISI-channel;**

I. Introduction

Ungerboeck's trellis-coded modulation (TCM) [1] is a bandwidth efficient digital transmission scheme when very low overall latency is desired. Low latency is ensured by the use of convolutional codes instead of block codes (*cf.* [2]) and the dispense with interleaving (as opposed to conventional bit-interleaved coded modulation [3]).

Ungerboeck showed that a significant increase in the Asymptotic Coding Gain (ACG) can be achieved when considering channel coding and modulation jointly. By expanding the constellation from 2^{n-1} to 2^n signal points and employing a rate-$\frac{n-1}{n}$ convolutional encoder one can improve the robustness of the transmission against noise by up to 6 dB without any further costs besides some computational effort. However, TCM is strictly limited to integer transmission rates.

Our approach applies *punctured* TCM (P-TCM) with an arbitrary rate. We extend P-TCM to intersymbol interference (ISI)-channel scenarios. In this case, ML-decoding can be performed by efficiently incorporating the ISI-channel into the trellis.

We show that reduced-state sequence estimation (RSSE) can be applied in order to reduce computational complexity. We were able to show in [4], [5] that for minimum phase channels, the number of states to decode must *not* be significantly higher than the number of states in the channel encoder. In this paper we will describe the application of RSSE to P-TCM and discuss the partitioning of the time-variant trellis into hyperstates.

This paper is structured as follows: In Sec. II we first introduce notation and present the system model. Sec. III briefly recapitulates a presentation technique that enables the implementation of punctured encoding. The application of RSSE for P-TCM is given in Sec. IV. Final results of numerical simulation and conclusions are given in Sec. V and Sec. VI, respectively.

II. System Model

This paper deals with convolutionally encoded pulse-amplitude modulated (PAM) transmission as depicted in Fig. 1. (Here, the term PAM is used for complex-valued signal constellations $\mathcal{A}$ as well including amplitude-shift keying (ASK), phase-shift keying (PSK) or quadrature-amplitude modulation (QAM).) A binary data sequence $\langle u \rangle$ is encoded using a rate-$\frac{n_\mathrm{c}-1}{n_\mathrm{c}}$ binary convolutional encoder $\mathcal{C}$ with generator polynomials $g_{ij}(D),\ 1 \leq i \leq n_\mathrm{c}; 1 \leq j \leq n_\mathrm{c}-1$, with delay operator D, $n_\mathrm{c}-1$ parallel binary-input symbols and n_c parallel output symbols at each time instant.

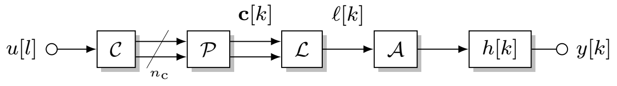


Fig. 1. Concatenation of a rate-$\frac{1}{2}$ convolutional encoder $\mathcal{C}$ and puncturing $\mathcal{P}$ with labeling and modulation ($n_\mathrm{u} = 0$, $n_\mathrm{c} = 2$).

At each output of the encoder, the symbols traverse through a puncturing system with puncturing matrix $\mathbf{P} = [P_{ij}]$, $P_{ij} \in \{0, 1\}$; $1 < i \leq n_\mathrm{c}$; $1 < j < \Omega$ and period Ω. For each (n_c)-tuple of encoder output symbols the puncturing scheme cyclically advances by one step. Where P_{ij} is zero, the current symbol at the output is discarded, accordingly.

The punctured encoded output symbols $\mathbf{c}[k]$ are labeled to $\ell[k]$ before being mapped to the $M = 2^{n_\mathrm{u}+n_\mathrm{c}} = 2^n$-ary signal constellation $\mathcal{A}$.

The modulated (possibly complex-valued) transmit signal traverses through a memory-L discrete-time ISI-channel with $L+1$ channel coefficients $h[k]$ with k denoting the discrete time index.

The task of the receiver is to estimate for the information bits given the transmit signal $y[k]$ plus additive noise. Here, we focus on perfect channel knowledge at the receiver-side.

III. Punctured Trellis-Coded Modulation

In the following, we will briefly recapitulate punctured convolutional trellis coded transmission over ISI-channel scenarios as introduced in [5].

In contrast to classical TCM, our approach using *punctured* convolutional codes results in nontrivial mapping of coded bits to modulation symbols. As a consequence, the trellis is time-variant as already described in [5]–[7].

In order to briefly recapitulate decoding concept for punctured trellis coded modulation, we focus on 4-ary ASK-modulation, a memory-2 convolutional code, and a short puncturing scheme namely $\mathbf{P} = \left[(1\,1)^\top \, (0\,1)^\top \right]$. Note that, whenever the number of erased bits in one period of the puncturing scheme is not dividable by $\log_2(M)$, the puncturing scheme has to be repeated until this condition is fulfilled. This restriction ensures that entire modulation symbols can be constructed by the finite state machine (FSM). In our example, the puncturing period (*e.g.*, $\left[(1\,1)^\top \, (0\,1)^\top \right]$) has to be applied twice. As can be seen from the encoding process in Fig. 2, the third and the seventh encoded symbol are punctured and do not contribute to the labeling and modulation process. Thus, the second symbol, *i.e.*, $a[k+1]$, contains information about $u[l+1]$ and $u[l+2]$ and the third symbol, *i.e.*, $a[k+2]$, contains information about $u[l+2]$ and $u[l+3]$.

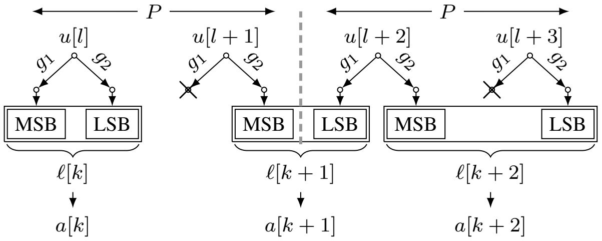


Fig. 2. Encoding process for a rate-$\frac{2}{3}$ punctured convolutional code and natural labeling. Overall transmission rate $R = \frac{4}{3}$.

When decoding the second symbol (*e.g.*, $a[k+1]$), a decision can be made for $u[l+1]$ but not for $u[l+2]$, as a portion of information will be received in the consecutive symbol. Thus, the trellis has to be expanded in order to use the symbols $a[k+1]$ and $a[k+2]$ when decoding $u[l+2]$. As sample trellis is given in Fig. 3.

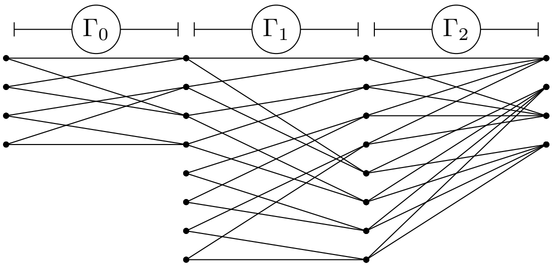


Fig. 3. Time-variant trellis for a punctured rate-2/3 convolutional code. In the first to VA steps, two transitions arrive at each state, *i.e.*, one bit can be estimate, whereas the third step allows an estimation for two bits.

To algorithmically handle the time-variant mapping we introduced a set of so-called generator offsets $\mathcal{T}_i$ which describe, depending on the puncturing scheme, modulation size, and time instant, the relations between generator polynomials, input value, FSM state, and mapping to MSB or LSB, respectively. For each new generator offset $\mathcal{T}_i$ a new trellis segment arises, *e.g.*, the number of generator offsets equals the number of trellis segments in one trellis period.

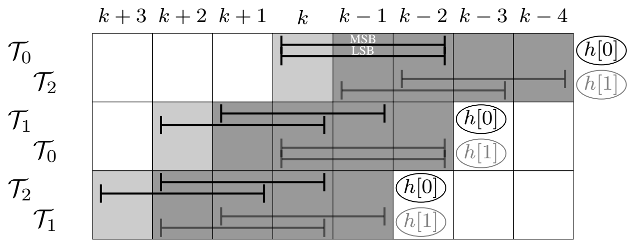


Fig. 4. State transitions of the transmitter FSM with $R = 4/3$ and the relations between generator polynomials, FSM-state/input and channel state for a memory-1 ISI-channel.

When transmitting over an ISI-channel, several symbols are stored in the memory of the ISI-channel independently from the encoding and puncturing process. Thus, multiple generator offsets $\mathcal{T}_i$ have to be considered simultaneously. This can be seen from Fig. 4 for a memory-1 channel. There, the resulting sequence of used generator offsets is depicted. This scheme can easily be extended to arbitrary lengths of the ISI-channel. A detailed description as well as an algorithm to construct such trellises will be given in a separate full-length paper.

IV. Reduced-State Sequence Estimation

Reduced-state sequence estimation (RSSE) [8] is proposed to reduce the number of states at the cost of small loss in Euclidean distance. In order to introduce RSSE for P-TCM, we first briefly recapitulate *delayed decision-feedback sequence estimation* (DFSE) [9].

A. Delayed Decision Feedback Sequence Estimation

When equalizing (uncoded) digital PAM signaling over a discrete-time ISI-channel with $L+1$ tabs using DFSE (*i.e.*, no decoding), the trellis is constructed from the first $\tilde{L} \le L$ tabs only. Thus, the number of states is reduced from M^L to $M^{\tilde{L}}$.

The remaining $L + 1 - \tilde{L}$ channel tabs are considered in a delayed decision-feedback equalization (DFE) that is performed in each trellis state using the *delayed* path register of the corresponding state.

The main difference to full state equalization appears in the metric computation for each time instant. From (1) it becomes clear that the state specific path register $p_{\mathrm{reg}}[k, \mathbf{s}]$ is delayed by $\tilde{L}$ and its elements are multiplied by the subsequent channel coefficients $h_{\mathrm{dfe}}[h]$ which have not been considered in the trellis. The branch metric $\lambda(\mathbf{s}, \mathbf{u})$ (*e.g.*, Euclidean distance of the received symbol $y[k]$ to the hypotheses $h(\mathbf{s}, \mathbf{u})$ for the states $\mathbf{s}$ and bits $\mathbf{u}$) thus includes the correction factor δ:

$$\delta = \sum_{\kappa} p_{\mathrm{reg}}[k - \tilde{L} + \kappa, \mathbf{s}] \cdot h_{\mathrm{dfse}}[\kappa] \qquad (1)$$

$$\lambda(\mathbf{s}, \mathbf{u}) = \left| y[k] - h(\mathbf{s}, \mathbf{u}) - \delta \right|^2$$

B. Reduced State Sequence Estimation

For our coded transmission over ISI-channel we consider RSSE instead. Here, Z arbitrary MLSE states, each with $M = 2^K$ possible branches to adjacent states, are combined into $Z_\mathrm{R} = \frac{Z}{2^J}$; $J \in \mathbb{N}$ *hyperstates* [10] each having 2^J substates and $2^K \cdot 2^J$ branches. A certain assignment of states to hyperstates is called a *partitioning* [10].

Instead of having 2^K arriving branches at each of the Z MLSE states we get a set of $2^K \cdot 2^J$ branches at each of the Z_R hyperstates. The total number of available branches remains $2^K \cdot Z$. However, when using RSSE only 2^K branches are possible (*i.e.*, *enabled*) from each state, at a given time instant. The availability of branches is determined by the path registers, and thus form a decision-feedback.

The metric computation in this case can be implemented as depicted in Algorithm 1. Note that with line 6 only M branches are activated. Thus, the VA has to decide between M survivor branches at each state giving an estimate for $\log_2(M)$ bits.

Algorithm 1 Metric calculations for RSSE

```
1: $\tilde{L} \leftarrow \log_2(\text{nr. hyperstates})/\log_2(M)$
2: $\ell \leftarrow \log_2(\text{nr. substates})/\log_2(M)$
3: for all $\mathbf{s} \in \mathcal{S}$ do
4:   for all $\mathbf{u} \in \mathcal{A} \cdot K$ do
5:     for all $\kappa = 1 \to \ell$ do
6:       $\zeta[\kappa] = p_\mathrm{reg}(\mathrm{end} - \tilde{L} + \kappa, \mathbf{s})$      ▷ active branches
7:     end for
8:     $\lambda(\mathbf{s},\mathbf{u}) \leftarrow \left|y[k] - h(\boldsymbol{\zeta},\mathbf{u})\right|^2$      ▷ branch metric
9:     $\lambda(\mathbf{s},\mathbf{u}) \leftarrow \lambda(\mathbf{s},\mathbf{u}) + \Gamma(\mathbf{s})$      ▷ acc. path metric
10:  end for
11: end for
```

For time-variant trellises some modifications to the underlying VA are necessary, which are described in [5].

The main difference to MLSE is, that we decide for a surviving path prematurely resulting in a truncation of error events. A loss in Euclidean distance appears if an error event with minimum Euclidean distance gets truncated. Therefore the performance of RSSE strongly depends on the partitioning of the states into hyperstates. Instead of exhaustively search for the optimum state partitioning, which maximizes the intra-hyperstate distance [10], we exploit the minimum phase characteristics of the ISI-channel which is, as described above, fully integrated into our trellis.

For a minimum phase channel impulse response the prior channel input symbols are weighted less than more recent ones and, thus, affect the metric less. The elder the symbols, the further back it is stored in the vector presentation of a particular trellis state. Hence, the intra-hyperstate distance is maximized when states are combined with respect to elder positions in the state number. This particular partitioning is equivalent to DFSE for ISI-channels (which is the optimum partitioning for equalization of minimum phase ISI-channels [10]) and will in the latter be called *DFSE partitioning*. As the minimum phase ISI-channel is the last element to affect the transmitted symbols and is also fully integrated into the trellis, we can apply the *DFSE partitioning* to use RSSE for P-TCM over ISI-channels. An implementation of this set partitioning for the J^th level exploiting the minimum phase characteristics is shown in Algorithm 2. The columns in the resulting matrix $p(z, i)$ define the states that have to be grouped into hyperstates.

Algorithm 2 DFSE Partitioning for RSSE

```
$i, z \leftarrow 1$
for $\ell = 1 \to Z$ do
  if $\ell \bmod J = 0$ then
    $z \leftarrow z + 1$
  end if
  $p(z, i) = \ell$
  $i \leftarrow i + 1$
end for
```

In the following we will focus on our state design and two possibilities to apply *DFSE partitioning* to time-variant trellises.

C. State Design

In Fig. 5 a single trellis state in the VA is depicted as a FIFO. Input values to the FSM are represented by the branches at the left-hand side, whereas values that drop of the FIFO are stored within the state-specific path register $p_\mathrm{reg}(k, \mathbf{s})$ at the right-hand side. During decoding when entering a trellis segment that is split, *i.e.*, has an increased number of states, the two possibilities for the punctured bit are tracked using an additional delay element, indicated by the hatched block (▨).

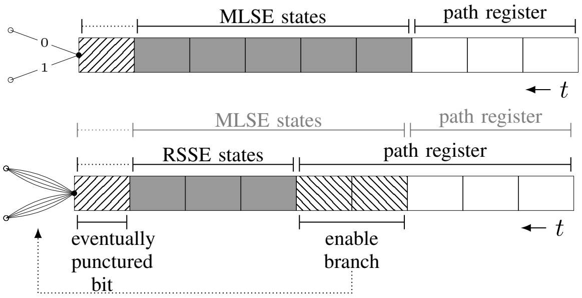


Fig. 5. Graphical representation of our state design for punctured convolutional coding with and without reduced-state sequence estimation.

When DFSE partitioning is performed, the states can be reduced as shown in Fig. 5. There, fewer FIFO elements are used to define the trellis state, while the remaining ones are used as feedback to enable branches for the next trellis step for that particular state.

An implementation of this algorithm needs to ensure that when entering a split trellis segment, *i.e.*, increased number of states, the right path register is chosen as source for the decision feedback.

D. State Partitioning

As already mentioned, a DFSE partitioning of the first order, *i.e.*, reducing the number of states in each trellis segment by a factor of two, combining states that differ in the eldest position, into hyperstates. Hence, when applied to a punctured TCM, each trellis segment undergoes the same partitioning. A resulting reduced-state trellis is depicted on the left-hand side of Fig. 6 for $J = 1$ and $J = 2$. As a consequence, non-existing states from the original trellis are also partitioned (*cf.*, first trellis segment).

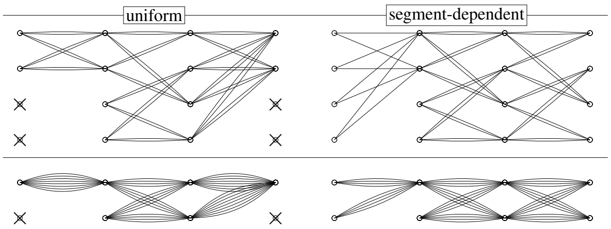


Fig. 6. Illustration of uniform and segment-dependent partitioning for $J = \{1, 2\}$.

Due to the reduction of non-existing states, and hence reduced minimum Euclidean distance, we propose to partition those segments first that are split and keeping the others unpartitioned. As can be seen from the first-level state reduction in Fig. 6 (right-hand side, above), the first segment contains two transitions in each state, and thus is not partitioned, whereas in segment two and three, each state has four transitions, due to the state partitioning.

Apparently, the first segment may be handled with a full state VA, while the other two segments need to be decoded via RSSE using the path register in each state.

The right-hand side of Fig. 6 show the segment-depending partitioning and state reduction for $J = 1$ and $J = 2$. The segments for $J = 1$ show four, and eight transitions per state, respectively. At this point RSSE has to consider a different amount of information in the path register for each segment.

The resulting trellis shows a less decreased minimum Euclidean distance when compared to the uniform partitioning but has a slightly higher computational complexity because of the extra states. Thus, the segment-dependent partitioning technique enables an even more flexible way to trade between complexity and performance.

V. Numerical Results

In this section we will give numerical simulation results and investigate several ISI-channels for *punctured* TCM. We analyse the decoder complexity as *number of branch metric calculations* per *information bit* and show that this approach enables a flexible trade-off between computational complexity and performance.

Due to the minimum phase characteristics of the ISI-channel the partitioning of the trellis states into hyperstates leads to the smallest possible loss in Euclidean distance. Hence, if, for instance, the ISI-channel is an equal tab delay line, the loss in Euclidean distance is higher than for an exponentially decaying channel because of the premature decisions for a surviving path.

To see this effect we conducted simulations over different ISI-channels of unit energy and plotted the complexity number over the required $\frac{E_\mathrm{b}}{N_0}$ to achieve a bit error probability of less than 10^{-3}. The unit energy channels are defined as:

$$h_\mathrm{exp}[\kappa] = \frac{1}{\sqrt{\sum_\kappa |h_\mathrm{exp}[\kappa]|^2}} \, \mathrm{e}^{(-\kappa/\kappa_0)} \qquad \text{for } 0 \le \kappa \le L$$

$$h_\mathrm{lin}[\kappa] = \frac{1}{\sqrt{\sum_\kappa |h_\mathrm{lin}[\kappa]|^2}} \, \frac{L-\kappa+1}{L+1} \qquad \text{for } 0 \le \kappa \le L$$

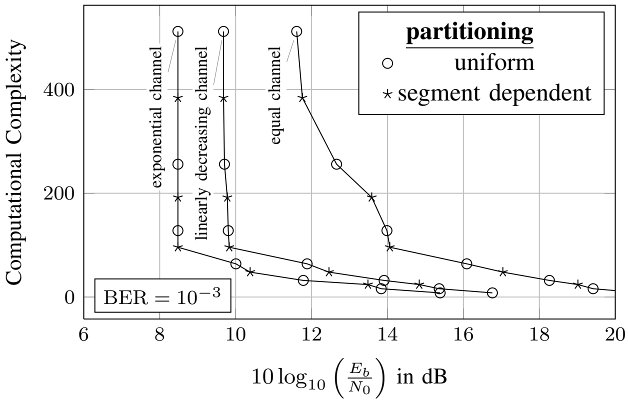


Fig. 7. Decoding complexity for a P-TCM transmission scheme with generator polynomials $[13\,15]_\mathrm{oct}$, puncturing scheme $\left[\,(1\,1)^\top\,(0\,1)^\top\,\right]$ (Rate: $\frac{4}{3}$) natural labeling and 4-ASK signaling over three different ISI-channels $h_\mathrm{exp}[\kappa]$ ($\kappa_0 = 1$), $h_\mathrm{lin}[\kappa]$, $h_\mathrm{equ}[\kappa]$ (solid: Uniform partitioning, dashed: Segment-dependent partitioning).

$$h_\mathrm{equ}[\kappa] = \frac{1}{\sqrt{\sum_\kappa |h_\mathrm{equ}[\kappa]|^2}} = \frac{1}{\sqrt{L}} \qquad \text{for } 0 \le \kappa \le L$$

The results can be seen in Fig. 7. As should be clear to the reader, the loss in Euclidean distance is smallest for an exponentially decaying ISI-channel.

VI. Conclusion

It has been shown that TCM can be extended by puncturing. Furthermore, an efficient MLSE decoder for P-TCM over ISI-channels was proposed and investigated.

The numerical simulation results clearly show that we can achieve a soft trade-off between spectral and power efficiency easier and more flexibly than by means of traditional TCM.

References

[1] G. Ungerboeck, “Trellis-coded modulation with redundant signal sets; part i: Introduction; part ii: State of the art,” *Communications Magazine, IEEE*, vol. 25, no. 2, pp. 5 –21, February 1987.

[2] T. Hehn and J. B. Huber, “LDPC Codes and Convolutional Codes with Equal Structural Delay: A Comparison,” *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1683–1692, June 2009.

[3] E. Zehavi, “8-PSK Trellis Codes for a Rayleigh Channel,” *IEEE Transactions on Communications*, vol. 40, no. 5, pp. 873 –884, may 1992.

[4] F. Schuh, A. Schenk, and J. Huber, “Reduced complexity Super-Trellis decoding for convolutionally encoded transmission over ISI-Channels,” in *2013 International Conference on Computing, Networking and Communications, Signal Processing for Communications Symposium (ICNC’13 - SPC)*, San Diego, USA, Jan. 2013.

[5] ——, “Matched decoding for punctured convolutional encoded transmission over ISI-Channels,” in *9th International ITG Conference on Systems, Communications and Coding 2013 (SCC’2013)*, Munich, Germany, Jan. 2013.

[6] T. Woerz and R. Schweikert, “Performance of punctured pragmatic codes,” in *Global Telecommunications Conference, 1995. GLOBECOM ’95., IEEE*, vol. 1, nov 1995, pp. 664 –669 vol.1.

[7] F. Schuh, A. Schenk, and J. Huber, “Punctured Trellis-Coded Modulation,” in *submitted to ICC 2014*, Jun. 2014.

[8] M. Eyuboglu and S. Qureshi, “Reduced-State Sequence Estimation With Set Partitioning and Decision Feedback,” *IEEE Trans. Commun.*, vol. 36, no. 1, pp. 13–20, Jan. 1988.

[9] A. Duel-Hallen and C. Heegard, “Delayed Decision-Feedback Sequence Estimation,” *IEEE Trans. Commun.*, vol. 37, no. 5, pp. 428–436, May 1989.

[10] B. Spinnler and J. Huber, “Design of Hyper States for Reduced-State Sequence Estimation,” in *Proc. IEEE Int. Conf. Communications ICC ’95 Seattle*, vol. 1, 1995, pp. 1–6.

An Efficient Length- and Rate-Preserving Concatenation of Polar and Repetition Codes

Mathis Seidl and Johannes B. Huber
Lehrstuhl für Informationsübertragung
Universität Erlangen-Nürnberg
Erlangen, Germany
Email: {seidl, huber}@LNT.de

***Abstract*—We improve the method in [1] for increasing the finite-lengh performance of polar codes by protecting specific, less reliable symbols with simple outer repetition codes. Decoding of the scheme integrates easily in the known successive decoding algorithms for polar codes. Overall rate and block length remain unchanged, the decoding complexity is at most doubled. A comparison to related methods for performance improvement of polar codes is drawn.**

I. Introduction

Polar coding is known as a channel coding construction that is able to achieve the capacity of many symmetric discrete memoryless channels under low-complexity $\mathcal{O}(N \log N)$ encoding and successive decoding [2]. Unfortunately, the error performance of polar codes for finite block lengths is quite moderate. The key feature of polar coding – when compared to other existing channel block coding schemes – clearly lies in its low decoding complexity. Therefore, the trade-off between computational complexity and error performance for polar codes is of interest, i.e., the development of efficient methods that allow for better performance at moderate additional complexity.

Optimizing the decoding algorithm for polar codes has been the subject of various work, e.g. [3], [4] and has led to substantial improvements. Though, apart from the decoder, the code itself leaves room for improvement as well.

To this end, we propose a modified polar code construction by means of a serial concatenated scheme with the polar code used as an inner code. In contrast to many existing concatenation schemes based on polar codes as inner codes (as considered, e.g., in [5], [6], [7]), we focus here on coding schemes that do not change overall rate and block length, thus facilitating a pure trade-off of complexity and error performance. The approach is based on our prior attempt [1] where block codes of small dimension were chosen as outer codes. In this paper we show that – by an efficient and systematic design – an equivalent, significant performance gain is achieved by protecting an inner polar code with only one-dimensional outer codes, i.e., repetition codes, resulting in an quite smaller increase of complexity. Furthermore, we relate the results to methods where the decoder is modified instead of the code.

The paper is organized as follows: After a brief review on polar codes and their decoding strategies in Sec. II, we describe our concatenated code construction in Sec. III, followed by simulation results in Sec. IV and some conclusive remarks in Sec. V.

II. Polar Codes and their Decoding

A. Code Construction

Since the concept of polar coding is widely known, we only give a brief overview, focussing on the aspects of importance for this paper. We follow the original approach in [2] where the generator matrix is chosen as a subset (indexed by $\mathcal{A}$) of the rows of the binary matrix

$$\boldsymbol{G}_N = \boldsymbol{F}^{\otimes n} \quad , \quad \boldsymbol{F} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{1}$$

with $n = \log_2 N$ and $\otimes n$ denoting the n-th Kronecker power.

Under successive decoding, the transmission of the particular source symbols u_i may be described by their own binary-input channels (*bit channels*) which show a polarization effect in the sense that their capacities are almost all either near 0 or near 1. These capacities – or equivalently, the corresponding failure probabilities $p_\mathrm{e}(i)$) – can be easily determined. The channels with high capacities are chosen to form the set $\mathcal{A}$ whereas the residual channels (*frozen channels*) transmit fixed values that are known to the decoder.

B. Successive Decoding

In the successive cancellation (SC) decoding approach [2], estimates $\hat{u}_i$ on the source symbols u_i ($i \in \mathcal{A}$) are calculated successively, according to the recursion formula

$$\hat{u}_i := \underset{b \in \{0,1\}}{\operatorname{argmax}} \left\{ \Pr\bigl(U_i = b | \boldsymbol{Y}, \hat{U}_0 \cdots \hat{U}_{i-1}\bigr) \right\} . \tag{2}$$

Thus, in each step i the decoder checks which of the possible two values for u_i is more likely, given the received vector $\boldsymbol{y}$ as well as the sequence $\hat{u}_0 \cdots \hat{u}_{i-1}$ of data symbols already decided in the previous steps. Due to the special structure of $\boldsymbol{G}_N$, the calculation of the probabilities in (2) can be implemented in an FFT-like fashion, resulting in a low $\mathcal{O}(N \log N)$ overall decoding complexity. With increasing SNR, the performance of the SC algorithm is known to converge to that of an optimum Maximum-Likelihood (ML) decoder. The decoding process as a path search is illustrated in Fig. 1a).

The word error performance under SC decoding can be precisely determined. It is given by the term

$$\mathrm{WER}_\mathrm{SC} = 1 - \prod_{i \in \mathcal{A}} (1 - p_\mathrm{e}(i)) \tag{3}$$

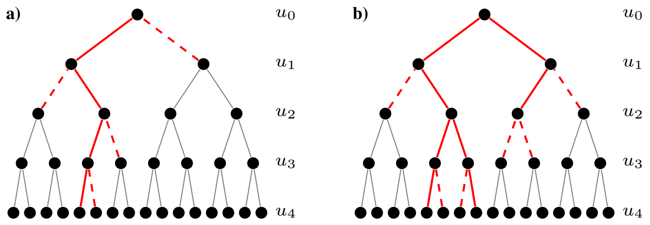


Fig. 1. a) SC decoding. b) Successive list decoding (list size $L = 2$). Bold face lines: Inspected (and tentatively selected) paths in the SC decoding process. Dashed lines: Inspected but discarded paths. Thin lines: Never inspected paths.

where $p_\mathrm{e}(i)$ denotes the probability of a wrong decision at stage i of the decoder provided that all previous decisions have been correct. From (3) it is clear that for an optimal code construction, $\mathcal{A}$ should consist of the bit channels with lowest failure probabilities $p_\mathrm{e}(i)$.

C. Successive List Decoding

As an improved version of the SC decoder for increased performance in the low-SNR regime, list decoding for polar codes has been proposed [3]. The successive list decoder does not take hard decisions on the u_i immediately. Instead, both possible values are examined in separate decoding branches, and the corresponding likelihood values are determined. If the number of branches exceeds a certain design parameter L (the *list size*), the least probable branches are discarded, as examplary visualized in Fig. 1b) for $L = 2$. The complexity of decoding scales linearly with the list size L and is of order $\mathcal{O}(LN \log N)$. Note that only the decoder is modified here while the code does not change.

In an extended version of the above-mentioned paper [8], the authors propose a serial concatenation scheme with an inner polar code and a very high-rate outer CRC (cyclic redundancy check) code. Decoding for this scheme is accomplished in two steps: First, the successive list decoder generates a list of L possible codewords. After that, the CRC sums for each entry of the list are calculated in order to check for the correct codeword. By this means, correct decoding is possible in principle even when another polar codeword in the list belongs to a more likely path, enabling successful decoding beyond the performance of an ML decoder for the inner polar code alone – as long as the correct codeword is part of the output list. It has been shown [8] that by this means, a significant performance gain is achieved.

III. Concatenated Code Construction

Here, we follow a different approach based on the varying bit channel capacities under successive decoding. The proposed coding scheme follows the conventional serial concatenation principle where the source symbols are first encoded by an outer code, followed by an inner encoding. Thus, the overall rate is given as $R = R_\mathrm{o} R_\mathrm{i}$ with R_o and R_i being the rate of outer and inner code, respectively. In our approach, outer and inner code are decoded jointly by a single algorithm.

The generator matrix $\boldsymbol{G}$ of a (N, K) polar code constructed in the conventional way may be represented as

$$\boldsymbol{G} = \boldsymbol{P}_\mathcal{A} \cdot \boldsymbol{G}_N \tag{4}$$

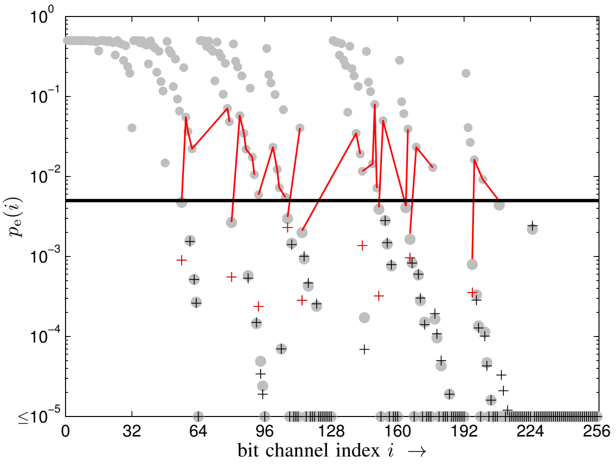


Fig. 2. Failure probabilities $p_\mathrm{e}(i)$ for a polar code ($R = 1/2$, $N = 256$, BI-AWGN channel at $E_\mathrm{s}/N_0 = -0.5$ dB). Gray circles: original code. Markers: concatenated code. Red lines: Repetition blocks of outer code.

where $\boldsymbol{P}_\mathcal{A}$ is a $(K \times N)$ projection matrix with rows built from the i-th unit vectors of length N ($i \in \mathcal{A}$, $|\mathcal{A}| = K$). We now aim to construct an optimized generator matrix of equal dimensions by a serial concatenation of the form

$$\boldsymbol{G}^* = \boldsymbol{G}_\mathrm{o} \cdot (\boldsymbol{P}_{\mathcal{A}^*} \cdot \boldsymbol{G}_N) \tag{5}$$

based on an enlarged set of channel indices $\mathcal{A}^*$ with $K < |\mathcal{A}^*| \leq N$. The $(K \times |\mathcal{A}^*|)$ matrix $\boldsymbol{G}_\mathrm{o}$ serves as a generator matrix of a suitably chosen outer code [1].

In the following, we demonstrate the code construction by means of an example considering a rate-$1/2$, length-256 polar code.

A. Inner Code Design

The gray circles in Fig. 2 show the failure probabilities $p_\mathrm{e}(i)$ of the bit channels after transmission over a binary-input AWGN channel at $E_\mathrm{s}/N_0 = -0.5$ dB. The black line corresponds to the design rate $R = 1/2$ of the original code. The indices i with $p_\mathrm{e}(i)$ below this threshold form the set $\mathcal{A}$.

For construction of a concatenated code with equal rate and block length from a given (N, K) polar code, we choose the inner code as a polar code with same length N but with a higher rate $R_\mathrm{i} > R$. This is easily accomplished by enlarging the set $\mathcal{A}$ of information symbols, i.e., using additional (previously frozen) bit channels for transmission to form the set $\mathcal{A}^*$.

B. Outer Code Design

As can be derived from (3) and Fig. 2, the word error rate is dominated by a comparatively small fraction of bit channels close to the threshold. We now aim to protect these least-reliable bit channels by a suited outer code including some

[1] While the code is designed to operate *inside* the successive decoding process, from the encoding procedure (5) it becomes clear that it serves in fact as an outer code. In our prior paper [1], the denotation ”inner code“ had been used from the decoding perspective which indeed is misleading.

additional (formerly frozen) channels that are provided by the inner enlarged polar code.

In contrast to our previous approach [1], in this paper we aim to minimize the additional complexity introduced by outer decoding which imposes a number of constraints on the code construction that are explained in the following:

First, we focus on simple one-dimensional codes, i.e., repetition codes. Thus, the outer coding actually consists in setting some of the source symbols to the same value and building small sets of combined channel indices from $\mathcal{A}^*$ (*repetition blocks*). For further complexity reduction, we require that these blocks do not overlap (in the sense that the contained indices do not overlap).

The use of more than one bit channel for transmitting a single bit of information may be represented by a single equivalent bit channel (red markers in Fig. 2). We found that protecting (merging) two channels in this way always leads to an improvement w.r.t. the first-positioned channel, but not necessarily when compared to the second one. This is due to the successive decoding strategy: The decision on a repetition block is made at reaching the end of the block, like explained in detail in the next subsection. If already at the first index of a block a wrong codeword corresponds to the more likely path, decoding of the following symbols (that are in general not protected by the outer code) is quite likely to fail, even when both possible values for the first symbol are pursued. Therefore, a high misdecoding probability at the first index of a repetition block has a more fatal influence than an unreliable decision at the end. Consequently, a repetition block should always start with the most reliable bit channel. Blocks of larger length are built in an analog fashion. Here, also the most reliable bit channel should be put in front.

Finally, the rate of the original code has to be preserved, which leads to further obvious restrictions on the number of possible repetition blocks. Finding the optimum from the remaining possible outer coding schemes is easily accomplished by an exhaustive search.

Fig. 2 shows an example of an outer coding scheme constructed according to the above-mentioned constraints. The markers represent the bit channels used by the concatenated code. Here, the repetition blocks are visualized by red lines, the red markers denote the corresponding equivalent bit channels while the black markers stand for the unmodified bit channels of the concatenated code. We remark that further increasing the rate R_i of the inner code has no significant effect on the performance.

Clearly, the proposed scheme can easily be extended to using higher-dimensional outer codes for increased performance, as considered in [1], though at the cost of an increased complexity. Moreover, the results from [1] indicate that the possible additional gain will not be large.

C. Decoding

For joint decoding of inner and outer code, we apply the original SC algorithm with a slight modification, only. Decoding of a received vector starts as usual. Assume now that two source bits u_i and u_j are protected by a repetition code, i.e., $u_i = u_j$ for some $j > i$. On reaching stage i, instead of taking a hard decision on u_i, the decoder creates a new branch and tests both possibilities by determining the sequences

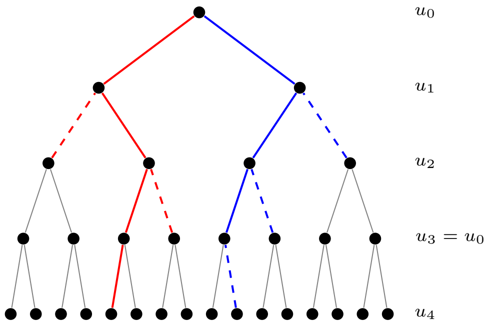


Fig. 3. SC decoding of an outer repetition code operating on u_0 and u_3. Bold face lines: Inspected (and tentatively selected) paths in the SC decoding process. Dashed lines: Inspected but discarded paths. Thin lines: Never inspected paths.

$$\boldsymbol{s}_0 = \langle 0, \hat{u}_{0,i+1} \cdots \hat{u}_{0,j-1}, 0 \rangle \ ,$$
$$\boldsymbol{s}_1 = \langle 1, \hat{u}_{1,i+1} \cdots \hat{u}_{1,j-1}, 1 \rangle \ .$$

For decisions on the symbols $\hat{u}_{0,i+1}, \ldots, \hat{u}_{0,j-1}$ and $\hat{u}_{1,i+1}, \ldots, \hat{u}_{1,j-1}$, the conventional SC decision rule (2) is applied. Afterwards, the more likely of the two sequences is selected:

$$\langle \hat{u}_i \cdots \hat{u}_j \rangle := \boldsymbol{s}_{b^*} \tag{6}$$

where

$$b^* = \operatorname*{argmax}_{b \in \{0,1\}} \left\{ \Pr\bigl(\boldsymbol{S}_b = \boldsymbol{s}_b | \boldsymbol{Y}, \hat{U}_0 \cdots \hat{U}_{i-1}, \boldsymbol{S}_b\bigr) \right\} . \tag{7}$$

The other path is discarded. The decoding scheme is visualized in Fig. 3 for a simple example code with $u_0 = u_3$. Repetition codes of larger length are decoded in an analog fashion. Clearly, the decoding complexity is at most doubled since we exclude overlapping blocks. Furthermore, the decoding of outer repetition codes in this way can easily be integrated into improved versions of the SC decoder, e.g., list decoding.

IV. Simulation Results

Fig. 4 shows simulation results for polar coding schemes of block length $N = 256$ and $N = 1024$ transmitted over a BPSK-AWGN channel. The shorter and longer code have been optimized (according to (3)) for $E_\mathrm{b}/N_0 = 2.5$ dB and $E_\mathrm{b}/N_0 = 2.0$ dB, respectively.

Compared to the original, SC-decoded polar codes (blue), the use of an improved decoder like the successive list decoder (green) shows an SNR-dependent effect: In the low-SNR region, significant gains are achieved while for increasing SNR the performance advantage vanishes. Here, both decoders perform close to ML decoding.

The proposed scheme (red) leads to an improved performance in a similarly efficient way with a complexity comparable to that of a list decoder with $L = 2$. However, in contrast to an optimized decoder, it achieves a constant coding gain of approx. 0.3 dB (0.2 dB for the longer code) over the SC-decoded code at all SNR regimes. Therefore, at high SNR it

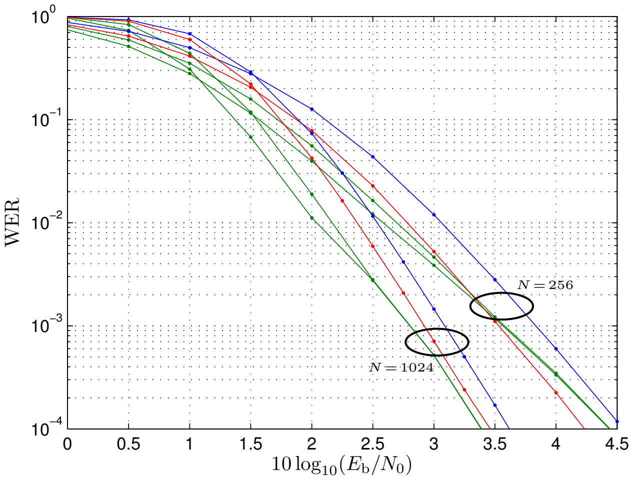


Fig. 4. Simulation results: BPSK-AWGN channel, polar code block length $N = 256, 1024$, rate $R = 1/2$. Blue: SC decoding. Green: successive list decoding (list size $L = 2, 4$). Red: proposed concatenation scheme.

is able to outperform even a list decoder with large list size or an ML decoder, but with much lower complexity, because the rate- and length-preserving concatenation yields an improved code.

V. Conclusion

The proposed concatenation scheme may be seen as a method to overcome the quantization effect when constructing a polar code that is caused by a hard selection of the bit channels (each channel is either used for information transmission or frozen).

As this quantization vanishes with increasing block length and polarization, the scheme is certainly restricted to polar codes of short to moderate length. Although the achievable performance gain is not too large, it comes at very small additional costs. When used together with an improved polar decoder, the beneficial effects of both approaches are combined. Furthermore, the proposed scheme can itself be used as an inner code in other concatenation approaches – at least if inner and outer decoding are performed separately there like in [6]. In this case, the coding gain in error performance is preserved.

References

[1] M. Seidl and J. Huber, “Improving successive cancellation decoding of polar codes by usage of inner block codes,” in *Proc. Int. Symp. Turbo Codes & Related Topics*, Sep. 2010, pp. 103–106.

[2] E. Arıkan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. Inf. Theory*, vol. 55, pp. 3051–3073, Jul. 2009.

[3] I. Tal and A. Vardy, “List decoding of polar codes,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 1–5.

[4] A. Alamdar-Yazdi and F. Kschischang, “A simplified successive-cancellation decoder for polar codes,” *IEEE Commun. Lett.*, vol. 15, no. 12, pp. 1378–1380, Dec. 2011.

[5] M. Bakshi, S. Jaggi, and M. Effros, “Concatenated polar codes,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2010, pp. 918–922.

[6] H. Kurzweil, M. Seidl, and J. B. Huber, “Reduced-complexity collaborative decoding of interleaved Reed-Solomon and Gabidulin codes,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 2563–2567.

[7] H. Mahdavifar, M. El-Khamy, J. Lee, and I. Kang, “On the construction and decoding of concatenated polar codes,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 952–956.

[8] I. Tal and A. Vardy, “List decoding of polar codes,” *CoRR*, vol. abs/1206.0050, 2012.

Sierpinski Prefactors in the Guruswami–Sudan Interpolation Step

Christian Senger
ECE, University of Toronto
Toronto, Ontario, Canada
Email: csenger@comm.utoronto.ca

Abstract—**Sierpinski prefactors are introduced, a concept that exploits the fact that many binomial coefficients in the Hasse derivative that appears in the Guruswami–Sudan interpolation step are zero modulo the base field characteristic. A reduced Guruswami–Sudan interpolation step for generalized Reed–Solomon codes with significantly fewer unknowns than the original interpolation step is formulated.**

I. Introduction

The re-encoding projection [1], [2] is a well-known method to reduce the complexity of the Guruswami–Sudan interpolation step for decoding generalized Reed–Solomon codes. It is based on projecting received vectors to a subspace with beneficial properties, which allows useful predictions about the solution space of the interpolation step to be made. The computational overhead of the projection is negligible, as it basically consists of one additional erasures-only decoding step. We show in this paper that similar predictions about the solution space can be made by exploiting an inherent property of finite fields, namely, that all multiples of the field characteristic are zero.

II. Preliminaries

Definition 1. For a prime power q and $n,k\in\mathbb{N}\setminus\{0\}$ with $k\leq n\leq q$ let $\mathcal{A}=\{\alpha_0,\dots,\alpha_{n-1}\}$ be an ordered set of distinct elements from the finite field $\mathbb{F}_q$ with q elements and let $\mathcal{B}=\{\beta_0,\dots,\beta_{n-1}\}$ be an ordered set of nonzero (not necessarily distinct) elements from $\mathbb{F}_q$. Then the set of vectors

$$\mathcal{GRS}_{\mathcal{A},\mathcal{B}}(\mathbb{F}_q;n,k)\triangleq\{(\beta_0u(\alpha_0),\dots,\beta_{n-1}u(\alpha_{n-1})):\\ u(x)\in\mathbb{F}_q[x],\deg[u(x)]<k\}$$

constitutes a *generalized Reed–Solomon (GRS) code* over $\mathbb{F}_q$.

When possible, we write $\mathcal{GRS}$ for $\mathcal{GRS}_{\mathcal{A},\mathcal{B}}(\mathbb{F}_q;n,k)$. *Conventional Reed–Solomon (RS)* codes are special cases of GRS codes with $\mathcal{A}=\{1,\alpha,\dots,\alpha^{n-1}\}$ and $\mathcal{B}=\{1,\alpha^b,\dots,\alpha^{b(n-1)}\}$, where $\alpha\in\mathbb{F}_q$ is primitive and $b\in\mathbb{N}$.

The *Guruswami–Sudan algorithm (GSA)* [3] for decoding GRS codes can be divided into two steps: the interpolation step (Problem 1) and the factorization step (which is not the focus of this paper). Let $\boldsymbol{c}\in\mathcal{GRS}$ be a codeword, $\boldsymbol{e}\in\mathbb{F}_q^n$ be an error vector of Hamming weight $\mathrm{wt}_\mathrm{H}[\boldsymbol{e}]=\varepsilon$, and $\boldsymbol{y}=\boldsymbol{c}+\boldsymbol{e}$ be the corresponding received vector obtained from the transmission channel. Furthermore, let $\mathcal{I}\triangleq\{0,\dots,n-1\}$ and let $r,\ell\in\mathbb{N}\setminus\{0\}$ be two parameters of the GSA with $r\leq\ell$. We associate the polynomial $P_\mathcal{I}(x)\triangleq\prod_{i\in\mathcal{I}}(x-\alpha_i)$ with $\mathcal{I}$.

Problem 1 (GSA Interpolation Step). Given a received vector $\boldsymbol{y}$ and $\varepsilon_0\in\mathbb{N}$, find a nonzero bivariate polynomial $Q(x,z)=Q_0(x)+Q_1(x)z+\dots+Q_\ell(x)z^\ell\in\mathbb{F}_q[x,z]$ such that

$$\deg[Q_\nu(x)]\leq r(n-\varepsilon_0)-\nu(k-1)-1\triangleq d_{Q_\nu}$$

for $\nu=0,\dots,\ell$ and
$\forall i\in\mathcal{I}\;\forall s,t\in\mathbb{N}:s+t<r$ and

$$\sum_{\nu=t}^{\ell}\binom{\nu}{t}z^{\nu-t}\sum_{\mu=s}^{d_{Q_\nu}}\binom{\mu}{s}x^{\mu-s}Q_{\nu,\mu}\Big|_{(x,z)=(\alpha_i,y_i)}=0,\quad(1)$$

where $Q_\nu(x)=\sum_{\mu=0}^{d_{Q_\nu}}Q_{\nu,\mu}x^\mu$.

The nested sum in (1) is called the (s,t)th *mixed partial Hasse derivative* of $Q(x,z)$. The condition that all (s,t)th Hasse derivatives with $s+t<r$ evaluate to zero for all tuples (α_i,y_i), $i\in\mathcal{I}$, means that these tuples are zeros of multiplicity r of $Q(x,z)$. For that reason, we refer to the parameter r as the *multiplicity* of the GSA. The parameter ℓ is called the *list size*. The linear system of equations associated with Problem 1 has a nonzero solution $Q(x,z)$ as long as

$$\varepsilon<\frac{n(2\ell-r+1)}{2(\ell+1)}-\frac{\ell(k-1)}{2r}\triangleq\varepsilon_0.$$

Naively solving the system with Gaussian elimination in order to obtain a solution is in $\mathcal{O}\left[\ell^6n^3\right]$, however, accelerated algorithms can be found in the literature, e.g., *Kötter interpolation* in [4]. Without loss of generality we assume in the following that the columns of the coefficient matrix (from left to right) are associated with the unknown coefficients $Q_{0,0},\dots,Q_{0,d_{Q_0}},Q_{1,0},\dots,Q_{1,d_{Q_1}},\dots,Q_{\ell,0},\dots,Q_{\ell,d_{Q_\ell}}$.

III. Sierpinski Prefactors

In this section, we introduce a new technique that results in structured solutions of Problem 1. In contrast to the well-known re-encoding projection [1], [2], this approach does *not* require any additional computations, it simply exploits basic properties of the GRS code's base field $\mathbb{F}_q$. The main idea is to exploit the fact that many of the binomial coefficients in (1) are zero modulo the characteristic of $\mathbb{F}_q$.

To see this, let us consider the left-aligned Pascal triangle in Fig. 1, where the entries are calculated modulo 3. Obviously, the zero entries of the triangle follow regular patterns. The triangle resembles variants of the left-aligned *Sierpinski gasket*, one of the most basic examples of a self-similar set. In the following, we will refer to a Pascal triangle with entries

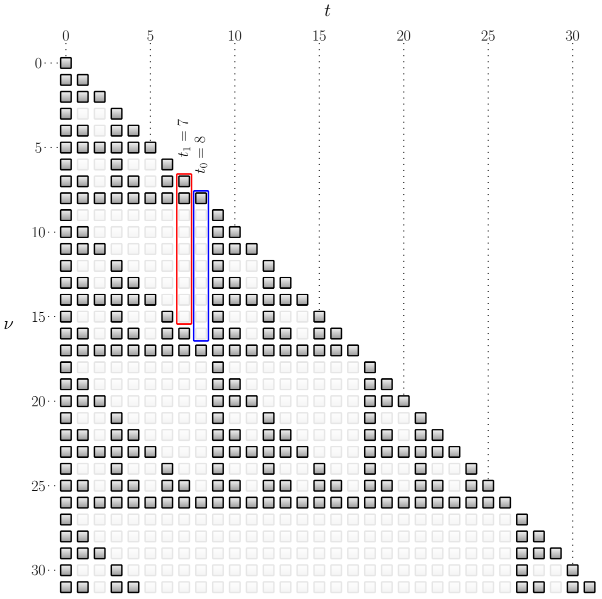


Fig. 1. Sierpinski triangle $\mathfrak{S}_3$, where the binomial coefficients are calculated modulo 3. Pale boxes represent zero entries, the actual values of the nonzero entries are irrelevant for our purposes which is why they are represented by generic bold boxes. Column $t_0 = 8$ is a zero column for $\ell \in \{9, \dots, 16\}$ and $t_1 = 7$ is a zero column with resolvable spoiler $\binom{8}{7}$ for $\ell \in \{9, \dots, 15\}$.

modulo any positive integer p as a *Sierpinski triangle* and we denote it by $\mathfrak{S}_p$.

The summands of the outer sum in (1) are weighted by the binomial coefficients $\binom{\nu}{t}$, $\nu = t, \dots, \ell$. These are exactly the binomial coefficients that appear at the first $\ell - t + 1$ entries in column t of a Sierpinski triangle. For any $\beta \in \mathbb{F}_q$, $p\beta = 0$, where $p \triangleq \operatorname{char}[\mathbb{F}_q]$ is the characteristic of $\mathbb{F}_q$. Thus, summands for which $\binom{\nu}{t}$ is a multiple of p are zero.

For given ℓ, let us assume a column t_0 in $\mathfrak{S}_p$ with

$$\binom{t_0+1}{t_0}, \dots, \binom{\ell}{t_0} \equiv 0 \bmod p,$$

i.e., all entries except the first one (which is $\binom{t_0}{t_0} \equiv 1 \bmod p$) are zero. We refer to such columns as *zero columns*. Fig. 1 shows that zero columns actually exist, e.g., column $t_0 = 8$ for $\ell \in \{9, \dots, 16\}$ in $\mathfrak{S}_3$.

Assume that $Q(x, z)$ is a solution of Problem 1 for a received vector $\boldsymbol{y}$ and consider (1) for zero column t_0. This equation simplifies to

$$\forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - t_0 \text{ and}$$
$$\underbrace{\binom{t_0}{t_0}}_{=1} z^{t_0 - t_0} \sum_{\mu=s}^{d_{Q_{t_0}}} \binom{\mu}{s} x^{\mu-s} Q_{t_0,\mu}\Big|_{(x,z)=(\alpha_i, y_i)} = 0, \quad (2)$$

since all summands of the outer sum except the first one are annihilated by the zero binomial weights. This means that the 0th to $r - t_0 - 1$th Hasse derivatives of $Q_{t_0}(x)$ evaluate to zero at all α_i, $i \in \mathcal{I}$. But then, these α_i are roots of multiplicity $r - t_0$ of $Q_{t_0}(x)$ and thus $Q_{t_0}(x)$ can be factored as $Q_{t_0}(x) = V_{t_0}(x) P_{\mathcal{I}}(x)^{r-t_0}$, where $\deg[V_{t_0}(x)] \leq d_{Q_{t_0}} - n(r - t_0)$.

The following lemma, whose proof is based on Lucas' Theorem [5], specifies the conditions for the existence of a zero column t_0 and its location.

Lemma 1. *Let $r, \ell \in \mathbb{N} \setminus \{0\}$, $r \leq \ell$. If $\ell < p$ or*

$$\exists\, a \in \{1, \dots, p-1\}, j \in \mathbb{N} : r \leq a p^j - 1 \leq \ell$$

then a zero column does not *exist. Otherwise, find the least significant base-p digit ℓ_i of ℓ such that $\ell_i < p - 1$. Then,*

$$t_0 = \left\lfloor \frac{r}{p^{i+1}} \right\rfloor p^{i+1} - 1$$

is a zero column. In particular, t_0 is the maximal zero column.

We will now generalize the concept to zero columns with resolvable spoilers. This will reveal additional structure in solutions $Q(x, z)$ of Problem 1, i.e., to factorizations of additional univariate polynomials $Q_\nu(x)$ besides $Q_{t_0}(x)$.

Assume there exists a zero column t_0 in $\mathfrak{S}_p$. Further assume there is a column t_1, $t_1 < t_0$, in $\mathfrak{S}_p$ such that

$$\binom{t_0}{t_1} \not\equiv 0 \bmod p \text{ and}$$
$$\forall \nu = t_1 + 1, \dots, \ell, \nu \neq t_0 : \binom{\nu}{t_1} \equiv 0 \bmod p.$$

We refer to such a column as *zero column with spoiler at* $\binom{t_0}{t_1}$, an example is shown in Fig. 1.

If $Q(x, z)$ is a solution of Problem 1 for a received vector $\boldsymbol{y}$ then (1) for column t_1, $t_1 < t_0$, becomes

$$\forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - t_0 \text{ and}$$
$$\underbrace{\binom{t_1}{t_1}}_{=1} z^{t_1 - t_1} \sum_{\mu=s}^{d_{Q_{t_1}}} \binom{\mu}{s} x^{\mu-s} Q_{t_1,\mu} +$$
$$\binom{t_0}{t_1} z^{t_0 - t_1} \sum_{\mu=s}^{d_{Q_{t_0}}} \binom{\mu}{s} x^{\mu-s} Q_{t_0,\mu}\Big|_{(x,z)=(\alpha_i, y_i)} = 0, \quad (3)$$

since all summands except the ones for $\nu = t_1$ and $\nu = t_0$ are annihilated. But, according to (2), the second sum evaluates to zero at all α^{-i}, $i \in \mathcal{I}$, since t_0 is by assumption a zero column. We refer to $\binom{t_0}{t_1}$ as a *resolvable spoiler* for t_1 because the sum associated with $\binom{t_0}{t_1}$ vanishes. As a result, we obtain

$$\forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - t_0 \text{ and}$$
$$\sum_{\mu=s}^{d_{Q_{t_1}}} \binom{\mu}{s} x^{\mu-s} Q_{t_1,\mu}\Big|_{x=\alpha_i} = 0,$$

i.e., the α_i, $i \in \mathcal{I}$, are roots of multiplicity $r - t_0$ of $Q_{t_1}(x)$ and thus it can be factored as $Q_{t_1}(x) = V_{t_1}(x) P_{\mathcal{I}}(x)^{r-t_0}$, where $\deg[V_{t_1}(x)] \leq d_{Q_{t_1}} - n(r - t_0)$.

It is easy to see that if $\binom{t_1}{t_2}$ is a spoiler for t_2, $t_2 < t_1$, then it is also resolvable. Furthermore, it is easy to see that the concept generalizes to multiple spoilers. This leads to the following recursive definition:

Definition 2. Consider $\mathfrak{S}_p$ and $r, \ell \in \mathbb{N} \setminus \{0\}$. For $\nu \le t_0$ let

$$\begin{aligned}\mathcal{R}_\nu^{(\ell)} \triangleq \{&t \in \mathbb{N} : \nu < t \le t_0,\\ &t \text{ is a zero column with resolvable spoilers}\}\end{aligned}$$

and

$$\mathcal{S}_\nu^{(\ell)} \triangleq \left\{ t \in \mathbb{N} : \binom{t}{\nu} \text{ is a spoiler for } \nu \right\}.$$

Then ν is a *zero column with resolvable spoilers* if and only if $\mathcal{S}_\nu^{(\ell)} \subseteq \mathcal{R}_\nu^{(\ell)}$. The basic case is $\mathcal{R}_{t_0}^{(\ell)} = \{t_0\}$ if t_0 exists.

Note that zero columns are special cases of zero columns with resolvable spoilers where $\mathcal{S}_\nu^{(\ell)} = \emptyset$. The sets of zero columns with resolvable spoilers are non-increasing with ν, i.e., $\mathcal{R}_{\nu+1}^{(\ell)} \subseteq \mathcal{R}_\nu^{(\ell)}$. We stress that $\mathcal{R}_0^{(\ell)}$ contains all zero columns with resolvable spoilers in $\mathfrak{S}_p$, i.e., it is the set that we are interested in.

Lemma 2. *Let $r, \ell \in \mathbb{N} \setminus \{0\}$, $r \le \ell$. If according to Lemma 1 a maximal zero column t_0 of $\mathfrak{S}_p$ exists then the set of zero columns with resolvable spoilers of $\mathfrak{S}_p$ is*

$$\mathcal{R}_0^{(\ell)} = \left\{ t \in \mathbb{N} : t < r \text{ and } \sum_{\ell'=t_0+1}^{\ell} \left(\binom{\ell'}{t} \bmod p \right) = 0 \right\},$$

otherwise it is $\mathcal{R}_0^{(\ell)} = \emptyset$.

Example 1. Consider the conventional RS code $\mathcal{GRS}_{\mathcal{A},\mathcal{B}}(\mathbb{F}_{27}; 26, 16)$. The characteristic of the code's base field is $p = 3$. Lemma 1 yields maximal zero column $t_0 = 8$. Fig. 1 shows that for $t_1 = 7$ we have $\mathcal{S}_7^{(13)} = \{8\} = \mathcal{R}_7^{(13)}$, i.e., $t_1 = 7$ is a zero column with resolvable spoilers and we can set $\mathcal{R}_6^{(13)} = \{7, 8\}$. For $t_2 = 6$ we have $\mathcal{S}_6^{(13)} = \{7, 8\} = \mathcal{R}_6^{(13)}$ and thus it is a zero column with resolvable spoilers as well. This gives $\mathcal{R}_5^{(13)} = \{6, 7, 8\}$. For $t_3 = 5$ we have $\mathcal{S}_5^{(13)} = \{8\} \subseteq \mathcal{R}_5^{(13)}$ and thus it is a zero column with resolvable spoilers as well. It turns out that for all $t_4 < t_3$ holds $\mathcal{S}_{t_4}^{(13)} \not\subseteq \{5, 6, 7, 8\} = \mathcal{R}_{t_4}^{(13)}$ and thus the only zero columns with resolvable spoilers in $\mathfrak{S}_3$ with respect to $\ell = 13$ are $t_0 = 8$, $t_1 = 7$, $t_2 = 6$, and $t_3 = 5$. It can be readily checked that Lemma 2 confirms this result and delivers $\mathcal{R}_0^{(13)} = \{5, 6, 7, 8\}$.

The following map will turn out to be useful in the following, it returns either ν itself or its greatest spoiler:

$$g : \left\{ \begin{array}{ccl} \mathbb{N} & \to & \mathbb{N} \\ \nu & \mapsto & \left\{ \begin{array}{ll} \max\{\mathcal{S}_\nu^{(\ell)}\} & \mathcal{S}_\nu^{(\ell)} \neq \emptyset \\ \nu & \mathcal{S}_\nu^{(\ell)} = \emptyset \end{array} \right. \end{array} \right. . \qquad (4)$$

Theorem 1. *Let $\mathcal{GRS}_{\mathcal{A},\mathcal{B}}(\mathbb{F}_q; n, k)$ be a GRS code and r, ℓ such that the GSA can correct at most ε_0 errors. Let further $\boldsymbol{c} \in \mathcal{GRS}$, $\boldsymbol{e} \in \mathbb{F}_q^n$ with $\mathrm{wt}_{\mathrm{H}}[\boldsymbol{e}] \le \varepsilon_0$ and $\boldsymbol{y} = \boldsymbol{c} + \boldsymbol{e}$. When the GSA is applied to $\boldsymbol{y}$ it yields a bivariate result polynomial $Q(x, z) = Q_0(x) + Q_1(x)z + \cdots + Q_\ell(x)z^\ell \in \mathbb{F}_q[x, z]$ whose constituent univariate polynomials $Q_\nu(x)$, $\nu \in \mathcal{R}_0^{(\ell)}$, can be factored as*

$$Q_\nu(x) = V_\nu(x) P_{\mathcal{I}}(x)^{r - g[\nu]}, \qquad (5)$$

where $\deg[V_\nu(x)] \le d_{Q_\nu} - n\,(r - g[\nu]) \triangleq d_{V_\nu}$.

Proof: Let $\nu \in \mathcal{R}_0^{(\ell)}$. Since ν is a zero column with resolvable spoilers, all $\binom{t}{\nu}$ with $t \in \mathcal{S}_\nu^{(\ell)}$ are resolvable, i.e.,

$$\begin{aligned}&\forall t \in \mathcal{S}_\nu^{(\ell)} \forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - t \text{ and}\\ &\qquad \sum_{\mu=s}^{d_{Q_t}} \binom{\mu}{s} x^{\mu - s} Q_{t,\mu} \Big|_{x = \alpha_i} = 0. \qquad (6)\end{aligned}$$

Since by definition all terms except the ones weighted by the spoilers vanish, we can write (1) as[1]

$$\begin{aligned}&\forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - \nu \text{ and}\\ &\sum_{t \in \mathcal{S}_\nu^{(\ell)}} \binom{t}{\nu} z^{t-\nu} \sum_{\mu=s}^{d_t} \binom{\mu}{s} x^{\mu-s} Q_{t,\mu} +\\ &\qquad \sum_{\mu=s}^{d_{Q_\nu}} \binom{\mu}{s} x^{\mu-s} Q_{\nu,\mu} \Big|_{(x,z)=(\alpha_i, y_i)} = 0.\end{aligned}$$

In order to let the sum over t vanish in the case $\mathcal{S}_\nu^{(\ell)} \neq \emptyset$ (i.e., to exploit (6)), we must guarantee $s < r - t$ for all $t \in \mathcal{S}_\nu^{(\ell)}$, i.e., $s < r - \max\{\mathcal{S}_\nu^{(\ell)}\}$. In case $\mathcal{S}_\nu^{(\ell)} = \emptyset$ the sum over t is empty and thus it is sufficient to guarantee $s < r - \nu$. Due to the definition (4) of $g[\nu]$ we have $s < r - g[\nu]$ in both cases and thus

$$\begin{aligned}&\forall i \in \mathcal{I}\ \forall s \in \mathbb{N} : s < r - g_\nu \text{ and}\\ &\qquad \sum_{\mu=s}^{d_{Q_\nu}} \binom{\mu}{s} x^{\mu-s} Q_{\nu,\mu} \Big|_{x=\alpha_i} = 0\end{aligned}$$

and the α_i, $i \in \mathcal{I}$, are roots of multiplicity $r - g[\nu]$ of $Q_\nu(x)$ and thus it can be factored as in (5). The bound on the degrees of the $V_\nu(x)$ follows from a comparison of the involved polynomial degrees. ∎

We stress that the *Sierpinski prefactors* $P_{\mathcal{I}}(x)^{r-g[\nu]}$, $\nu \in \mathcal{R}_0^{(\ell)}$, are fixed a-priori and do not depend on the received vector $\boldsymbol{y}$. This justifies the term *pre*factor.

Example 2. From Example 1 we have $\mathcal{R}_0^{(13)} = \{5, 6, 7, 8\}$, i.e., Theorem 1 guarantees factorizations of $Q_5(x)$, $Q_6(x)$, $Q_7(x)$, and $Q_8(x)$. They are $Q_5(x) = V_5(x)P_{\mathcal{I}}(x)^2$, $Q_6(x) = V_6(x)P_{\mathcal{I}}(x)^2$, $Q_7(x) = V_7(x)P_{\mathcal{I}}(x)^2$, and $Q_8(x) = V_8(x)P_{\mathcal{I}}(x)^2$, with $d_{V_5} = 72$, $d_{V_6} = 57$, $d_{V_7} = 42$, and $d_{V_8} = 27$, respectively, because $g[5] = g[6] = g[7] = g[8] = 8$, $d_{Q_5} = 124$, $d_{Q_6} = 109$, $d_{Q_7} = 94$, and $d_{Q_8} = 79$.

Theorem 1 states that some of the univariate constituent polynomials of the GSA interpolation step (Problem 1) have certain prefactors. We will now show how this knowledge can be exploited in order to simplify solving the interpolation step. More precisely, we show that the associated linear system of equations in $\sum_{\nu=0}^{\ell}(d_{Q_\nu} + 1)$ unknowns can be reduced to a linear system of smaller size.

As noted before, GSA interpolation (Problem 1) amounts to finding the solution of a linear system of equations. This can be done naively using Gaussian elimination. Several faster methods have been developed, all of which exploit the *structure of the involved coefficient matrix*. Such methods can be

[1] Note that this is the generalization of (3) to multiple spoilers.

applied to a reduced linear system of equations, which can be obtained using Sierpinski prefactors. The key idea here is to exploit the a-priori known *structure of the solutions*.

In the following, it will be convenient to have the sets

$$\mathcal{F} \triangleq \{\nu : 0 \le \nu \le \ell, Q_\nu(x) \text{ has prefactor}\} \text{ and}$$
$$\mathcal{F}^c \triangleq \{\nu : 0 \le \nu \le \ell, \nu \notin \mathcal{F}\}.$$

Let us consider the factorization of a univariate constituent polynomial $Q_\nu(x)$, $\nu \in \mathcal{F}$, into a Sierpinski prefactor $P_{\mathcal{I}}(x)^{r-g[\nu]}$ and the corresponding quotient polynomial $V_\nu(x)$. For simplicity, let us denote the $P_{\mathcal{I}}(x)^{r-g[\nu]}$ by $F_\nu(x) = \sum_{\mu=0}^{\deg[F_\nu(x)]} F_{\nu,\mu}x^\mu$. This gives $Q_\nu(x) = V_\nu(x)F_\nu(x)$ for $\nu \in \mathcal{F}$ with coefficients

$$Q_{\nu,\mu} = \sum_{i=0}^{\mu} V_{\nu,\mu-i}F_{\nu,i}, \quad \mu = 0, \ldots, d_{Q_\nu}, \tag{7}$$

where we implicitly used that the ith coefficient of a polynomial with $i < 0$ or i greater than the degree of the polynomial is zero. In order to simplify the following description, let us agree on trivial prefactors $F_\nu(x) = 1$ for all $Q_\nu(x)$, $\nu \in \mathcal{F}^c$. In these cases, the quotient polynomials are $V_\nu(x) = Q_\nu(x)$. Note that the constant term $F_{\nu,0}$ of any prefactor is nonzero. This allows us to write

$$\begin{aligned} V_{\nu,\mu} &= \frac{Q_{\nu,\mu} - \sum_{i=1}^{\mu} V_{\nu,\mu-i}F_{\nu,i}}{F_{\nu,0}} \\ &= \frac{Q_{\nu,\mu}}{F_{\nu,0}} - \sum_{i=1}^{\mu} \frac{F_{\nu,i}V_{\nu,\mu-i}}{F_{\nu,0}}, \ \mu = 0, \ldots, d_{Q_\nu}, \end{aligned} \tag{8}$$

which shows that $V_{\nu,\mu}$ is a linear combination of the $V_{\nu,i}$, $i = 0, \ldots, \mu-1$, and $Q_{\nu,\mu}$.

We can exploit (8) for $\mu = 0, \ldots, d_{V_\nu}$ in order to obtain a solvable linear system whose solution comprises the coefficients of $V_\nu(x)$ (*preparation*) and then exploit (7) for $\mu = d_{V_\nu} + 1, \ldots, d_{Q_\nu}$ in order to dispose of the redundant columns of the coefficient matrix (*reduction*). Both steps — preparation and reduction — are based on applying a simple lemma from linear algebra with certain parameters.

As a result of this process, the original coefficient matrix associated with (1), whose $\sum_{\nu=0}^{\ell}(d_{Q_\nu} + 1)$ columns are associated with $Q_{0,0}, \ldots, Q_{0,d_{Q_0}}, Q_{1,0}, \ldots, Q_{1,d_{Q_1}}, \ldots, Q_{\ell,0}, \ldots, Q_{\ell,d_{Q_\ell}}$ is converted into a reduced matrix, whose $\sum_{\nu=0}^{r-1}(d_{V_\nu} + 1) + \sum_{\nu=r}^{\ell}(d_{Q_\nu} + 1)$ columns are associated with $V_{0,0}, \ldots, V_{0,d_{Q_0}}, V_{1,0}, \ldots, V_{1,d_{Q_1}}, \ldots, V_{\ell,0}, \ldots, V_{\ell,d_{Q_\ell}}$. This allows the following reformulation of a reduced GSA interpolation step:

Problem 2 (Reduced GSA Interpolation Step)**.** Given a received vector $\boldsymbol{y}$ with Sierpinski prefactors $F_\nu(x) = P_{\mathcal{I}}(x)^{r-g[\nu]}$, $\nu \in \mathcal{F}$, find a nonzero bivariate polynomial

$$\widetilde{Q}(x,z) = \sum_{\nu\in\mathcal{F}} V_\nu(x)y^\nu + \sum_{\nu\in\mathcal{F}^c} Q_\nu(x)y^\nu \in \mathbb{F}_q[x,z]$$

such that $\deg[V_\nu(x)] \le d_{V_\nu}$ and $\deg[Q_\nu(x)] \le d_{Q_\nu}$ and

$\forall i \in \mathcal{I}\ \forall s,t \in \mathbb{N} : s+t < r$ and

$$\sum_{\substack{\nu\in\mathcal{F}\\ \nu\ge t}} \binom{\nu}{t} z^{\nu-t} \sum_{\mu=0}^{d_{V_\nu}} \underbrace{\left(\sum_{\Delta=0}^{d_{Q_\nu}-d_{V_\nu}} \binom{\mu+\Delta}{s} x^{\mu+\Delta-s} F_{\nu,\Delta}\right)}_{\triangleq \mathrm{LUT}[s,i,\nu,\mu]} V_{\nu,\mu}$$
$$+ \sum_{\substack{\nu\in\mathcal{F}^c\\ \nu\ge t}} \binom{\nu}{t} z^{\nu-t} \sum_{\mu=s}^{d_{Q_\nu}} \underbrace{\binom{\mu}{s} x^{\mu-s}}_{\triangleq \mathrm{LUT}[s,i,\nu,\mu]} Q_{\nu,\mu}\Big|_{(x,z)=(\alpha_i,y_i)} = 0.$$

The sums with summation index μ are independent of the received vector $\boldsymbol{y}$ and thus their addends can be pre-calculated and stored in a lookup table $\mathrm{LUT}[s,i,\nu,\mu]$. A solution $Q(x,z)$ of Problem 1 can easily be recovered from a solution $\widetilde{Q}(x,z)$ of Problem 2 using the prefactors. $Q(x,z)$ can then be used as input to the GSA factorization step in order to complete the decoding. A *reduced* GSA factorization step that can operate directly on $\widetilde{Q}(x,z)$ in order to construct the result list was proposed in [1], [2].

Example 3. Sierpinski prefactors work particularly well for the two conventional RS codes $\mathcal{GRS}_1(\mathbb{F}_{255}; 255, 191, 65)$ and $\mathcal{GRS}_2(\mathbb{F}_{255}; 255, 144, 112)$ considered by Kötter and Vardy in [6]. The GSA for $\mathcal{GRS}_1$ can correct up to $\varepsilon_0 = 34$ errors with multiplicity $r = 16$ and list size $\ell = 18$. This requires solving a linear system in 34694 unknowns. Sierpinski prefactors reduce the system to 31379 unknowns. The GSA with $r = 4$ and $\ell = 5$ for $\mathcal{GRS}_2$ can correct up to $\varepsilon_0 = 59$ errors. The associated linear system has 2559 unknowns, which can be diminished to 2049 unknowns using Sierpinski prefactors.

We emphasize that Sierpinski prefactors can easily be combined with the re-encoding projection, resulting in a significant reduction of the Guruswami–Sudan interpolation step beyond the reduction enabled by re-encoding alone. More details and the missing proofs are provided in [7].

Acknowledgment

The author is grateful to Frank R. Kschischang for his valuable comments on the manuscript. This work has been supported by DFG under grant SE 2279/1-1.

References

[1] R. Koetter, J. Ma, A. Vardy, and A. Ahmed, "Efficient interpolation and factorization in algebraic soft-decision decoding of Reed–Solomon codes," in *International Symposium on Information Theory. ISIT 2003.* IEEE, Jun. 2003, p. 365, doi: 10.1109/isit.2003.1228381.

[2] R. Koetter, J. Ma, and A. Vardy, "The re-encoding transformation in algebraic list-decoding of Reed–Solomon codes," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 633–647, Feb. 2011, doi: 10.1109/tit.2010.2096034.

[3] V. Guruswami and M. Sudan, "Improved decoding of Reed–Solomon and algebraic-geometry codes," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1757–1767, Sep. 1999, doi: 10.1109/18.782097.

[4] R. Koetter, "On Algebraic Decoding of Algebraic-Geometric and Cyclic Codes," Ph.D. dissertation, University of Linköping, 1996.

[5] E. Lucas, "Théorie des fonctions numériques simplement périodiques," *American Journal of Mathematics*, vol. 1, no. 2-4, Jan. 1878, doi: 10.2307/2369308.

[6] R. Koetter and A. Vardy, "Algebraic soft-decision decoding of Reed–Solomon codes," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2809–2825, Nov. 2003, doi: 10.1109/tit.2003.819332.

[7] C. Senger, "Prefactor reduction of the Guruswami–Sudan interpolation step," *preprint* `arXiv:1309.7901 [cs.IT]`, 2013.

Improved Decoding of Partial Unit Memory Codes Using List Decoding of Reed–Solomon Codes

Sven Puchinger*, Antonia Wachter-Zeh◇ and Martin Bossert*
*Institute of Communications Engineering, University of Ulm, Ulm, Germany
◇Computer Science Department, Technion–Israel Institute of Technology, Haifa, Israel
{sven.puchinger | martin.bossert}@uni-ulm.de, antonia@cs.technion.ac.il

Abstract—**An existing bounded minimum distance decoding algorithm for Partial Unit Memory codes is improved by using list decoding of Reed–Solomon codes. Furthermore, a sufficient decoding condition is given and upper bounds on the complexity and error probability are derived.**

Index Terms—**Convolutional codes, Partial Unit Memory Codes, Reed–Solomon codes, list decoding**

I. Introduction

(Partial) Unit Memory ((P)UM) codes are special convolutional codes with memory $m = 1$ and were introduced by Lee [1] and Lauer [2] in the 1970s. Their construction is based on block codes, providing a better algebraic structure than general convolutional codes. In the 1990s, Dettmar and Sorger [3] introduced a *bounded minimum distance* (BMD) decoding algorithm for (P)UM codes that guarantees to correct up to an error bound that can be derived from the minimum distances of the used block codes. This decoding method can decode a (P)UM code sequence polynomially in the field size, whereas the Viterbi algorithm [4] might not be able to find the *Maximum Likelihood* (ML) sequence in sufficiently short time, because (P)UM codes are usually defined over large fields.

Reed–Solomon (RS) codes were discovered in 1960 and are commonly used in a broad spectrum of applications due to the existence of fast and efficient decoding algorithms. Some of these algorithms, like the *Guruswami–Sudan* algorithm, are able to decode beyond half the minimum distance.

This paper combines the advantages of both, (P)UM codes and list decoding of RS codes. We introduce an improved version of the BMD algorithm given by Dettmar and Sorger [3] for (P)UM codes based on RS codes using Guruswami–Sudan list decoders instead of BMD block decoders. We also state a sufficient decoding condition for this algorithm and derive upper bounds on the error probability and complexity.

Section II provides basic notations, defines RS and (P)UM codes and outlines the Guruswami–Sudan algorithm. In Section III, we introduce the improved algorithm and prove that the stated condition is indeed a sufficient decoding condition. Upper bounds on the error probability and complexity are shown in Section IV and Section V illustrates simulation results. Finally, Section VI gives a conclusion.

Due to space restrictions, some of the proofs are skipped, but can be found in the long version of this paper [5].

This work was supported by the German Research Council "Deutsche Forschungsgemeinschaft" (DFG) under Grant No. Bo 867/21-2.

II. Definitions and Notations

A. Notations

Let q be a power of a prime, let $\mathbb{F}$ denote the finite field of order q, let $\mathbb{F}[x]$ denote the polynomial ring over $\mathbb{F}$ and $\mathbb{F}[x,y]$ the bivariate polynomial ring over $\mathbb{F}$. We denote by $\mathbb{F}^n = \mathbb{F}^{1\times n}$ the set of all *row* vectors of length n over $\mathbb{F}$ and the elements of a vector $\mathbf{a}_j \in \mathbb{F}^n$ by $\mathbf{a}_j = (a_j^{(0)}, a_j^{(1)}, \dots, a_j^{(n-1)})$. Furthermore, we introduce a helpful notation to use parts of a vector $\mathbf{a}_j$, namely $\mathbf{a}_j^{[\ell_1,\ell_2]} := (a_j^{(\ell_1)}, a_j^{(\ell_1+1)}, \dots, a_j^{(\ell_2-1)})$ for all $0 \leq \ell_1 < \ell_2 \leq n$ and $\mathbf{a}_j^{[\ell]} := \mathbf{a}_j^{[0,\ell]}$ for all $0 \leq \ell \leq n$.

B. Reed–Solomon Codes

Let $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ be distinct elements of $\mathbb{F}$ with $n \leq q$. A *Reed–Solomon* (RS) code $\mathcal{RS}(n,k)$ of length n and dimension k over $\mathbb{F}$ with $n \leq q$ is given by

$$\mathcal{RS}(n,k) = \Big\{ (f(\alpha_0) \ \dots f(\alpha_{n-1})) : f(x) \in \mathbb{F}[x], \deg f(x) < k \Big\}.$$

RS codes are *Maximum Distance Seperable* (MDS) codes, i.e., their minimum Hamming distance is $d = n - k + 1$.

C. The Guruswami–Sudan Algorithm

The Guruswami–Sudan list decoding algorithm solves the following problem.

Problem 1 *Given $\mathbf{r} \in \mathbb{F}^n$, find a bivariate polynomial $Q(x,y) \in \mathbb{F}[x,y]$ of the form $Q(x,y) = \sum_{j=0}^{\ell} Q_j(x)y^j$, such that for given integers s, τ and ℓ:*

1) *(α_i, r_i) are zeros of $Q(x,y)$ of multiplicity s, $\forall i = 1,\dots,n$,*
2) $\deg Q_j(x) \leq s(n-\tau) - 1 - j(k-1)$, $\forall j = 0,\dots,\ell$,
3) $Q(x,y) \neq 0$.

The Guruswami–Sudan algorithm returns a list of polynomials that are y-roots of $Q(x,y)$, i.e., they satisfy $(y - f(x)) | Q(x,y)$. It was proven in [6] that these polynomials include all evaluation polynomials $f(x)$, which generate codewords of Hamming distance less than or equal to τ to $\mathbf{r}$. The maximum value of τ for given s and ℓ can be found in [7, page 131], and is greater than half the minimum distance for sufficiently large s and ℓ.

Due to the restriction of the y-degree, the *list size* is upper bounded by ℓ. However, it turns out that for most parameters the average list size is notably smaller than this parameter, see e.g., McEliece [8].

D. (Partial) Unit Memory Codes

The encoding rule for a code block of a (P)UM code of length n is given by $\mathbf{c}_j = \mathbf{i}_j \cdot \mathbf{G}_0 + \mathbf{i}_{j-1} \cdot \mathbf{G}_1$, for $\mathbf{i}_j, \mathbf{i}_{j-1} \in \mathbb{F}^k$ and $\mathbf{G}_0$ and $\mathbf{G}_1$ are $k \times n$ matrices. Both matrices have full rank if we construct an (n, k) UM code. For an $(n, k \mid k_1)$ PUM code, $\mathrm{rank}(\mathbf{G}_0) = k$ and $\mathrm{rank}(\mathbf{G}_1) = k_1 < k$ hold, such that $\mathbf{G}_0 = \left(\begin{smallmatrix}\mathbf{G}_{00}\\ \mathbf{G}_{01}\end{smallmatrix}\right)$ and $\mathbf{G}_1 = \left(\begin{smallmatrix}\mathbf{G}_{10}\\ \mathbf{0}\end{smallmatrix}\right)$,where $\mathbf{G}_{00}$ and $\mathbf{G}_{10}$ are $k_1 \times n$ matrices and $\mathbf{G}_{01}$ is a $(k - k_1) \times n$-matrix. As notation, let the generator matrices

$$\mathbf{G}_0,\ \begin{pmatrix}\mathbf{G}_{10}\\ \mathbf{G}_{01}\end{pmatrix},\ \mathbf{G}_{01} \text{ and } \mathbf{G}_\alpha = \begin{pmatrix}\mathbf{G}_{00}\\ \mathbf{G}_{01}\\ \mathbf{G}_{10}\end{pmatrix}$$

define the block codes $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_{01}$ and $\mathcal{C}_\alpha$ and τ_0, τ_1, τ_{01} and τ_α the decoding radii of corresponding block decoding algorithms. In this paper, we assume that these codes are RS codes and the decoders realize the Guruswami–Sudan algorithm with the parameters (τ_i, s_i, ℓ_i), $i \in \{0, 1, 01, \alpha\}$. A concrete construction scheme for these codes can be found in [9, page 30]. We denote such codes as $(n, k \mid k_1)$ RS PUM codes.

III. Decoding Algorithm

A. Decoding Condition

Let the received sequence $\mathbf{r} = \mathbf{c} + \mathbf{e} = (\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{N-1})$ be given, where $\mathbf{r}_h = \mathbf{c}_h + \mathbf{e}_h$, $h = 0, \dots, N-1$ is in $\mathbb{F}^n$, $\mathbf{c} = (\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{N-1})$ is a codeword of the $(n, k \mid k_1)$ RS PUM code as in Section II-D, with $\mathbf{i}_j = 0$ for all $j \geq N-1$, and $\mathbf{e}_h$ is an error block of Hamming weight $\mathrm{wt}(\mathbf{e}_h)$. In the following, we assume that for each of the underlying RS block codes $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_\alpha$ and $\mathcal{C}_{01}$, we have a list decoder, which can correct up to τ_0, τ_1, τ_α and τ_{01} errors respectively. In order to give a sufficient decoding condition for RS PUM codes we need the following definition.

Definition 1 *We define the following integers:*

$$\tau_j^r := \begin{cases}\tau_{01}, & j = 1,\\ \tau_0 + (j-2)\tau_\alpha + \tau_1, & j > 1,\end{cases}$$
$$\tau_j^c := \tau_0 + (j-1)\tau_\alpha \quad j \geq 1,$$
$$\tau_j^{rc} := (j-1)\tau_\alpha + \tau_1 \quad j \geq 1.$$

Section III-C shows that the first one of the following conditions is sufficient that the reduced trellis contains the ML path after decoding step 3 of the improved algorithm

$$\sum_{i=k}^{k+j-1} \mathrm{wt}\,\mathbf{e}_i \leq \tau_j^r, \quad \forall j, k \tag{1}$$

and that the block of the ML sequence in block $\mathbf{r}_t$ is in the reduced trellis after step 4 if

$$\sum_{i=k}^{k+j-1} \mathrm{wt}\,\mathbf{e}_i \leq \tau_j^r, \quad \forall j, k \text{ with } k \leq t \leq j+k-1. \tag{2}$$

B. Algorithm

Our algorithm is an improved version of the BMD PUM decoding method by Dettmar and Sorger [3] using list decoding of RS codes. It first constructs a reduced trellis and then applies the Viterbi Algorithm to it.

We show that if Condition 1 is satisfied for a received sequence $\mathbf{r}$, then the ML path is contained in the reduced trellis and thus, can be found by the Viterbi algorithm.

In the following, the algorithm for decoding a received sequence $\mathbf{r} = (\mathbf{r}_0, \dots, \mathbf{r}_{N-1})$ of length N is presented in detail.

In the **first step** of the algorithm all received words $\mathbf{r}_j$ for $j = 0, \dots, N-1$ are decoded individually in $\mathcal{C}_\alpha$. Actually, $\mathbf{r}_0$ can be decoded in $\mathcal{C}_0$ and $\mathbf{r}_{N-1}$ in $\mathcal{C}_{10}$ because we know that $\mathbf{i}_{-1} = \mathbf{i}_{N-1} = \mathbf{0}$. If decoding does not fail, we obtain a list of information word tuples $\left(\hat{\mathbf{i}}_{j-1}^{\mu,[k_1]}, \hat{\mathbf{i}}_j^{\mu}\right)$ and trellis edges $\hat{\mathbf{c}}_j^{\mu}$ with $\mu \in \{1, \dots, \ell_j\}$ for each level j of the trellis, where ℓ_j is upper bounded by the maximum list size ℓ_α of the list decoder of $\mathcal{C}_\alpha$. Moreover, we define a metric

$$m_j^{(\alpha)} = \begin{cases}\tau_\alpha + 1, & \text{if decoding fails,}\\ \max\limits_{\mu}\left\{d(\mathbf{r}_j, \hat{\mathbf{c}}_j^{\mu})\right\}, & \text{else.}\end{cases}$$

Step two of the algorithm uses the results of step one for decoding in forward and backward direction. In particular, that means that we take all "left nodes" $\hat{\mathbf{i}}_{j-1}^{\mu,[k_1]}$ of step one and decode $l_B^{(j)}$ steps in $\mathcal{C}_1$ in backward direction:

$$\mathbf{r}_{j-i} - \hat{\mathbf{i}}_{j-i}^{\mu,[k_1]}\mathbf{G}_{00} = \left(\hat{\mathbf{i}}_{j-i-1}^{\nu(\mu),[k_1]}, \hat{\mathbf{i}}_{j-i}^{\nu(\mu),[k_1,k]}\right)\begin{pmatrix}\mathbf{G}_{10}\\ \mathbf{G}_{01}\end{pmatrix} + \mathbf{e}_{j-i},$$

where the left side of the equation is known and the information words $\left(\hat{\mathbf{i}}_{j-i-1}^{\nu(\mu),[k_1]}, \hat{\mathbf{i}}_{j-i}^{\nu(\mu),[k_1,k]}\right)$ are the result of decoding it in $\mathcal{C}_1$. Furthermore, we take all "right nodes" $\hat{\mathbf{i}}_j^{\mu}$ and decode $l_F^{(j)}$ steps in $\mathcal{C}_0$ in forward direction:

$$\mathbf{r}_{j+i} - \hat{\mathbf{i}}_{j+i-1}^{\mu,[k_1]}\mathbf{G}_{10} = \hat{\mathbf{i}}_{j+i}^{\nu(\mu)}\mathbf{G}_0 + \mathbf{e}_{j+i},$$

where $l_B^{(j)}$ and $l_F^{(j)}$ are defined as follows:

$$l_B^{(j)} = \min\left\{i : \sum_{t=1}^{i}(2\tau_\alpha + 1 - m_{j-t}^{(\alpha)}) > \tau_i^{rc}\right\},$$
$$l_F^{(j)} = \min\left\{i : \sum_{t=1}^{i}(2\tau_\alpha + 1 - m_{j+t}^{(\alpha)}) > \tau_i^{c}\right\}.$$

The following **Step three** of the algorithm makes sure that even a certain class of error patterns with scattered peaks of error blocks with up to τ_{01} errors can be decoded. For every block j we have to take all nodes $\hat{\mathbf{i}}_{j-1}$ and $\hat{\mathbf{i}}_j^{[k_1]}$ from Steps 1 and 2. Then we calculate $\hat{\mathbf{i}}_j^{[k_1,k]}$ by decoding in $\mathcal{C}_{01}$:

$$\mathbf{r}_j - \hat{\mathbf{i}}_{j-1}^{[k_1]}\mathbf{G}_{10} - \hat{\mathbf{i}}_j^{[k_1]}\mathbf{G}_{00} = \hat{\mathbf{i}}_j^{[k_1,k]}\mathbf{G}_{01} + \mathbf{e}_j$$

The **fourth step** is needed to ensure correct decoding of a certain block $\mathbf{r}_t$ that satisfies Condition (2) if the entire sequence does not fulfill (1). We have to define an erasure node for every level j of the reduced trellis and link it to every node in the levels $j-1$ and $j+1$ (including erasure nodes) using the designed edge costs of Definition 2.

Finally, **Step 5** applies the Viterbi algorithm (cf. [4]) to the reduced trellis in order to obtain the ML sequence.

Step 4 of the algorithm needs the following designed edge costs.

Definition 2 (Designed edge costs for Step 4) *The edge cost between an erasure node $\mathbf{i}_j^{(e)}$ at level j of the trellis and information words $\mathbf{i}_{j-1}$ and $\mathbf{i}_{j+1}$ at levels $j-1$ and $j+1$, which are found in Steps 1 and 2 are defined as follows:*

1) *The edge cost $m_j^{(ef)}$ between every information word $\mathbf{i}_{j-1}$ and the erasure node $\mathbf{i}_j^{(e)}$ is given by*
$$m_j^{(ef)} = \max\left(\tau_0+1, 2\tau_0+1-\zeta_F\right)$$
where ζ_F is defined as the smallest Hamming distance between received block and estimated codeblock of all edges which connect $\mathbf{i}_{j-1}$ with any node of the reduced trellis at level j.
2) *The edge cost $m_j^{(eb)}$ between the erasure node $\mathbf{i}_j^{(e)}$ and every information word $\mathbf{i}_{j+1}$ is given by*
$$m_j^{(eb)} = \max\left(\tau_1+1, 2\tau_1+1-\zeta_B\right)$$
where ζ_B is defined as the smallest estimated number of errors of all edges which connect any node of the reduced trellis at level j with $\mathbf{i}_{j+1}$.
3) *The edge cost $m_j^{(ee)}$ between the erasure nodes $\mathbf{i}_j^{(e)}$ and $\mathbf{i}_{j+1}^{(e)}$ is given by*
$$m_j^{(ee)} = \begin{cases} 2\tau_\alpha+1-\zeta_\alpha, & \text{if } \exists\, \mathbf{i}_j, \mathbf{i}_{j+1} \text{ which are connected through an edge,} \\ \tau_\alpha+1, & \text{else,} \end{cases}$$
where ζ_α is defined as the smallest estimated number of errors of all edges which connect any two nodes i_j and i_{j+1} of the reduced trellis.

C. Proof of Correctness

Lemma 1 *If* (1) *is satsified, then the gap between two adjacent correctly decoded blocks in Step 1 is smaller than $L := min\{L_B, L_F\}$, where*
$$L_B^{(j)} = \min\left\{ i : \sum_{t=1}^{i} (2\tau_\alpha + 1 - m_{j-t}^{(\alpha)}) > \tau_i^r \right\}, \tag{3}$$
$$L_F^{(j)} = \min\left\{ i : \sum_{t=1}^{i} (2\tau_\alpha + 1 - m_{j+t}^{(\alpha)}) > \tau_i^r \right\}. \tag{4}$$

Proof: If Step 1 cannot decode a block successfully, there occured more than $(2\tau_\alpha+1-m_j^\alpha)$ errors in this block. Suppose that block $\mathbf{r}_t$ was decoded correctly and decoding in all following L blocks failed. Then,
$$\sum_{j=t+1}^{t+L} \mathrm{wt}\,\mathbf{e}_j > \sum_{j=1}^{L} (2\tau_\alpha + 1 - m_{t+j}^{\alpha}) \overset{(3)\&(4)}{>} \tau_L^r$$
in contradiction to (1). ■

Lemma 2 *Step 2 is able to find the correct path between two adjacent correct decoding decisions from Step 1 ($\mathbf{r}_t$ and $\mathbf{r}_{t+i}$) if* (1) *holds and*
$$\mathrm{wt}(\mathbf{e}_j) \le \min\{\tau_0, \tau_1\} \quad \forall j \in \{t+1, \ldots, t+i-1\},$$
is satsified.

Lemma 3 *If* (1) *is satisfied between two adjacent correct decisions from Step 1 ($\mathbf{r}_t$ and $\mathbf{r}_{t+i}$), the following holds*
$$\mathrm{wt}(\mathbf{e}_j) + \mathrm{wt}(\mathbf{e}_k) \le \tau_0 + \tau_1,$$
$$\forall j,k \in \{t+1, \ldots, t+i-1\}, j \neq k.$$

Theorem 1 *If* (1) *is satisfied, the ML sequence is in the reduced trellis.*

Proof: Note that there are always at least two correctly decoded blocks found in Step 1, namely $\mathbf{c}_{-1} = \mathbf{c}_N = \mathbf{0}$, and from Lemma 1 it is clear that it is sufficient that Step 2 corrects only $l_F^{(j)}$ steps in forward and $l_B^{(j)}$ steps in backward direction.

From Lemmas 2 and 3, we know that after Step 2 there is at most a gap of one block between two correctly decoded blocks from Step 1, in which decoding both in forward and in backward direction fails.

Since $\mathrm{wt}(\mathbf{e}_j) \le \tau_{01}$ for every block j, if (1) is satisfied, we are able to close this gap in Step 3 and the complete ML path is in the reduced trellis. ■

Lemma 4 *If Condition* (2) *is satisfied for block $\mathbf{r}_t$, the most likely code block $\mathbf{c}_t$ is in the reduced trellis.*

Theorem 2 *If Condition* (2) *is satisfied for block $\mathbf{r}_t$, the Viterbi Algorithm finds the most likely code block $\mathbf{c}_t$.*

IV. Error Probability and Complexity Analysis

A. Error Probability

By proceeding in the same way as Dettmar in [9], page 74, and redefining $\rho_\nu := \tau_\nu^r + 1$, we obtain the following upper bound on the block error probability of the code for a binary symmetric channel with crossover probability p:
$$P_W^{(P)UM} \le \sum_{\nu=1}^{\infty} \nu \sum_{i=\rho_\nu}^{\rho_{\nu+1}-1} \binom{\nu n}{i} p^i (1-p)^{\nu n - i}.$$

B. Complexity Analysis

By counting the maximum number of necessary block decoding iterations, we can derive an upper bound on the decoding complexity, which holds under the assumption of Condition 2.

A detailed proof of the following theorem can be found in the long version of this paper [5].

Theorem 3 *If Condition 2 is satisfied for a certain block* $\mathbf{r}_t$ *of a RS (P)UM code sequence, the decoding complexity of this block is upper bounded by*

$$C_{(P)UM} \leq \frac{4}{L-1} L^{(\tau_0-\tau_\alpha+1)(\tau_\alpha+3)-1} C_B,$$

where C_B *is the maximum of the complexities of the decoders used in Steps 1-3 and* L *denotes the maximum size of the lists of code words which are the results of the decoders of each step of the algorithm.*

There are two important cases which we want to discuss. Theorem 3 shows that in general, the decoding complexity of a RS (P)UM code list decoder is exponential in τ_0 and τ_α.

1) In the *worst case*, the list size L is always the maximum list size of all block decoders.
2) However, McEliece [8] showed that for almost every RS code and its Guruswami–Sudan decoder, the average list size is close to 1. In *average case*, we can replace L by the upper bound

$$L \leq 1 + \overline{L}_0(\tau_B) = 1 + \sum_{s=0}^{\tau_B} (q-1)^{s-n+k} \binom{n}{s}$$

for the average list size (cf. [8]), where we define $\tau_B := \max\{\tau_\alpha, \tau_0, \tau_1, \tau_{01}\}$.

Hence, the decoding complexity is notably smaller in the average case than in the worst case. Taking the limit of $C_{(P)UM}$ for $L \to 1+$ helps to illustrate the average complexity:

$$\lim_{L\to 1+} C_{(P)UM} \leq \big[1 + 4(2\tau_0 + \tau_1 + \tau_\alpha(\tau_0 - \tau_\alpha - 2))\big] C_B,$$

which is polynomial in τ_α, τ_0 and τ_1. The derivation of this limit can be found in the long version of this paper [5].

V. Simulation Results

We now illustrate the improvements of the algorithm by simulation results. A $(31, 11 \mid 6)$ RS PUM sequence over $\mathbb{F}_{2^5}$ of length $N = 50$ is sent via *BPSK* modulation over an *AWGN* channel. Figure 1 shows the block error probability over the signal-to-noise ratio. The plot illustrates the differences between the actual decoding capability, the decoding condition and the equivalent block-by-block decoding performance of both, the BMD decoding algorithm by Dettmar and Sorger [3] and the improved (P)UM decoder.

Since the improved algorithm is able to correct more error patterns than those fulfilling condition 2, there is a difference between considering only the sufficient decoding condition and evaluating the actual decoding capability.

One can notice that the E_b/N_0*-gain* between the actual BMD PUM decoder and the improved PUM algorithm at an error probability of 10^{-4} is about 1.2 dB.

VI. Conclusion

We presented an improved version of the (P)UM decoding algorithm of Dettmar and Sorger [3] and introduced a sufficient decoding condition. Moreover, we derived upper bounds on the complexity and error probability and illustrated the improvements with simulation results.

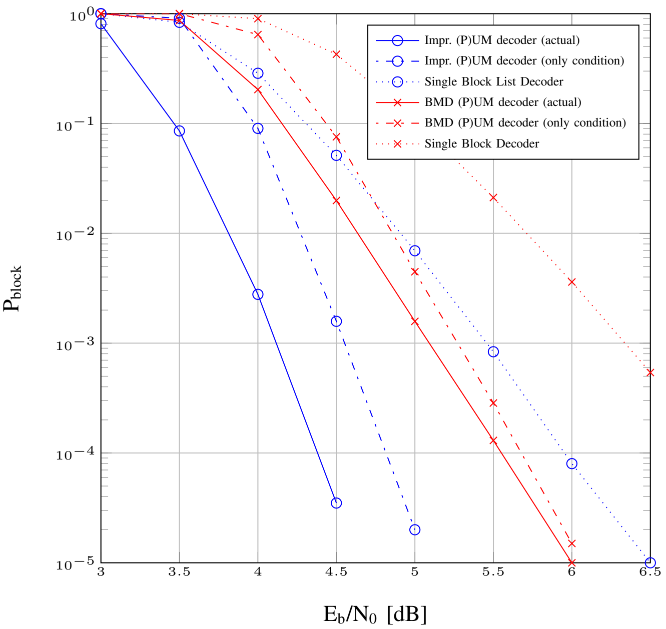


Fig. 1. Error probability simulation result of a $(31, 11|6)$ PUM code

The decoding condition as well as the simulation results show a significant improvement compared to the original algorithm.

A possible further modification of the algorithm could use the *Kötter–Vardy* soft-decision list decoding algorithm [10] instead of the *Guruswami–Sudan* algorithm as block decoders.

Acknowledgment

The authors thank Vladimir Sidorenko and Alexander Zeh for the valuable discussions.

References

[1] L.-N. Lee, “Short Unit-Memory Byte-Oriented Binary Convolutional Codes Having Maximal Free Distance,” *IEEE Transactions on Information Theory*, pp. 349–352, May 1976.

[2] G. S. Lauer, “Some Optimal Partial-Unit Memory Codes,” *IEEE Transactions on Information Theory*, vol. 23, no. 2, pp. 240–243, Mar. 1979.

[3] U. Dettmar and U. K. Sorger, “Bounded Minimum Distance Decoding of Unit Memory Codes,” *IEEE Transactions on Information Theory*, vol. 41, no. 2, pp. 591–596, 1995.

[4] A. J. Viterbi, “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,” *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 260–269, April 1967.

[5] S. Puchinger, A. Wachter-Zeh, and M. Bossert, “Improved Decoding of Partial Unit Memory Codes Using List Decoding of Reed-Solomon Codes,” 2013. [Online]. Available: http://nt.uni-ulm.de/~puchinger/papers/pum_list_decoding_long.pdf

[6] V. Guruswami and M. Sudan, “Improved Decoding of Reed-Solomon and Algebraic-Geometry Codes,” *IEEE Transactions on Information Theory*, no. 6, pp. 1757–1767, Sept.

[7] J. Justesen and T. Høholdt, *A Course in Error-Correcting Codes*. Zürich: European Mathematical Soc., 2004.

[8] R. J. Mceliece, “On the Average List Size for the Guruswami-Sudan Decoder,” in *International School on Coding Theory and Applications*, 2003.

[9] U. Dettmar, “Partial Unit Memory Codes,” Ph.D. dissertation, University of Darmstadt, June 1994.

[10] R. Kötter and A. Vardy, “Algebraic Soft-Decision Decoding of Reed-Solomon Codes,” *IEEE Transactions on Information Theory*, no. 11, pp. 2809–2825.

Cyclic Codes over the Matrix Ring $M_2(\mathbb{F}_p)$ and Their Isometric Images over $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$

Dixie F. Falcunit, Jr. and Virgilio P. Sison
Institute of Mathematical Sciences and Physics
University of the Philippines, Los Baños
College, Laguna 4031, Philippines
Email: {dffalcunitjr, vpsison}@uplb.edu.ph

***Abstract*—Let $\mathbb{F}_p$ be the prime field with p elements. We derive the homogeneous weight on the Frobenius matrix ring $M_2(\mathbb{F}_p)$ in terms of the generating character. We also give a generalization of the Lee weight on the finite chain ring $\mathbb{F}_{p^2}+u\mathbb{F}_{p^2}$ where $u^2 = 0$. A non-commutative ring, denoted by $\mathcal{F}_{p^2} + \mathrm{v}_p\mathcal{F}_{p^2}$, v_p an involution in $M_2(\mathbb{F}_p)$, that is isomorphic to $M_2(\mathbb{F}_p)$ and is a left $\mathbb{F}_{p^2}$-vector space, is constructed through a unital embedding τ from $\mathbb{F}_{p^2}$ to $M_2(\mathbb{F}_p)$. The elements of $\mathcal{F}_{p^2}$ come from $M_2(\mathbb{F}_p)$ such that $\tau(\mathbb{F}_{p^2}) = \mathcal{F}_{p^2}$. The irreducible polynomial $f(x) = x^2 + x + (p-1) \in \mathbb{F}_p[x]$ required in τ restricts our study of cyclic codes over $M_2(\mathbb{F}_p)$ endowed with the Bachoc weight to the case $p \equiv 2$ or 3 mod 5. The images of these codes via a left $\mathbb{F}_p$-module isometry are additive cyclic codes over $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$ endowed with the Lee weight. New examples of such codes are given.**

***Index Terms*—Frobenius matrix ring, finite chain ring, homogeneous weight, cyclic codes.**

I. Introduction

The theory of codes over finite rings has gained much attention since the significant result in [8] showed that several well-known families of good nonlinear binary codes can be identified as Gray images of linear codes over the quaternary ring $\mathbb{Z}_4$ of integers modulo 4. Several recent papers dealt with codes over finite Frobenius rings. These rings are considered the most appropriate coding alphabet since the two classical theorems, namely the extension theorem and the MacWilliams identities, generalize neatly in the case of finite Frobenius rings.

Let p be a prime and $r \geq 1$ an integer. We denote by $\mathbb{F}_{p^r}$ the Galois field of order p^r and characteristic p. In this study we restrict ourselves to a small class of finite Frobenius rings, the matrix rings over a finite field, in particular the ring of 2×2 matrices over $\mathbb{F}_p$, denoted by $M_2(\mathbb{F}_p)$. The multiplicative group $GL(2,p)$ of invertible matrices in $M_2(\mathbb{F}_p)$ will be of much use in the ensuing discussion as well. Until now very few publications on codes over non-commutative rings have been seen. It was only in 2012 that the theory of cyclic codes over $M_2(\mathbb{F}_2)$ was developed [1]. The idea for the construction of cyclic codes over $M_2(\mathbb{F}_2)$ came from [2] in which was defined an isometric map ϕ from $\mathbb{F}_4^2$ onto $M_2(\mathbb{F}_2)$ where

$$\phi((a + b\omega, c + d\omega)) = \begin{pmatrix} a+d & b+c \\ b+c+d & a+b+d \end{pmatrix}$$

using the usual Hamming weight $w_{\mathtt{Ham}}$ on $\mathbb{F}_4$ extended component-wise, and the Bachoc weight $w_{\mathtt{B}}$ on $M_2(\mathbb{F}_2)$ such that $w_{\mathtt{Ham}}(\alpha) = w_{\mathtt{B}}(\phi(\alpha))$ for all α in $\mathbb{F}_4^2$. Here ω is a root of the monic irreducible polynomial $x^2+x+1 \in \mathbb{F}_2[x]$ such that $\mathbb{F}_4$ is seen as an extension of $\mathbb{F}_2$ by ω. The Bachoc weight on $M_2(\mathbb{F}_p)$ as given in [2] is defined as follows.

$$w_{\mathtt{B}}(A) = \begin{cases} 0 & \text{if } A = \mathbf{0} \\ 1 & \text{if } A \in GL(2,p) \\ p & otherwise \end{cases}$$

The study of codes over $\mathbb{Z}_4$ and $M_2(\mathbb{F}_2)$ reveals the importance of weight functions that are different from the Hamming weight. Here we derive the homogeneous weight on $M_2(\mathbb{F}_p)$ using the formula introduced by T. Honold for arbitrary finite Frobenius rings [9]. Likewise we extend the definition of the Lee weight on $\mathbb{F}_2 + u\mathbb{F}_2$, $u^2 = 0$ given in [4] to the finite chain ring $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$, $u^2 = 0$. The connection between the minimal left ideals and the idempotent elements of $M_2(\mathbb{F}_p)$ is used to generalize the homogeneous weight on $M_2(\mathbb{F}_p)$. We also employ the well known representation of the field by matrices by giving a unital embedding τ from $\mathbb{F}_{p^2}$ to $M_2(\mathbb{F}_p)$ to construct a non-commutative ring that is isomorphic to $M_2(\mathbb{F}_p)$ and is a left $\mathbb{F}_{p^2}$-vector space. This ring is denoted by $\mathcal{F}_{p^2} + \mathbf{v}_p\mathcal{F}_{p^2}$ where $\mathbf{v}_p$ is an involution in $M_2(\mathbb{F}_p)$ and the elements of $\mathcal{F}_{p^2}$ come from $M_2(\mathbb{F}_p)$ such that $\tau(\mathbb{F}_{p^2}) \cong \mathcal{F}_{p^2}$. The unital embedding τ comes from a characterization of $\mathbb{F}_p$ in terms of an irreducible polynomial $f(x) = x^2 + x + (p-1) \in \mathbb{F}_p[x]$. The property of this polynomial restricts our study to the case where $p \equiv 2$ or 3 mod 5. As a consequence certain structural properties of cyclic codes over $M_2(\mathbb{F}_p)$ that are similar to those of cyclic codes over $M_2(\mathbb{F}_2)$ are derived. The structure theorems used the transformation of the non-commutative ring $\mathcal{F}_{p^2} + \mathbf{v}_p\mathcal{F}_{p^2}$ to $\mathcal{F}_{p^2} + \mathbf{u}_p\mathcal{F}_{p^2}$ by introducing a matrix $\mathbf{i}_p \in M_2(\mathbb{F}_p)$ such that $\mathbf{u}_p = \mathbf{i}_p + \mathbf{v}_p$, where $\mathbf{u}_p^2$ is the zero matrix. Also we define a left $\mathbb{F}_p$-module isometry from $M_2(\mathbb{F}_p)$ to $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$ using their respective Bachoc weight and Lee weight.

II. Homogeneous Weight on $M_2(\mathbb{F}_p)$

Let R be a finite ring and $\mathbb{R}$ the set of real numbers. A weight function $w\colon R \longrightarrow \mathbb{R}$ is called *left homogeneous* provided $w(0) = 0$ and the following hold:

(H1) If $Rx = Ry$ for $x, y \in R$, then $w(x) = w(y)$.

(H2) There exists $\Gamma > 0$ such that for every nonzero $x \in R$ there holds

$$\sum_{y \in Rx} w(y) = \Gamma |Rx|.$$

The definition for a right homogeneous weight follows analogously, and we say that w is *homogeneous* if it is both left homogeneous and right homogeneous. The number Γ is called the *average value* of w. The weight w is said to be *normalized* if $\Gamma = 1$. It is well known that the normalized homogeneous weight on $\mathbb{F}_q$, $q = p^r$, is given by

$$w_{\mathtt{nhom}}(x) = \begin{cases} 0 & \text{if } x = 0 \\ \dfrac{q}{q-1} & \text{if } x \neq 0. \end{cases}$$

This idea comes from the generalization of the homogeneous weight on a finite chain ring [7]. But our goal is to give a generalization of the homogeneous weight on $M_2(\mathbb{F}_p)$ which is not a finite chain ring but is a finite (non-commutative) Frobenius ring. We shall use the generating character instead of the Möbius inversion formula for homogeneous weight that was employed in [5].

For a finite Frobenius ring R, Honold [9] observed that every homogeneous weight on R with generating character χ must have the form

$$w : R \longrightarrow \mathbb{R}, x \mapsto \Gamma \left[1 - \frac{1}{|R^{\times}|} \sum_{u \in R^{\times}} \chi(ux) \right]$$

where $R^{\times}$ is the group of units of R. Note that every finite Frobenius ring has a generating character [12]. The generating character of $M_n(\mathbb{F}_q)$ is

$$\chi(A) = exp \left\{ \frac{2\pi i \cdot tr(Tr(A))}{p} \right\}$$

where tr is the trace map from $\mathbb{F}_q$ down to $\mathbb{F}_p$, that is, $tr(\alpha) = \alpha + \alpha^p + \cdots + \alpha^{p^{r-1}}$ for $\alpha \in \mathbb{F}_q$, and Tr is the classical trace of the matrix $A \in M_n(\mathbb{F}_q)$. The homogeneous weight on $M_n(\mathbb{F}_q)$ is given by

$$w : M_n(\mathbb{F}_q) \longrightarrow \mathbb{R}, A \mapsto \Gamma \left[1 - \frac{1}{|GL(n,q)|} \sum_{u \in GL(n,q)} \chi(uA) \right]$$

where $GL(n,q)$ is the group of nonsingular matrices in $M_n(\mathbb{F}_q)$. It is known that $|GL(n,q)| = q^{n(n-1)/2} \prod_{i=1}^{n} (q^i - 1)$ [3].

The main concern in this section is to derive the homogeneous weight on $M_2(\mathbb{F}_p)$. First we discuss the structure of $M_2(\mathbb{F}_p)$.

Remark 2.1: The matrix ring $M_n(\mathbb{F}_q)$ has no proper ideals but it has proper left ideals [10]. In particular $M_2(\mathbb{F}_p)$ has $p+1$ minimal left ideals [2]. This is essential in this section so we take it as a theorem.

Theorem 2.2: $M_2(\mathbb{F}_p)$ has $p+1$ minimal left ideals and each minimal left ideal contains p^2 elements.

Proof: Let $A \in M_2(\mathbb{F}_p)$ where $A = \begin{pmatrix} a_0 & a_1 \\ a_2 & a_3 \end{pmatrix}$. Note that $\begin{pmatrix} 1 & r \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ are nonzero nonunit idempotents of $M_2(\mathbb{F}_p)$ where $r \in \mathbb{F}_p$. Thus the proper left ideals are of the form $\begin{pmatrix} a_0 & ra_0 \\ a_2 & ra_2 \end{pmatrix}$ and $\begin{pmatrix} 0 & a_1 \\ 0 & a_3 \end{pmatrix}$. Hence there are $p+1$ minimal left ideals in $M_2(\mathbb{F}_p)$ since the intersection of any two minimal left ideals of $M_2(\mathbb{F}_p)$ is the zero matrix. It follows immediately that every minimal left ideal of $M_2(\mathbb{F}_p)$ has p^2 elements. □

In order to generalize the homogeneous weight on $M_2(\mathbb{F}_p)$ we need to get the value of the sum $\sum_{u \in GL(2,p)} \chi(uA)$ where $A \in M_2(\mathbb{F}_p)$. The case when A is the zero matrix is obvious. Theorem 2.3 below deals with the invertible matrices while Theorem 2.4 involves the zero divisors.

Theorem 2.3: $\sum_{u \in GL(2,p)} \chi(u) = \sum_{u \in GL(2,p)} \chi(uA) = p$ where $A \in GL(2,p)$.

Proof: Let D be the set of all the zero divisors in $M_2(\mathbb{F}_p)$. We have $\sum_{A \in M_2(\mathbb{F}_p)} \chi(A) = 0$ [9]. So,

$$\sum_{u \in GL(2,p)} \chi(u) = -\sum_{B \in D} \chi(B) - \chi(0) \tag{1}$$

and since $M_2(\mathbb{F}_p)$ has $p+1$ minimal left ideals, $\chi_{I_L}(A) = \chi(A)$ for all $A \in I_L$ and $\chi_{I_L}(0) = 1$ in [9], where χ_{I_L} is a character of the minimal left ideal I_L of $M_2(\mathbb{F}_p)$. Hence,

$$\begin{aligned} \sum_{u \in GL(2,p)} \chi(u) &= -(p+1) \sum_{B \in I_L \backslash \{0\}} \chi_{I_L}(B) - 1 \qquad (2) \\ &= -(p+1)(-1) - 1 \\ &= p. \end{aligned}$$

□

Theorem 2.4: $\sum_{u_k \in GL(2,p)} \chi(u_k B) = p - p^2$ for all $B \in I_L \backslash \{0\}$.

Proof: $-\sum_{u_k \in GL(2,p)} \chi(u_k B)$

$$= \left[\sum_{B_j \in I_L \backslash \{0\}} \chi(B_j) \right] \left[\sum_{u_k \in GL(2,p)} \chi(u_k B) \right] \tag{3}$$

$$= \sum_{B_j \in I_L \backslash \{0\}} \sum_{u_k \in GL(2,p)} \chi(u_k B) \chi(B_j) \tag{4}$$

$$= \sum_{B_j \in I_L \backslash \{0\}} \sum_{u_k \in GL(2,p)} \chi(u_k B + B_j) \tag{5}$$

$$= \sum_{u_k \in GL(2,p)} \sum_{B_j \in I_L \backslash \{0\}} \chi(u_k B + B_j) \tag{6}$$

For each $u_r \in GL(2,p)$, there exists $B_s \in I_L \backslash \{0\}$ such that $u_r B + B_s = 0$ (Note: B_s is not unique for every u_r). So,
$-\sum_{u_k \in GL(2,p)} \chi(u_k B)$

$$= \sum_{u_r \in GL(2,p)} \chi(u_r B + B_s) \tag{7}$$

$$+ \sum_{u_k \in GL(2,p)} \sum_{B_j \in I_L \backslash \{0\}} \chi(u_k B + B_j) \tag{8}$$

where $u_kB + B_j \neq 0$

$$= \sum_{u_r \in GL(2,p)} \chi(0) + \sum_{u_k \in GL(2,p)} \sum_{B_j \in I_L \backslash \{0\}} \chi(u_kB + B_j) \tag{9}$$

where $u_kB + B_j \neq 0$

$$= |GL(2,p)| + \sum_{u_k \in GL(2,p)} \sum_{B_j \in I_L \backslash \{0\}} \chi(u_kB + B_j) \tag{10}$$

where $u_kB + B_j \neq 0$.

For every $B_t \in I_L \backslash \{0, B_s\}$ we have $u_rB + B_t \in I_L \backslash \{0, u_rB\}$ $(i.e. B_t + \{I_L \backslash \{0, B_s\}\} = I_L \backslash \{0, u_rB\})$ and for fixed B_t and u_r we can always find l such that $u_l \neq u_r$ and $u_rB + B_t = u_rB$. Thus, we can collect all the elements of $I_L \backslash \{0\}$. And since $|I_L \backslash \{0\}|$ divides $|GL(2,p)|$,

$-\sum_{u_k \in GL(2,p)} \chi(u_kB)$

$$= |GL(2,p)| \tag{11}$$

$$+ \frac{|GL(2,p)||I_L \backslash \{0\}| - |GL(2,p)|}{|I_L \backslash \{0\}|} \sum_{B_j \in I_L \backslash \{0\}} \chi(B_j) \tag{12}$$

$$= (p^2 - p)(p^2 - 1) + (p^2 - p)(p^2 - 2)(-1) \tag{13}$$

$$= p^2 - p. \tag{14}$$

Thus,

$$\sum_{u_k \in GL(2,p)} \chi(u_kB) = p - p^2. \tag{15}$$

□

Theorem 2.5: The homogeneous weight on $M_2(\mathbb{F}_p)$ is given by

$$w_{\mathtt{hom}}(A) = \begin{cases} 0 & \text{if } A = \mathbf{0} \\ \Gamma\left(1 - \dfrac{1}{(p^2-1)(p-1)}\right) & \text{if } A \in GL(2,p) \\ \Gamma\left(\dfrac{p^2}{p^2-1}\right) & \mathit{otherwise.} \end{cases}$$

Proof: The proof is straightforward from the two preceding theorems. □

III. Lee Weight on $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$, $u^2 = 0$

In [4] the Lee weight w_{L} of $x = (x_1, \ldots, x_n) \in (\mathbb{F}_2 + u\mathbb{F}_2)^n$ is defined as $n_1(x) + 2n_2(x)$, where $n_2(x)$ and $n_1(x)$ are, respectively, the number of u symbols and the number of 1 or $1+u$ symbols present in x. So when $n = 1$, $w_{\mathrm{L}}(0) = 0$, $w_{\mathrm{L}}(1) = w_{\mathrm{L}}(1+u) = 1$ and $w_{\mathrm{L}}(u) = 2$.

Consider the finite chain ring $\mathbb{F}_3 + u\mathbb{F}_3$, $u^2 = 0$ then we can define $w_{\mathrm{L}}(x) = n_1(x) + 3n_2(x)$, for all $x \in \mathbb{F}_3 + u\mathbb{F}_3$, where $n_2(x)$ and $n_1(x)$ are, respectively, the number of u symbols and the number of 1 or $1+u$ symbols present in x, as can be seen in Table II.

Now consider the subset $\mathcal{B}_2$ of $\mathbb{F}_4 + u\mathbb{F}_4$, $u^2 = 0$ where

$$\mathcal{B}_2 = \{(\alpha a_1 + \alpha b_1\omega) + u(\beta a_1 + \beta b_1\omega) | \alpha = 1, a_1, b_1, \beta \in \mathbb{F}_2\}.$$

TABLE I
Bachoc Weight and Normalized Homogeneous Weight on $M_2(\mathbb{F}_2)$

$M_2(\mathbb{F}_2)$	w_{B}	$w_{\mathtt{nhom}}$	$M_2(\mathbb{F}_2)$	w_{B}	$w_{\mathtt{nhom}}$
$\begin{pmatrix}0&0\\0&0\end{pmatrix}$	0	0	$\begin{pmatrix}0&1\\0&1\end{pmatrix}$	2	4/3
$\begin{pmatrix}1&0\\0&1\end{pmatrix}$	1	2/3	$\begin{pmatrix}1&1\\0&0\end{pmatrix}$	2	4/3
$\begin{pmatrix}0&1\\1&1\end{pmatrix}$	1	2/3	$\begin{pmatrix}0&0\\0&1\end{pmatrix}$	2	4/3
$\begin{pmatrix}1&1\\1&0\end{pmatrix}$	1	2/3	$\begin{pmatrix}1&0\\0&0\end{pmatrix}$	2	4/3
$\begin{pmatrix}1&0\\1&1\end{pmatrix}$	1	2/3	$\begin{pmatrix}1&1\\1&1\end{pmatrix}$	2	4/3
$\begin{pmatrix}1&1\\0&1\end{pmatrix}$	1	2/3	$\begin{pmatrix}0&0\\1&1\end{pmatrix}$	2	4/3
$\begin{pmatrix}0&1\\1&0\end{pmatrix}$	1	2/3	$\begin{pmatrix}1&0\\1&0\end{pmatrix}$	2	4/3
$\begin{pmatrix}0&0\\1&0\end{pmatrix}$	2	4/3	$\begin{pmatrix}0&1\\0&0\end{pmatrix}$	2	4/3

TABLE II
Lee weight on $\mathbb{F}_3 + u\mathbb{F}_3$, $u^2 = 0$

$\mathbb{F}_3 + u\mathbb{F}_3$	w_{L}
0	0
1	1
2	1
$1+u$	1
$2+2u = 2(1+u)$	1
u	3
$2+u$	3
$2u$	3
$1+2u$	3

Similarly we can define the Lee weight on $\mathbb{F}_4 + u\mathbb{F}_4$, $u^2 = 0$ to be $w_{\mathrm{L}}(x) = n_1 + 2n_2(x)$ where again $n_2(x)$ and $n_1(x)$ are, respectively, the number of u symbols and the number of 1 or $1+u$ symbols present in x.

$$w_{\mathrm{L}}(x) = \begin{cases} 0 & \text{if } x = \mathbf{0} \\ 1 & \text{if } A \in \mathcal{B}_2 \backslash \{0\} \\ 2 & \mathit{otherwise} \end{cases}$$

This can also be seen in Table III. In general we can define the Lee weight on $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$, $u^2 = 0$ as $w_{\mathrm{L}}(x) = n_1 + pn_2(x)$.

$$w_{\mathrm{L}}(x) = \begin{cases} 0 & \text{if } x = \mathbf{0} \\ 1 & \text{if } A \in \mathcal{B}_p \backslash \{0\} \\ p & \mathit{otherwise} \end{cases}$$

where

$$\mathcal{B}_p = \{(\alpha a_1 + \alpha b_1\omega) + u(\beta a_1 + \beta b_1\omega) | \alpha \in \mathbb{F}_p^{\times}, a_1, b_1, \beta \in \mathbb{F}_p\}.$$

IV. $\mathbb{F}_{p^2}$-Linear Map

In this section we give the conditions on the finite field $\mathbb{F}_p$ for the polynomial $f(x) = x^2 + x + (p-1)$ to be irreducible over $\mathbb{F}_p$. Using the well known representation of fields by matrices, Theorem 4.2 shows the corresponding cyclic algebra that is isomorphic to $M_2(\mathbb{F}_p)$ and is a left $\mathbb{F}_{p^2}$-vector space.

TABLE III
LEE WEIGHT ON $\mathbb{F}_4 + u\mathbb{F}_4$, $u^2 = 0$

$\mathbb{F}_4 + u\mathbb{F}_4$	w_{L}
0	0
1	1
ω	1
$1+\omega$	1
$1+u$	1
$\omega + u\omega = \omega(1+u)$	1
$(1+\omega) + u(1+\omega) = (1+\omega)(1+u)$	1
u	2
$\omega + u$	2
$(1+\omega) + u$	2
$u\omega$	2
$1 + u\omega$	2
$(1+\omega) + u\omega$	2
$u(1+\omega)$	2
$1 + u(1+\omega)$	2
$\omega + u(1+\omega)$	2

Lemma 4.1: Let $p \equiv 2$ or 3 (mod 5) then the polynomial $f(x) = x^2 + x + (p-1)$ is irreducible over $\mathbb{F}_p$.

Proof: The case when $p = 2$ is trivial. Note that the discriminant of the polynomial $f(x)$ is equal to $5 \in \mathbb{F}_p$. Then $f(x)$ is reducible over $\mathbb{F}_p$ if there exists $y \in \mathbb{F}_p$ such that $y^2 \equiv 5$ (mod p). By the Law of Quadratic Reciprocity of elementary number theory, when p is odd, $y^2 \equiv 5$ (mod p) is solvable if and only if $p \equiv 1$ or -1 (mod 5). □

Theorem 4.2: Let $f(x) = \sum_{i=0}^{n} a_i x^i \in \mathbb{F}_q[x]$ be a monic irreducible polynomial. Then the mapping $\pi\colon \mathbb{F}_q[x] \to M_n(\mathbb{F}_q)$, $g(x) \mapsto g(X)$ induces a unital embedding of $\mathbb{F}_q[x]/(f)$ into $M_n(\mathbb{F}_q)$ where

$$X = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}.$$

Remark 4.3: The matrix X is known as the *companion matrix.*

Corollary 4.4: Let $\mathbb{F}_{p^2} = \mathbb{F}_p[\omega]$ where $\omega^2 + \omega + (p-1) = 0$ then $\tau\colon \mathbb{F}_{p^2} \longrightarrow M_2(\mathbb{F}_p)$ defined by

$$a + b\omega \mapsto \begin{pmatrix} a & b \\ b & a + (p-1)b \end{pmatrix}$$

is an embedding.

Proof: The proof follows immediately from Lemma 4.1 and Theorem 4.2. □

Theorem 4.5: If ω is a root of $f(x) = x^2 + x + (p-1)$ then $\omega^p \equiv (p-1)\omega + (p-1)(\mathrm{mod}(\omega^2 + \omega + (p-1))$.

Proof: First we show that $(p-1)\omega + (p-1)$ is also a root of $f(x)$, that is,

$$\begin{aligned} & f[(p-1)\omega + (p-1)] \\ &= [(p-1)\omega + (p-1)]^2 + [(p-1)\omega + (p-1)] + (p-1) \\ &= [(p-1)^2\omega^2 + 2\omega + 1] + [(p-1)\omega + (p-1)] + (p-1) \\ &= \omega^2 + 2\omega + 1 - \omega - 2 \\ &= \omega^2 + \omega + (p-1) \\ &= 0. \end{aligned}$$

Now, let $h(x) = x^p$. By the Division Algorithm, there exist $g(x)$ and $r_1x + r_2$ such that $h(x) = g(x)f(x) + r_1x + r_2$ where $r_1x + r_2$ is the remainder when $h(x)$ is divided by $f(x)$. Since ω and $(p-1)\omega + (p-1)$ are roots of $f(x)$ then we have $\omega^p = r_1\omega + r_2$ and

$$[(p-1)\omega + (p-1)]^p = r_1[(p-1)\omega + (p-1)] + r_2$$

or equivalently,

$$(p-1)\omega^p + (p-1) = r_1(p-1)\omega + r_1(p-1) + r_2.$$

Since the characteristic of $\mathbb{F}_p$ is p, then

$$[(p-1)\omega + (p-1)]^p = [(p-1)\omega]^p + (p-1)^p = [(p-1)^p\omega^p] + (p-1)^p.$$

By Fermat's Little Theorem,

$$[(p-1)^p\omega^p] + (p-1)^p = (p-1)\omega^p + (p-1).$$

Adding equations $\omega^p = r_1\omega + r_2$ and $(p-1)\omega^p + (p-1) = r_1(p-1)\omega + r_1(p-1) + r_2$ modulo p, the resulting equation is $(p-1) = r_1(p-1) + 2r_2$ or simply $r_1 + (p-2)r_2 = 1$.

Note that $gcd(1, p-2) = 1$. And we have $1 = (p-1) - (p-2) = (p-1) + (p-2)(p-1)$. So, $r_1 = p-1$ and $r_2 = p-1$. Thus, $\omega^p \equiv (p-1)\omega + (p-1)(\mathrm{mod}(\omega^2 + \omega + (p-1))$. □

Theorem 4.6: $\tau^p(\omega) = \begin{pmatrix} p-1 & p-1 \\ p-1 & 0 \end{pmatrix}$.

Proof: Since τ is a homomorphism we have

$$\begin{aligned} \tau(\omega^p) &= \tau(\underbrace{\omega\omega\cdots\omega}_{p\quad \omega's}) \\ &= \underbrace{\tau(\omega)\tau(\omega)\cdots\tau(\omega)}_{p\quad \tau(\omega)'s} \\ &= \tau^p(\omega). \end{aligned}$$

$$\begin{aligned} \tau^p(\omega) &= \tau(\omega^p) \\ &= \tau[(p-1)\omega + (p-1)] \\ &= \tau[(p-1)\omega] + \tau(p-1) \\ &= \tau(p-1)\tau(\omega) + \tau(p-1) \\ &= \begin{pmatrix} p-1 & 0 \\ 0 & p-1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & p-1 \end{pmatrix} + \begin{pmatrix} p-1 & 0 \\ 0 & p-1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & p-1 \\ p-1 & 1 \end{pmatrix} + \begin{pmatrix} p-1 & 0 \\ 0 & p-1 \end{pmatrix} \\ &= \begin{pmatrix} p-1 & p-1 \\ p-1 & 0 \end{pmatrix}. \end{aligned}$$

□

Theorem 4.7: Let $\mathcal{F}_p$ be the set of all scalar matrices in $M_2(\mathbb{F}_p)$, $p \equiv 2$ or $3 \bmod 5$, $\tau(\mathbb{F}_{p^2}) = \mathcal{F}_{p^2}$ and $\mathbf{v}_p = \begin{pmatrix} 1 & 0 \\ p-1 & p-1 \end{pmatrix}$. Then $\mathbf{v}_p\tau(\omega) = \tau^p(\omega)\mathbf{v}_p$, $\mathcal{F}_p[\tau(\omega)] = \mathcal{F}_{p^2}$ and $M_2(\mathbb{F}_p) = \mathcal{F}_{p^2} + \mathbf{v}_p\mathcal{F}_{p^2}$.

Proof: It is easy to show that $\mathbf{v}_p\tau(\omega) = \tau^p(\omega)\mathbf{v}_p$ and $\mathcal{F}_p[\tau(\omega)] = \mathcal{F}_{p^2}$ since $\tau^2(\omega) + \tau(\omega) + \tau(p-1) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

$$\begin{aligned} M_2(\mathbb{F}_p) &= \mathcal{F}_{p^2} + \mathbf{v}_p\mathcal{F}_{p^2} \\ &= \left\{ \begin{pmatrix} a+c & b+d \\ b-c-d & a-b-c) \end{pmatrix} |a,b,c,d \in \mathbb{F}_p \right\}. \end{aligned}$$

□

V. Cyclic Codes over $M_2(\mathbb{F}_p)$

Structure theorems for cyclic codes over $\mathcal{A}_2 = M_2(\mathbb{F}_2)$ were established in [1] by introducing two matrices $\tau(\omega)$ and v in $\mathcal{A}_2$ satisfying the relation $\mathrm{v}\tau(\omega) = \tau^2(\omega)\mathrm{v}$. A possible choice would be those given by Bachoc [2] which are

$$\mathrm{v} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } \tau(\omega) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

such that $\mathcal{A}_2 = \mathcal{F}_2[\tau(\omega)] + \mathrm{v}\mathcal{F}_2[\tau(\omega)]$ where $\mathcal{F}_2[\tau(\omega)] = \mathcal{F}_4 \cong \mathbb{F}_4$ and $\mathrm{v} \neq \mathbf{v}_2$. Setting $\mathrm{u} = \tau(1) + \mathrm{v}$ gives $\mathrm{u}^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $A_2 = \mathcal{F}_4 + \mathrm{u}\mathcal{F}_4$. Alamadhi et.al. [1] used the ring $\mathcal{F}_4 + \mathrm{u}\mathcal{F}_4$ to develop structure theorems for cyclic codes over $M_2(\mathbb{F}_2)$ by simply extending from cyclic codes over $\mathbb{F}_2 + u\mathbb{F}_2$, $u^2 = 0$ [4].

It seems that a construction of cyclic codes over $\mathbb{F}_p + u\mathbb{F}_p$, $u^2 = 0$, will result in the construction of cyclic codes over $\mathcal{A}_p = M_2(\mathbb{F}_p)$. Fortunately, Qian, Zhang and Zhu [11] solved an open-ended question given in [4], that is, to extend the cyclic codes over $\mathbb{F}_2 + u\mathbb{F}_2$, $u^2 = 0$ to $\mathbb{F}_p + u\mathbb{F}_p + \cdots + u^{k-1}\mathbb{F}_p$, $u^k = 0$. Thus, the case when $k = 2$ gives the cyclic codes over $\mathbb{F}_p + u\mathbb{F}_p$, $u^2 = 0$.

$$\begin{tikzcd} \mathbb{F}_{p^2} + u\mathbb{F}_{p^2} \arrow[d, dash] \\ u\mathbb{F}_{p^2} \arrow[d, dash] \\ (0) \end{tikzcd}$$

Fig. 1. Lattice of ideals of $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}, u^2 = 0$

Let $p \equiv 2$ or 3 (mod 5), $\mathbf{i}_p = \begin{pmatrix} p-1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{u}_p = \mathbf{v}_p + \mathbf{i}_p$. Then $\mathbf{u}_p^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and

$$\begin{aligned} \mathcal{A}_p &= \mathcal{F}_{p^2} + \mathbf{u}_p\mathcal{F}_{p^2} \\ &= \left\{ \begin{pmatrix} a & b \\ b-c & a-b-d \end{pmatrix} |a,b,c,d \in \mathbb{F}_p \right\}. \end{aligned}$$

Let $\mathcal{A}_p[X]$ be the ring of polynomials over $\mathcal{A}_p$. We have a natural homomorphic mapping from $\mathcal{A}_p$ to its field $\mathcal{F}_{p^2}$. For any $a \in \mathcal{A}_p$, let $\hat{a}$ denote the polynomial reduction modulo $\mathbf{u}_p$.

$$\begin{tikzcd} \mathcal{F}_{p^2} + \mathbf{u}_p\mathcal{F}_{p^2} \arrow[d, dash] \\ (0) \end{tikzcd}$$

Fig. 2. Lattice of ideals of $\mathcal{F}_{p^2} + \mathbf{u}_p\mathcal{F}_{p^2}$,$\mathbf{u}_p^2$ is the zero matrix

Now define a polynomial reduction mapping $\mu \colon \mathcal{A}_p[X] \longrightarrow \mathcal{F}_{p^2}[X]$ such that

$$f(X) = \sum_{i=0}^{r} a_i X^j \mapsto \sum_{i=0}^{r} \hat{a}_i X^j.$$

A monic polynomial f over $\mathcal{A}_p[X]$ is said to be a basic irreducible polynomial if its projection $\mu(f)$ is irreducible over $\mathcal{F}_{p^2}[X]$. An $\mathcal{A}_p$-linear code C of length n is an $\mathcal{A}_p$-submodule of $\mathcal{A}_p^n$. As left modules we have the expansion $\mathcal{R}_{p,n} = \mathcal{A}_p[x]/(x^n - 1) = \oplus_{j=1}^{t}\mathcal{A}_{p,j}$, where the $\mathcal{A}_{p,j} = \mathcal{A}_p[x]/(f_j)$ are quotient $\mathcal{A}_p$-modules and $x^n - 1 = \prod_{j=1}^{t} f_j$ where f_j's are irreducible polynomials over $\mathcal{F}_{p^2}$.

We shall prove the lemma and the theorem below using the same techniques in [1] and [11] given the condition that p is not divisible by n.

Lemma 5.1: If f is an irreducible polynomial over $\mathcal{F}_{p^2}$ the only left $\mathcal{A}$-modules of $\mathcal{R}_p(f) = \mathcal{A}_p[X]/(f)$ are $(\tau(0))$, $(\mathbf{u}_p)$ and $(\tau(1))$. In particular this quotient ring is a non-commutative chain ring.

Proof: Let $I \neq (\tau(0))$ be an ideal of $\mathcal{R}_p(f)$. Pick g in $\mathcal{A}_p[X]$ such that $g + (f) \in I$, but $g \notin (f)$. Because f is irreducible the *gcd* of μg and f can only take two values, $\tau(1)$ and f. In the first case g is invertible mod f and $I = (\tau(1)) = \mathcal{R}_p(f)$. If this does not happen, $I \subseteq \mathbf{u}_p + (f)$. To show the reverse inclusion, let $g = \mathbf{u}_p r$ with $\mathbf{u}_p r + (f) \subseteq I$ and $\mathbf{u}_p r + (f) \neq \tau(0)$. We can assume by the latter condition that $\mu r \notin (f)$. Hence by the irreducibility of f we have that $gcd(\mu r, f) = \tau(1)$. This entails the existence of $a, b, c \in \mathcal{A}_p[X]$ such that $ra + fb = \tau(1) + \mathbf{u}_p c$. Multiplying both sides by $\mathbf{u}_p$ we get $\mathbf{u}_p ra = \mathbf{u}_p + \mathbf{u}_p fb$. The left hand side is in I, a right sided ideal. Thus the reverse inclusion follows. □

Theorem 5.2: Suppose C is a cyclic code of length n over $\mathcal{A}_p = \mathcal{F}_{p^2} + \mathbf{u}_p\mathcal{F}_{p^2}$ where p is not divisible by n. Then there are unique monic polynomials F_0, F_1, F_2 such that $C = \langle \hat{F}_1, \mathbf{u}_p\hat{F}_2 \rangle$, where $F_0F_1F_2 = X^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\hat{F}_1 = F_0F_2$, $\hat{F}_2 = F_0F_1$, and $|C| = p^{2s}$ where $s = 2\deg F_1 + \deg F_2$.

Proof: Let $X^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = f_1 f_2 \dots f_r$ be the unique factorization of $X^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ into a product of monic basic irreducible pairwise coprime polynomials. Note that C is a direct sum of right $\mathcal{A}_p$-modules of the form $(\mathbf{u}_p^j\hat{f}_i)$, $0 \leq j \leq 1$, $0 \leq i \leq r$ where $\hat{f}_i = \prod_{j=1,j\neq i}^{n} f_j$. After reordering, we can assume that C is a direct sum of any of the following

$$(\hat{f}_{t_1+1}), (\hat{f}_{t_1+2}), \dots, (\hat{f}_{t_1+t_2}), (\mathbf{u}_p\hat{f}_{t_1+t_2+1}), \dots, (\mathbf{u}_p\hat{f}_r).$$

That is,

$$C = \langle f_1 f_2 f_3 \dots f_{t_1} f_{t_1+t_2+1} \dots f_r, \mathbf{u}_p f_1 f_2 f_3 \\ \dots f_{t_1+t_2} f_{t_1+t_2+t_3} \rangle.$$

Let

$$\hat{F}_1 = f_1 f_2 f_3 \dots f_{t_1} f_{t_1+t_2+1} \dots f_r,$$
$$\hat{F}_2 = f_1 f_2 f_3 \dots f_{t_1+t_2} f_{t_1+t_2+t_3}.$$

where $t_1, t_2 \geq 0$ and $t_1 + t_2 + 1 \leq r$.

Then

$$F_i = \begin{cases} 1 & t_{i+1} = 0 \\ f_{t_0+t_1+\dots+t_i+1} \dots f_{t_0+t_1+\dots+t_{i+1}} & t_{i+1} \neq 0, \end{cases}$$

where $t_0 = 0$, $0 \leq i \leq 2$.

Then by our construction, it is clear that $C = \langle \hat{F}_1, \mathbf{u}_p \hat{F}_2 \rangle$ and $X^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = F_0 F_1 F_2 = f_1 f_2 \dots f_r$.

To prove uniqueness, we assume that G_0, G_1, G_2 are pairwise coprime monic polynomials in $\mathcal{A}_p[X]$ such that $G_0 G_1 G_2 = X^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $C = \langle \hat{G}_1, \mathbf{u}_p \hat{G}_2 \rangle$. Thus, $C = (\hat{G}_1) + (\mathbf{u}_p \hat{G}_2)$. Now there exist nonnegative integers $m_0 = 0, m_1, \dots, m_{d+1}$ with $m_0 + m_1 + \dots + m_{d+1} = r$, and a permutation of $\{f_1, f_2, \dots, f_r\}$ such that $G_i = f_{m_0+\dots+m_i+1} \dots f_{m_0+\dots+m_{i+1}}$ for $i = 0, 1, 2$. Hence,

$$C = (\hat{f}_{m_1+1}) \oplus \dots \oplus (\hat{f}_{m_1+m_2}) \oplus (\mathbf{u}_p \hat{f}_{m_1+m_2+1}) \dots \oplus (\mathbf{u}_p \hat{f}_r).$$

It follows that $m_i = t_i$ for $i = 0, 1, 2$. Furthermore, $(f_{m_0+\dots+m_d+1}, \dots, f_{m_0+\dots+m_{d+1}})$ is a permutation of $\{f_{t_0+\dots+t_d+1}, \dots, f_{t_0+\dots+t_{d+1}}\}$. Therefore, $F_i = G_i$ for $i = 0, 1, 2$. To calculate the order of C, note that

$$C = \langle \hat{F}_1, \mathbf{u}_p \hat{F}_2 \rangle = (\hat{F}_1) \oplus (\mathbf{u}_p \hat{F}_2).$$

Hence, $|C| = (p^2)^{2(n - deg\hat{F}_1)} (p^2)^{n - deg\hat{F}_2} = p^{2s}$. □

VI. Left $\mathbb{F}_p$-module isometry

Recall that $M_2(\mathbb{F}_p) = \mathcal{F}_{p^2} + \mathbf{u}_p \mathcal{F}_{p^2}$ and $B_p = \{(\alpha a_1 + \alpha b_1 \omega) + u(\beta a_1 + \beta b_1 \omega) | \alpha \in \mathbb{F}_p^\times, a_1, b_1, \beta \in \mathbb{F}_p\}$ is a subset of $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$ Consider the mapping Φ_p defined as

$$\Phi_p : M_2(\mathbb{F}_p) \longrightarrow \mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$$

where

$$\Phi_p \left[\begin{pmatrix} a & b \\ b - c & a - b - d) \end{pmatrix} \right] = (a + b\omega) + u(c + d\omega).$$

It is easy to show that Φ_p is a left $\mathbb{F}_p$-module isomorphism. Now let $\mathcal{D}_p$ be the set of matrices in $M_2(\mathbb{F}_p)$ with entries $a = \alpha a_1$, $b = \alpha b_1$, $c = \beta a_1$ and $d = \beta b_1$ where $\alpha \in \mathbb{F}_p^\times$ and $a_1, b_1, \beta \in \mathbb{F}_p$ then $\Phi_p^{-1}(\mathcal{B}_p \backslash \{0\}) = \mathcal{D}_p \backslash \{\mathbf{0}\} = GL(2, p)$. Therefore, Φ_p is a left $\mathbb{F}_p$- module isometry such that $w_{\mathtt{B}}(A) = w_{\mathtt{L}}(\Phi_p(A))$ for all $A \in M_2(\mathbb{F}_p)$. Thus, if C is a cyclic code over $M_2(\mathbb{F}_p)$ with minimum Bachoc distance $d_{\mathtt{B}}(C)$, the image $\Phi_p(C)$ is an additive cyclic code over $\mathbb{F}_{p^2} + u\mathbb{F}_{p^2}$, $u^2 = 0$ with minimum Lee distance $d_{\mathtt{L}}(\Phi_p(C)) = d_{\mathtt{B}}(C)$.

VII. Examples

For the following examples, MAGMA routines were created to construct cyclic codes over $M_2(\mathbb{F}_p)$ and their isometric images.

Example 7.1: Let $p = 2$ and $n = 3$. Then $x^3 - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = F_0 F_1 F_2$ where $F_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $F_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ and $F_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. Then $C_1 = \langle \hat{F}_1, \mathbf{u}_2 \hat{F}_2 \rangle$ is cyclic code of length 3 with $|C_1| = 2^6 = 64$, minimum normalized homogeneous distance $d_{\mathtt{nhom}} = 2$, minimum Bachoc distance $d_{\mathtt{B}} = 3$ and minimum Hamming distance $d_{\mathtt{Ham}} = 2$. The image $\Phi_2(C_1)$ is an additive cyclic code over $\mathbb{F}_4 + u\mathbb{F}_4$ of length 3, order 64, and minimum Lee distance $d_{\mathtt{L}} = 3$.

Example 7.2: Let $p = 3$ and $n = 4$. Then $x^4 - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = f_1 f_2 f_3 f_4$ where $f_1 = x - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $f_2 = x + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $f_3 = x + \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ and $f_4 = x + \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$. If we let $F_0 = f_2 f_4$, $F_1 = f_3$ and $F_2 = f_1$ then $C_2 = \langle \hat{F}_1, \mathbf{u}_3 \hat{F}_2 \rangle$ is a cyclic code of length 4 of order $|C_2| = 9^3 = 729$ with minimum normalized homogeneous distance $d_{\mathtt{nhom}} = 27/8$, minimum Bachoc distance $d_{\mathtt{B}} = 4$ and minimum Hamming distance $d_{\mathtt{Ham}} = 3$. The image $\Phi_3(C_2)$ is an additive cyclic code over $\mathbb{F}_9 + u\mathbb{F}_9$ with length 4, cardinality 729 and minimum Lee distance $d_{\mathtt{L}} = 4$.

References

[1] A. Alamadhi, H. Sboui, P. Solé and O. Yemen, “Cyclic Codes over $M_2(\mathbb{F}_2)$,” *available on arXiv:1201.6533v1*, 2012.

[2] C. Bachoc, “Application of coding theory to the construction of modular lattices,” *J. Combinatorial Theory*, vol. 78, pp. 92-119, 1997.

[3] C. Biggs and I. White, *Permutation groups and combinatorial structure*, London: Cambridge University Press, 1979.

[4] A. Bonnecaze and P. Udaya, “Cyclic codes and self-dual codes over $\mathbb{F}_2 + u\mathbb{F}_2$,” *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1250-1255,1999.

[5] M. Greferath and S.E. Schmidt, “Linear Codes and Rings of Matrices,” *Proceedings of AAECC 13 Hawaii Springer LNCS 1719*, pp. 160-169, 1999.

[6] M. Greferath and S.E. Schmidt, ” Finite-Ring Combinatorics and MacWilliams’ Equivalence Theorem,” *Journal of Combinatorial Theory, Series A*, vol. 92, pp. 17-28, 2000.

[7] M. Greferath and S.E. Schmidt, “Gray isometries for finite chain rings and a nonlinear ternary $(36, 3^{12}, 15)$ code,” *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2522–2524, November 2001.

[8] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, “The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals and related codes,” *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 301 - 319, January 1994.

[9] T. Honold, “ A characterization of finite Frobenius rings,” *Arch. Math. (Basel)*, vol. 76, pp. 406-415, 2001.

[10] T. Hungerford, *Algebra (Graduate Texts in Mathematics* **73***)*. New York: Springer-Verlag, 1974.

[11] J. Qian, L. Zhang and S. Zhu, “Cyclic codes over $\mathbb{F}_p + u\mathbb{F}_p + \dots + u^{k-1}\mathbb{F}_p$,” *IEICE Trans. Fundamentals*, vol. E88-A, no. 3, pp. 795-797, March 2005.

[12] J. Wood, “Duality for modules over finite rings and applications to coding theory,” *Amer. J. Math*, vol. 121, pp. 555-575 ,1999.

Databases for Biometric Identification

Frans M.J. Willems
Eindhoven University of Technology
Department of Electrical Engineering
Eindhoven, The Netherlands
Email: f.m.j.willems@tue.nl

ABSTRACT

In this lecture we consider databases that are used in biometric identification settings. We assume that the biometric sequences describing the individuals consist of independent and identically distributed variables. It is our objective to find the fundamental limits that characterize such systems.

First we consider an unprotected database. We determine the so-called identification capacity, i.e., the maximum rate of individuals that makes reliable identification of the individual based on a noisy observation of the corresponding enrolled sequence possible [1],[2].

Next we focus on search complexity. Since the database contains randomly generated enrollment sequences, exhaustive search procedures seem to be required to achieve the identification capacity. To find out whether smaller search complexities can be achieved, we investigate a clustering approach, in which a first decoder determines a clustering index and a second decoder does the identification, based on this index. The first decoder is unaware of the enrolled sequences, the second one has access to these sequences. For this setting we determine the fundamental limits. These limits give us an idea about the trade-off between search and memory complexity. Although the first encoder is ignorant of the enrolled sequences, it could use structured methods to form an index [6]. We also discuss a more advanced setting in which a first decoder sends a list of indices to a second one [12].

In the second part of the lecture we discuss protected databases, that are used for identification as well as for authentication. Each individual, by enrolling in the database, obtains a secret. Moreover helper data is stored in the database for each individual. During identification in addition to the identity index, this secret has to be reconstructed, from the noisy observation of the enrolled sequence and the helper data [9],[11]. We assume that the database does not leak information about the secrets and consider also the so-called privacy leakage, i.e. the information that the database contains about the biometric enrollment sequences [7],[10]. We investigate the fundamental limits for this setting and discuss the connections to earlier work, e.g. that of Westover and O'Sullivan [4] and Tuncel [5].

Collaborators: Tanya Ignatenko, Eindhoven University of Technology, and Farzad Farhadzadeh, University Geneva.

REFERENCES

[1] J. A. O'Sullivan and N.A. Schmid, "Large deviations performance analysis for biometrics recognition." *Proc. 40th Annual Allerton Conference on Communication, Control, and Computing,* Allerton House, Monticello, IL, Oct. 2-4, 2002.

[2] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the Capacity of A Biometrical System," *Proc. 2003 IEEE Int. Symp. Inform. Theory*, Yokohama, Japan, June 29 - July 4, 2003, p. 82.

[3] E. Tuncel, P. Koulgi, and K. Rose, "Rate-Distortion Approach to Databases: Storage and Content-Based Retrieval," *IEEE Trans. Inform. Th.,* Vol. IT - 50, No. 6, pp. 953 - 967, June 2004.

[4] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inform. Theory,* vol. 54, no. 1, pp. 299-320, Jan. 2008.

[5] E. Tuncel, "Capacity/Storage Tradeoff in High-Dimensional Identification Systems," *IEEE Trans. Inform. Theory,* Vol. 55, No. 5, May 2009, pp. 2097 - 2106.

[6] F.M.J. Willems, "Searching Methods for Biometric Identification Systems: Fundamental Limits," *Proc. 2009 IEEE Int. Symp. Information Theory*, pp. 2241 - 2245, June 28 - July 3, 2009, Seoul, South Korea.

[7] T. Ignatenko and F.M.J. Willems, "Biometric Systems: Privacy and Secrecy Aspects," *IEEE Trans. Information Forensics and Security,* Vol. 4, No. 4, pp. 956 - 973, Part 2, Dec. 2009.

[8] E. Tuncel and D. Gunduz, "Identification and Lossy Reconstruction in Noisy Databases, *2010 IEEE Int. Symp. Inform. Theory,* Austin Texas, June 13-18, 2010, pp. 191 - 195.

[9] F.M.J. Willems and T. Ignatenko, "Identification and Secret-Key Generation in Biometric Systems with Protected Templates," *Proc. 2010 ACM SIGMM Multimedia and Security Workshop*, pp. 63 - 66, Sept. 9 - 10, Roma, Italy.

[10] T. Ignatenko and F.M.J. Willems, "Fundamental Limits for Biometric Identification with a Database Containing Protected Templates," *Int. Symp. Information Theory and its Applications (ISITA)*, pp. 54 - 59, Oct. 17 - 20, 2010, Taichung, Taiwan.

[11] F.M.J. Willems and T. Ignatenko, "Identification and Secret-Key Binding in Binary-Symmetric Template-Protected Biometric Systems, *2010 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1 - 5, Dec. 12 - 15, 2010, Seattle, USA.

[12] F. Farhadzadeh, F.M.J. Willems and S. Voloshynovskiy, "Fundamental Limits of Identification: Identification Rate, Search and Memory Complexity Trade - Off", *2013 IEEE Int. Symp. Inform. Theory*, Istanbul, Turkey, July 7 - 12, 2013.

Information Theoretic Analysis of Storage, Identification, and Reconstruction in Noisy Data Management Systems

Ertem Tuncel

Department of Electrical Engineering, University of California, Riverside, CA. E-mail: ertem.tuncel@ucr.edu

I. Introduction

High-dimensional data such as biometric features or behavioral patterns are replacing classical identification documents for increased security. This replacement naturally brings about the need for huge disk storage space, and more importantly, the need for fast search algorithms for reliable identification of the entries in the database when queried by a user. Observation of biometric features is typically a noisy process in both the enrollment and the identification stages, making identification challenging even without storage constraints. The noise significantly limits the ability of the search engine to distinguish database entries, and therefore, reduces the identification performance.

We address some of the important problems that can be posed in this setting: What is the maximum number of entries that can be reliable identified when the entries are to be compressed at a certain rate? Obviously, the more we compress, the less we can identify, but what is the precise trade-off? If we are also to reconstruct the data from its compressed features, how does the reconstruction quality requirement affect this tradeoff? How can we leverage the correlation between the query and the identified data in this reconstruction?

Our analysis is information theoretic in the sense that expressions such as “number” or “speed” above all become exponential rates as the dimensionality of the data approaches infinity. “Reliability,” on the other hand, gets translated into the well-known *vanishing probability* of identification probability. Classical tools of information theory gives us interesting performance bounds in these problems.

The data management system is assumed to operate in three phases:

1) *Enrollment phase:* Noisy vectors $Y^n(m)$, $m = 1, 2, \ldots, M$, are observed. It is assumed that the underlying feature vectors $X^n(m)$ are independent and identically distributed (i.i.d.) with a known distribution P_X, and pass through a memoryless channel $P_{Y|X}$ to produce $Y^n(m)$. Depending on the specific problem, the vectors $Y^n(m)$ are either directly recorded or compressed before being recorded.
2) *Identification phase:* Nature chooses W uniformly from $1, 2, \ldots, M$, and the corresponding $X^n(W)$ passes through another memoryless channel $P_{Z|X}$, producing the query vector Z^n. The goal is then to identify W with high probability by using only the query Z^n and the data stored in the system.
3) *Reconstruction phase:* Once W is successfully identified, $X^n(W)$ is estimated with the help of Z^n. It is desired that this estimation is has as little *distortion* as possible.

II. Important Results

It was shown in [1] and [2] that if the observed vectors $Y^n(m)$ are stored directly, then for large n, $M \approx 2^{nR^i}$ objects can be reliably identified if and only if $R^i < C$, where C has a single-letter characterization given by

$$C = I(Y; Z) \ .$$

If, on the other hand, $Y^n(m)$ are to be compressed before recording, there will be a tradeoff between the identification capacity and the compression rate. This tradeoff was independently characterized in [3] and [4]: A compression/identification rate pair (R^c, R^i) is achievable if and only if there exists an auxiliary random variable U such that $Z - X - Y - U$ forms a Markov chain and

$$\begin{aligned} I(Y;U) &\leq R^c \\ I(Z;U) &\geq R^i \ , \end{aligned}$$

where U is distributed over some discrete alphabet $\mathcal{U}$ satisfying $|\mathcal{U}| \leq |\mathcal{Y}| + 1$.

If, in addition, one wants to reconstruct $X^n(W)$ after successfully identifying W, then as shown in [**?**], the compression/identification/distortion triplet (R^c, R^i, D) is achievable if and only if there exist an auxiliary random variable $U \in \mathcal{U}$ with joint distribution p_{UYXZ} and a function $\phi : \mathcal{U} \times \mathcal{Z} \to \hat{\mathcal{X}}$ such that $U - Y - X - Z$ forms a Markov chain and

$$\begin{aligned} R^i &\leq I(U; Z) \\ R^c - R^i &\geq I(U; Y|Z) \\ D &\geq E[d(X, \phi(U, Z))] \ . \end{aligned}$$

References

[1] J. A. O’Sullivan and N. A. Schmid, “Large deviations performance analysis for biometrics recognition,” Proc. of Allerton Conf. on Comm., Control, and Computing, Oct. 2002, Monticello, IL.

[2] F. Willems, T. Kalker, J. Goseling and J.-P. Linnartz, “On the capacity of a biometrical identification system,” *Proc. IEEE Int’l Symp. Inform. Theory*, Yokohama, Japan, July 2003.

[3] M. B. Westover and J. A. O’Sullivan, “Achievable rates for pattern recognition,” *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 299-320, Jan. 2008.

[4] E. Tuncel, “Capacity/storage tradeoff in high-dimensional identification systems,” *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2097-2106, May 2009.

[5] E. Tuncel and D. Gunduz “Identification and lossy reconstruction in noisy databases,” to appear in *IEEE Trans. Inform. Theory*.

Compression for Similarity Queries

Amir Ingber, Thomas Courtade, Idoia Ochoa and Tsachy Weissman

Traditionally, data compression deals with the problem of concisely representing a data source, e.g. a sequence of letters, for the purpose of eventual reproduction (either exact or approximate). In this work we are interested in the case where the goal is to answer *similarity queries* about the compressed sequence, i.e. to identify whether or not the original sequence is similar to a given query sequence.

We study the fundamental tradeoff between the compression rate and the reliability of the queries performed on compressed data. For i.i.d. sequences, we characterize the minimal compression rate that allows query answers, that are reliable in the sense of having a vanishing false-positive probability, when false negatives are not allowed. We term this fundamental limit the *identification rate* of the source. Our results for the case of Gaussian sources with quadratic distortion [1][2] are based on high-dimensional geometry of the Euclidean space. For general discrete memoryless sources (DMS's) and arbitrary distortion measures, our results [3] are partially based on the work of Ahlswede et al. [4], and their "inherently typical subset lemma" plays a key role in the converse proof. For compression rates above the identification rate, we show that the false positive probability vanishes exponentially, and we characterize this exponent (see [2] for Gaussian sources, [5] for the general DMS case and [6] for the special case of exact match identification).

We then study the relationship between classical lossy compression and compression for queries. In general, lossy compressors can be used as building blocks for constructing a scheme for compression for queries, but this should be done carefully. While the naive usage of lossy compressors is optimal in some cases (see [3]), it is not optimal in general. Nevertheless, such schemes can be constructed in practice. In [7], lossy compressors are successfully used in order to compress equiprobable sources, and are also applied for the task of compressing DNA sequences (taken from the Biozon database [8]) for similarity identification. In [9], an improved scheme is described, which is based on a lossy compressor that minimizes a distortion measure that differs from the one that measures the similarity between sequences. This approach forces the lossy compressor to match reconstruction sequences to source sequences, with a joint type that is carefully chosen in order to minimize the false positive probability and - by that - the identification rate.

References

[1] A. Ingber, T. Courtade, and T. Weissman, "Quadratic similarity queries on compressed data," in *Data Compression Conference (DCC)*, 2013, pp. 441–450.

[2] A. Ingber, T. A. Courtade, and T. Weissman, "Compression for quadratic similarity queries," Submitted to IEEE Trans. on Information Theory, 2013. [Online]. Available: http://arxiv.org/abs/1307.6609

[3] A. Ingber and T. Weissman, "The minimal compression rate for similarity identification," Submitted to IEEE Trans. on Information Theory, 2013. [Online]. Available: http://arxiv.org/abs/1312.2063

[4] R. Ahlswede, E.-h. Yang, and Z. Zhang, "Identification via compressed data," *IEEE Trans. on Information Theory*, vol. 43, no. 1, pp. 48 –70, Jan 1997.

[5] A. Ingber and T. Weissman, "The error exponent in compression for similarity identification," in prep., 2013.

[6] A. Ingber, T. Courtade, and T. Weissman, "Compression for exact match identification," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2013, pp. 654–658.

[7] I. Ochoa, A. Ingber, and T. Weissman, "Efficient similarity queries via lossy compression," in *51st Annual Allerton Conference on Communication, Control and Computing*, Monticelo, IL, Sep. 2013.

[8] A. Birkland and G. Yona, "BIOZON: a system for unification, management and analysis of heterogeneous biological data," *BMC bioinformatics*.

[9] I. Ochoa, A. Ingber, and T. Weissman, "Compression schemes for similarity queries," in *Data Compression Conference (DCC)*, 2014, to appear.

The authors are with the Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305.

Upper and Lower Bounds on the Reliability of Content Identification

Gautam Dasarathy
Electrical and Computer Engineering
University of Wisconsin - Madison
dasarathy@wisc.edu

Stark C. Draper
Electrical and Computer Engineering
University of Toronto
stark.draper@utoronto.ca

***Abstract*—In this paper we quantify upper and lower bounds on the reliability function for the problem of content identification from a large database based on noisy queries.**

I. Introduction

We consider the problem of content identification from a database. The database houses quantized representations of 2^{nR_I} length-n "enrollment" vectors, where $R_I > 0$ is what we call the *identification rate* (we will not worry about integer effects in this paper). When presented with a noisy observation of one of the (non-quantized) enrollment vectors, the goal is to identify (using only this observation and the stored data) which enrollment vector generated the noisy observation (or the "query")

The information-theoretic limits of this problem were found in [1] (see also [2] for related problem formulations) under the following model. The enrollment vectors, $\mathbf{X}_m$ for $m = 1, \ldots, 2^{nR_I}$ are chosen in an i.i.d. manner according to p_X. An index of a codeword in a pre-defined rate-R_C codebook $\mathcal{C}$ is used to represent each of these vectors in a database, where $R_C > 0$ is what we call the *compression rate*. As it takes nR_C bits to store the index of the representation of each enrollment, and 2^{nR_I} representations are stored, the entire database is of size $nR_C 2^{nR_I}$ bits. The query $\mathbf{Y}$ presented during the identification phase is a length-n observation of one of the 2^{nR_I} (unquantized) enrollment vectors observed via the discrete memoryless channel (DMC) $p_{Y|X}(\cdot|\cdot)$. The decoder's objective is to identify reliably the codeword corresponding to the enrollment vector from which the query was generated.

In [1], the capacity region of this problem was shown to be parameterized by a (rate-distortion) test channel $p_{U|X}$. Given the joint distribution $p_{U|X}(u|x)p_X(x)p_{Y|X}(y|x)$, the "compression/identification" rate pair (R_C, R_I) is achievable if

$$R_C > I(U;X) \qquad R_I < I(U;Y).$$

The achievable rate region is the convex hull of the union of achievable rate pairs over all test channels. The achievability is closely related the the Wynzer-Ziv problem. The codebook $\mathcal{C}$ needs to have good covering properties ($R_C > I(U;X)$) to ensure reliable encoding. The union of codewords stored in the database (which is a subset of $\mathcal{C}$) needs to form a code that has good packing properties ($R_I < I(U;Z)$) for reliable decoding.

In a previous paper [3], we quantified an achievable error exponent tradeoff for a particular encoding and decoding strategy, in other words, we derived a lower bound on the reliability function for the content identification problem. Our results used a novel lemma [3, Lemma 2] that characterized the *end-to-end* statistical relationship between the codeword and the channel output. In this paper, we will first quantify an upper bound on the error exponent over all encoding/decoding strategies. This upper bound can be shown to be strictly tighter than the naive sphere packing upper bound. We also derive a modification of the lower bound from [3] whose form parallels the upper bound expression obtained here

II. Problem Formulation

In this section we formally state the problem setting and define the notation that we will use throughout the paper.

Environment: We suppose that there are $M = 2^{nR_I}$ items to be represented in the database. To each item is associated a length-n "feature vector" or "enrollment vector" $\mathbf{X}(m) \in \mathcal{X}^n$, $m = 1, 2, \ldots, M$ which are drawn independently from p_X in an i.i.d manner where $\mathcal{X}$ is a finite alphabet.

Enrollment Phase: In the enrollment phase, each feature vector is mapped to a codeword selected from a pre-defined rate-R_C codebook $\mathcal{C} \triangleq \{\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(2^{nR_C})\}$ ignoring integer effects. The codewords are made up of symbols from the finite alphabet $\mathcal{U}$. We represent the operation of assigning a codeword $\mathbf{u}$ to the feature vector $\mathbf{x}$ by the function $f : \mathcal{X}^n \to \{1, 2, \ldots, L\}$, where we define $L = 2^{nR_C}$. The notation $J(m)$ will be used to denote the (random) quantity $f(\mathbf{X}(m))$. Observe that the codeword that gets assigned to object m is $\mathbf{u}(J(m))$. The set of all (possibly non-distinct) codewords corresponding to the enrolled items $\{\mathbf{u}(J(1)), \mathbf{u}(J(2)), \ldots, \mathbf{u}(J(M))\}$ will henceforth be called the *database* and will be denoted as $\mathcal{D}$. The database $\mathcal{D} \subset \mathcal{C}$ can be thought of as an analogue to the random bin of codewords in the Wyner-Ziv problem.

Identification Phase: An index W is selected uniformly at random from $\{1, 2, \ldots, M\}$. This corresponds to the item that the user wishes to query for. A noisy version of $\mathbf{X}^n(W)$, $\mathbf{Y}^n \in \mathcal{Y}^n$ is then observed at the database where the conditional distribution of $\mathbf{Y}^n$ is given by $\Pr[\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}] = \prod_{i=1}^n p_{Y|X}(y_i|x_i)$, i.e., we model the noise as a DMC $p_{Y|X} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is also assumed to be a finite alphabet. The objective at the identification phase is to produce an estimate $\widehat{W}$ from the observed random vector $\mathbf{Y}$ and the stored sequences $\mathcal{D}$. This estimation operation, which we will call *decoding*, is denoted by the function $g : \mathcal{Y}^n \times \mathcal{U}^{nM} \to \{1, 2, \ldots, M\}$. Our aim is to design an encoding function $f(\cdot)$ and a decoding function $g(\cdot)$ such that, with high probability, $g(\cdot)$ returns the correct value of W.

Given a finite alphabet $\mathcal{S}$, we write $\Pi(\mathcal{S})$ to denote the set of all distributions on $\mathcal{S}$. For other notation, we mainly follow [4].

III. Main Results

We will now state the main results of our paper. Given some joint distribution $p_{XY} \in \Pi(\mathcal{X} \times \mathcal{Y})$, let $P_e(f, g, p_{XY})$ denote the probability that a particular choice (f, g) of encoding and decoding function does not estimate W correctly. Then, we can define the *content identification reliability function* as follows

$$\rho(p_{XY}, R_I, R_C) = \lim_{\epsilon\downarrow 0}\limsup_{n\to\infty} -\frac{1}{n}\log\left[\min_{f,g} P_e\left(f, g, p_{XY}\right)\right],$$

where the minimization is over all encoding and decoding function pairs (f, g) such that $\log M \leq n(R_I - \epsilon)$ and $\log L \leq n(R_C + \epsilon)$.

The following theorem provides an upper bound for the reliability function ρ. That is, it gives us a lower bound on the probability of error over all choices of valid encoding and decoding functions for the problem.

Theorem 1. *Given $p_{XY} \in \Pi(\mathcal{X} \times \mathcal{Y})$ and $R_I, R_C > 0$, the reliability function $\rho(p_{XY}, R_I, R_C)$ is upper bounded by*

$$\rho_U(p_{XY}, R_I, R_C) \triangleq \inf_{q_X} \sup_{\substack{q_{S|X}:\\ I(X;S)\leq R_C}} \inf_{\substack{q_{Y|X}:\\ I(Y;S)\leq R_I}} D\left(q_{XY} \| p_{XY}\right). \tag{1}$$

Here Y, X, S have the joint distribution $q_Y q_{X|Y} q_{S|X}$ and the cardinality of S satisfies

$$|S| \leq |\mathcal{X}| \cdot |\mathcal{Y}| + |\mathcal{X}| + 2$$

Next, we will state a theorem that provides a lower bound to the reliability function ρ and therefore upper bounds the probability of error over all choices of encoders and decoders.

Theorem 2. *Given $p_{XY} \in \Pi(\mathcal{X} \times \mathcal{Y})$ and $R_I, R_C > 0$, the reliability function $\rho(p_{XY}, R_I, R_C)$ is lower bounded by the quantity*

$$\rho_L \triangleq \inf_{q_X} \sup_{\substack{q_{U|X}:\\ I(X,U)<R_C}} \inf_{q_{Y|X,U}} D(q_{XYU} \| p_{XY} q_{U|X}) + |I(U, Y) - R_I|^+,$$

where the joint distribution of X, Y, U is $q_X q_{U|X} q_{Y|X,U}$.

The rest of the paper will be devoted to proving these two results.

IV. Proofs

A. Proof of Theorem 1

In order to prove Theorem 1, we begin by picking an arbitrary encoder-decoder pair f, g that satisfies the rate constraints. That is, f and g are such that

$$\log M \leq nR_I \qquad \log L \leq nR_C. \tag{2}$$

Let us define the "error set"

$$\mathcal{E} = \{(\mathbf{x}(1), \ldots, \mathbf{x}(M), w, \mathbf{y}) : g(\mathbf{y}, \mathbf{x}(1), \ldots, \mathbf{x}(M)) \neq w\}. \tag{3}$$

Our proof will follow these steps:

(A) We will first fix a *bad* joint distribution q_{XY} such that $I(\mathbf{Y}, f(\mathbf{X})) \leq n(R_I - \epsilon)$. We will then show that $q^n_{XY}(\mathcal{E})$ is bounded away from zero.

(B) We will then use this fact to bound the probability of error from below over all valid encoder/decoder pairs. Notice that probability of error is the quantity $p^n_{XY}(\mathcal{E})$.

(C) We will next work to "single letterize" the expressions in the above bound. This will involve introducing the right auxiliary random variable and then ensuring that the corresponding alphabet size does not grow with n. Since this step is somewhat standard, we will only sketch the argument here.

(D) Finally, we use continuity arguments to show that we can take $\epsilon \to 0$, the details of which we will omit in this short manuscript.

Step (A): Let $q_{XY} \in \Pi(\mathcal{X} \times \mathcal{Y})$ be a distribution such that $I(\mathbf{Y}, f(\mathbf{X})) \leq n(R_I - \epsilon)$. Recall that g is an estimator for the random variable W and it has access to the random variables $\{J(i)\}_{i\in[M]}$ and $\mathbf{Y}$. Therefore, Fano's inequality (see e.g., [5]) tells us that

$$q^n_{XY}(\mathcal{E}) \geq \frac{H(W \mid J(1), \ldots, J(M), \mathbf{Y}) - 1}{\log M}. \tag{4}$$

Notice here that the conditional entropy is computed with respect to the joint distribution q^n_{XY}.

To get a lower bound on $q^n_{XY}(\mathcal{E})$, we will lower bound the conditional entropy term above.

$$\begin{aligned}
&H(W \mid J(1), \ldots, J(M), \mathbf{Y}) \\
&\quad = H(W) - I(W; J(1), \ldots, J(M), \mathbf{Y}) \\
&\quad \overset{(a)}{=} H(W) - I(W; \mathbf{Y} \mid J(1), \ldots, J(M)) \\
&\quad \overset{(b)}{=} H(W) - H(\mathbf{Y} \mid J(1), \ldots, J(M)) \\
&\qquad\quad + H(\mathbf{Y} \mid J(1), \ldots, J(M), W) \\
&\quad \overset{(c)}{\geq} H(W) - H(\mathbf{Y}) + H(\mathbf{Y} \mid J(W)) \\
&\quad \overset{(d)}{=} nR_I - I(f(\mathbf{X}); \mathbf{Y}) \\
&\quad \geq n\epsilon
\end{aligned}$$

In (a), we have used the definition of conditional mutual information and that fact that $I(W; J(1), \ldots, J(M)) = 0$ since these quantities are independent. In (b) we expand out the conditional mutual information term and in (c) we use the fact that conditioning cannot increase entropy (i.e., $H(\mathbf{Y} \mid \{J(i)\}_{i\in[M]}) \leq H(\mathbf{Y})$) and that $\mathbf{Y}$ only depends on $J(W)$ and therefore $H(\mathbf{Y} \mid \{J(i)\}_{i\in[M]}, W) = H(\mathbf{Y} \mid J(W))$. Finally, in (d) we use the fact that W is a uniform random variable over a set of size $M = 2^{nR_I}$ and the fact that, by definition, $J(W) = f(\mathbf{X})$. The final step follows from our assumption on q_{XY}. Let us assume that $n \geq 2\epsilon^{-1}$. Substituting this lower bound back in (4), we have

$$q^n_{XY}(\mathcal{E}) \geq \frac{n\epsilon - 1}{nR_I} \geq \frac{\epsilon}{2R_I} \tag{5}$$

Step (B): In order to lower bound the probability of error, $p^n_{XY}(\mathcal{E})$, we will use the so-called "change of measure" argument typically used in establishing *large deviation rate functions* (see e.g., [6] and [4, Page 268, Problem 13] for a version that is more applicable to an information theoretic setting). The idea is to lower bound the mass of a set (in this case $\mathcal{E}$) under one measure (in this case p^n_{XY}) using the mass of the same set under a different measure (in this case q^n_{XY}). For $\delta > 0$, let us define the set of all (q_{XY}, p_{XY}) – *divergence typical sequences* as follows

$$\mathfrak{D} = \left\{ (\mathbf{x}, \mathbf{y}) : \left| \frac{1}{n} \log \frac{q^n_{XY}(\mathbf{x}, \mathbf{y})}{p^n_{XY}(\mathbf{x}, \mathbf{y})} - D(q_{XY} \| p_{XY}) \right| \leq \delta \right\}.$$

Notice that by Chebyshev's inequality, we have that $q^n_{XY}(\mathfrak{D}^c)$ is no greater than $\frac{\mathbb{E}_{q_{XY}}\left[\log^2 \frac{q_{XY}}{p_{XY}}\right]}{n\delta^2}$. Assuming that $p_{XY} > 0$, this quantity can be uniformly upper bounded by some $\Delta > 0$ over all q_{XY} when $\mathcal{X}$ and $\mathcal{Y}$ are finite. Of course, one could obtain a tighter bound for $q^n_{XY}(\mathfrak{D}^c)$ using, for instance, Chernoff bounds. However, this weaker bound suffices for our current purpose. We can now proceed to bound the probability of error as follows for $n \geq \frac{4R_I\Delta}{\epsilon\delta^2}$,

$$\begin{aligned}
\mathbb{P}\,[\text{error}] &= p^n_{XY}(\mathcal{E}) \\
&\geq p^n_{XY}(\mathcal{E} \cap \mathfrak{D}) = \sum_{\mathbf{x},\mathbf{y} \in \mathcal{E}\cap\mathfrak{D}} q^n_{XY}(\mathbf{x}, \mathbf{y}) \frac{p^n_{XY}(\mathbf{x}, \mathbf{y})}{q^n_{XY}(\mathbf{x}, \mathbf{y})} \\
&\overset{(a)}{\geq} q^n_{XY}(\mathcal{E} \cap \mathfrak{D}) e^{-n[D(q_{XY}\|p_{XY})+\delta]} \\
&\overset{(b)}{\geq} \left(\frac{\epsilon}{2R_I} - \frac{\Delta}{n\delta^2} \right) e^{-n[D(q_{XY}\|p_{XY})+\delta]} \\
&\overset{(c)}{\geq} \frac{\epsilon}{4R_I} e^{-n[D(q_{XY}\|p_{XY})+\delta]}.
\end{aligned}$$

In (a) we use the definition of the divergence typical set $\mathfrak{D}$ and in (b) we use the fact that $P(A\cap B) \geq |P(A) - P(B^c)|^+$. Finally in (c) we use our assumption that $n \geq \frac{4R_I\Delta}{\epsilon\delta^2}$. Since this bound holds for any q_{XY} such that $I(f(\mathbf{X}), \mathbf{Y}) \leq n(R_I - \epsilon)$, it has to hold even if we take the supremum over all such q_{XY}'s. Also, since we want our lower bound to hold irrespective of the quality of the encoder/decoder pair picked, we can take the infimum of right side with respect to all valid coding functions. Therefore, the probability of error can be lower bounded by the following quantity

$$\inf_{\substack{f:\\ \log L \leq n(R_C+\epsilon)}} \; \sup_{\substack{q_{XY}:\\ I(\mathbf{Y};f(\mathbf{X}))\leq n(R_I-\epsilon)}} \frac{\epsilon \exp\Big(- n\,[D(q_{XY}\|p_{XY}) + \delta] \Big)}{4R_I}. \tag{6}$$

Step (C): Before we proceed, let us restrict our attention to the exponent. Notice that our objective will now be to upper bound the subsequent quantities.

$$\begin{aligned}
&\sup_{\substack{f:\\ \log L \leq n(R_C+\epsilon)}} \; \inf_{\substack{q_{XY}:\\ I(\mathbf{Y};f(\mathbf{X}))\leq n(R_I-\epsilon)}} D\left(q_{XY}\|p_{XY}\right) \\
&\overset{(a)}{\leq} \inf_{q_X} \sup_{\substack{f:\\ I(f(\mathbf{X});\mathbf{X})\leq n(R_C+\epsilon)}} \; \inf_{\substack{q_{Y|X}:\\ I(\mathbf{Y};f(\mathbf{X}))\leq n(R_I-\epsilon)}} D\left(q_{XY}\|p_{XY}\right) \\
&\overset{(b)}{\leq} \inf_{q_X} \sup_{\substack{q_{U|\mathbf{X}}:\\ I(U;\mathbf{X})\leq n(R_C+\epsilon)}} \; \inf_{\substack{q_{Y|X}:\\ I(\mathbf{Y};U)\leq n(R_I-\epsilon)}} D\left(q_{XY}\|p_{XY}\right)
\end{aligned} \tag{7}$$

In (a) we changed the order of minimization and maximization, and we used the fact that $\log L = H(f(\mathbf{X})) \geq I(f(\mathbf{X}); \mathbf{X})$. In (b), we are replacing the deterministic mapping $f(\mathbf{X})$ by a random variable U (whose alphabet has cardinality no less than the range of $f(\mathbf{X})$). Since deterministic functions are a special case of such random mappings, we are increasing the domain of maximization, and hence the inequality (this does not affect the inner infimization since the domain of that infimum is over $q_{\mathbf{Y}|\mathbf{X}}$). Towards "single-letterizing" the arguments of the optimizations above, we do the following two calculations. First, we observe that

$$\begin{aligned}
I(\mathbf{Y}; U) &= \sum_{i=1}^n H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid U, Y_1^{i-1}) \\
&\overset{(a)}{\leq} \sum_{i=1}^n H(Y_i) - H(Y_i \mid U, X_1^{i-1}) = \sum_{i=1}^n I(Y_i; U, X_1^{i-1}) \\
&\overset{(b)}{=} \sum_{i=1}^n I(Y_i; V_i) \overset{(c)}{=} I(Y_T; V_T, T),
\end{aligned}$$

where (a) follows from observing that $H(Y_i \mid Y_1^{i-1}) \leq H(Y_i)$ and that $H(Y_i \mid U, X_1^{i-1}) = H(Y_i \mid U, X_1^{i-1}, Y_1^{i-1}) \leq H(Y_i \mid U, Y_1^{i-1})$ because of the Markov chain $Y_i - (U, X_1^{i-1}) - Y_1^{i-1}$. In (b), we set $V_i \triangleq (U, X_1^{i-1})$ and in (c) we merely introduce a random variable T that is uniformly distributed in $[1, \ldots, n]$ (this is the so called "time sharing" random variable.) Second, using a similar calculation, we have

$$\begin{aligned}
I(\mathbf{X}; U) &= \sum_{i=1}^n I(X_i; U \mid X_1^{i-1}) = \sum_{i=1}^n I(X_i; X_1^{i-1}, U) \\
&= \sum_{i=1}^n I(X_i; V_i) = I(X_T; V_T, T)
\end{aligned}$$

Using these two calculations and defining $W \triangleq (V_T, T)$, we can upper bound (7) as

$$\inf_{q_X} \sup_{\substack{q_{U|\mathbf{X}}:\\ I(X_T,W)\leq(R_C+\epsilon)}} \; \inf_{\substack{q_{Y|X}:\\ I(Y_T;W)\leq(R_I-\epsilon)}} D\left(q_{XY}\|p_{XY}\right)$$

(Technically, since we are decreasing the size, the domain of infimization here might turn out to be empty. So, we will use the standard convention that an infimum over an empty set is ∞ to maintain consistency.) Now, notice that since X_T has the same distribution as X and that $H(X_T \mid W) = H(X_T \mid U, X_1^{T-1}, T) = H(X \mid W)$, we have that $I(X_T; W) = H(X_T) - H(X_T \mid W) = I(X; W)$. Similarly, $I(Y_T; W) = I(Y; W)$. Finally, we define the auxiliary random variable S so that $(X, W) = (X, X_1^{T-1}, T, U)$ has the same distribution as (X, S). Of course, this also implies that (Y_T, W) has the same distribution as (Y, S). Therefore, we obtain the upper bound

$$\tilde{\rho}_U\left(p_{XY}, R_C+\epsilon, R_I+\epsilon\right) \triangleq \inf_{q_X} \sup_{\substack{q_{S|X}: \\ I(X;S)\le(R_C+\epsilon)}} \inf_{\substack{q_{Y|X}: \\ I(Y;S)\le(R_I-\epsilon)}} D\left(q_{XY} \| p_{XY}\right). \tag{8}$$

In order to conclude, we must next show that the auxiliary random variable that we introduced has an alphabet size that does not grow with n.

We will only sketch the rest of the argument since the techniques are somewhat standard. Let $\bar{\rho}_U$ be the same expression as (8) but with an extra cardinality constraint $|\mathcal{S}| \le |\mathcal{X}|\,|\mathcal{Y}|+|\mathcal{X}|+2$. Our goal will be to show that $\bar{\rho}_U = \tilde{\rho}_U$. It is of course easy to see that $\bar{\rho}_U \le \tilde{\rho}_U$. In order to show the other direction, notice that it suffices to show that for each q_X and $q_{S|X}$ such that $I(X;S) \le R_C + \epsilon$, there exists a $\tilde{q}_{S|X}$ such that (a) it satisfies the same mutual information constraints, (b) it has $|\mathcal{S}| \le |\mathcal{X}| \cdot |\mathcal{Y}| + |\mathcal{X}| + 2$ and (c) if one defines the function

$$f(q_X, q_{S|X}) \triangleq \inf_{\substack{q_{Y|X}: \\ I(X;S)\le R_C+\epsilon}} D(q_{XY} \| p_{XY}), \tag{9}$$

then $f(q_X, q_{S|X}) \le f(q_X, \tilde{q}_{S|X})$.

Since $p_{XY} > 0$ and the optimization problem above has a convex bounded objective over a compact set, a minimizer $q^*_{Y|X}$ exists such that

$$f(q_X, q_{S|X}) = D(q_X q^*_{Y|X} \| p_{XY}). \tag{10}$$

Also, since the optimization problem is convex, we know that $q^*_{Y|X}$ satisfies the KKT conditions [7]. To conclude this step, we use the fact that the KKT conditions are essentially $|\mathcal{X}| \cdot |\mathcal{Y}|$ linear equalities involving $q^*_{Y|X}$ (which correspond to the gradient condition) and a set of $|\mathcal{X}|$ linear equalities corresponding to the mutual information constraint. Therefore, by Caratheodory's theorem, we have that the cardinality $|\mathcal{S}| \le |\mathcal{X}| \cdot |\mathcal{Y}| + |\mathcal{X}| + 2$ and that $\tilde{q}_{S|Y}$ satisfies the necessary constraints.

Step (D): Finally, using continuity arguments, we can show that $\lim_{\epsilon\to 0} \tilde{\rho}_U(p_{XY}, R_I+\epsilon, R_C+\epsilon) \to \rho_U(p_{XY}, R_I, R_C)$. This concludes the proof.

B. *Proof of Theorem 2*

In [3], we show, using an important lemma about the end-to-end behavior across a Markov chain, that the reliability function ρ of the problem is lower bounded by

$$\tilde{\rho}_L = \inf_{q_X} \sup_{\substack{q_{S|X}: \\ I(X;S)\le R_C}} \inf_{q_{Y|S}} D(q_X \| p_X) + D\left(\tilde{J}^*_{q_{Y|S}, q_X, q_{S|X}} \| J \mid q_S\right) + |I(S;Y) - R_I|^+ \tag{11}$$

where $q_S(\cdot) = \sum_{x\in\mathcal{X}} q_X(x) q_{S|X}(\cdot \mid x)$ is the distribution induced on $\mathcal{S}$ by q_X and $q_{S|X}$, $J : \mathcal{S} \to \mathcal{X} \times \mathcal{Y}$ is a stochastic matrix defined as $J(x, y \mid s) = \frac{q_{S|X}(s|x) q_X(x)}{q_S(s)} p_{Y|X}(y \mid x)$, for all $(s, x, y) \in \mathcal{S} \times \mathcal{X} \times \mathcal{Y}$, and $\tilde{J}^*_{q_{Y|S}, q_X, q_{S|X}} : \mathcal{S} \to \mathcal{X} \times \mathcal{Y}$ is a stochastic matrix defined as

$$\tilde{J}^*_{q_{Y|S}, q_X, q_{S|X}} = \arg\min_{\tilde{J}\in\mathcal{E}(q_{Y|S})} D(\tilde{J} \| J \mid q_S), \tag{12}$$

where $\mathcal{E}(q_{Y|S})$ is defined as the set $\left\{\tilde{J} : \sum_{x\in\mathcal{X}} \tilde{J}(x, y \mid s) = q_{Y|S}(y \mid s),\ \forall (s, y) \in \mathcal{S} \times \mathcal{Y}\right\}$.

It can be shown that this error exponent is positive in the capacity region indicated in [1]. We will now show that $\tilde{\rho}_L$ equals ρ_L so that the expression matches ρ_U more closely.

Towards this end, consider the first two terms in the expression for $\tilde{\rho}_L$ and observe that.

$$\begin{aligned} & D(q_X \| p_X) + D\left(\tilde{J}^* \middle\| J \middle| q_S\right) \\ & \overset{(a)}{=} D(q_X \| p_X) + D\left(q_X q_{S|X} q^*_{Y|X,S} \middle\| q_X q_{S|X} q^*_{Y|X,S}\right) \\ & \overset{(b)}{=} D(q_X q_{S|X} q^*_{Y|X,S} \| p_X q_{S|X} q^*_{Y|X,S}) \\ & \qquad + D\left(q_X q_{S|X} q^*_{Y|X,S} \middle\| q_X q_{S|X} q^*_{Y|X,S}\right) \\ & = D(q_X q_{S|X} q^*_{Y|X,S} \| p_{XY} q_{S|X}), \end{aligned}$$

where (a) follows from the definition of the conditional KL divergence, (b) follows from multiplying and dividing the appropriate term in the first KL divergence term, and the last line is a simple algebraic simplification of (b). Finally we note that every choice of q_X, $q_{Y|X}$, and $q_{Y|X,S}$ fixes the $q_{Y|S}$ distribution and therefore, one can equivalently optimize over $q_{Y|S,X}$ in the inner most infimum of (11). This concludes the proof.

References

[1] E. Tuncel. Capacity/storage tradeoff in high-dimensional identification systems. *IEEE Trans. Inform. Theory*, 55:2097–2106, November 2009.

[2] M. B. Westover and J. A. O'Sullivan. Achievable rates for pattern recognition. *IEEE Trans. Inform. Theory*, 54:299–320, Jan 2008.

[3] G. Dasarathy and S. C. Draper. On reliability of content identification from databases based on noisy queries. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 1066–1070. IEEE, 2011.

[4] I. Csiszár and J. Körner. *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, third edition, 1981.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.

[6] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38. Springer, 2010.

[7] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

On the Fluctuation of Mutual Information in Double Scattering MIMO Channels

Zhong Zheng*, Lu Wei†, Roland Speicher‡, Ralf Müller§, Jyri Hämäläinen*, and Jukka Corander†
* Department of Communications and Networking, Aalto University, Finland
Email: {zhong.zheng,jyri.hamalainen}@aalto.fi
† Department of Mathematics and Statistics, University of Helsinki, Finland
Email: {lu.wei,jukka.corander}@helsinki.fi
‡ Faculty of Mathematics and Computer Science, Saarland University, Germany
Email: speicher@math.uni-sb.de
§ Institute for Digital Communications, University of Erlangen-Nuremberg, Germany
Email: mueller@lnt.de

***Abstract*—In this paper, we study the fluctuation of mutual information in the presence of double-scattering MIMO channels. Based on techniques from free probability theory, the asymptotic variance of the mutual information is obtained. Using the derived results, we construct a Gaussian approximation to the channel outage probability. Numerical results show that the asymptotic analysis provides close approximations for realistic MIMO configurations.**

***Keywords*—Double scattering, free probability theory, multi-input multi-output, mutual information, outage probability, product of random matrices.**

I. Introduction

In recent years, multi-input multi-output (MIMO) wireless communication systems have received considerable attention since MIMO is seen as the most credible way to increase link level capacity. Extensive works have focused on the performance of MIMO channels with the assumption of a rich scattering environment. Therein, the presumed channel models are full rank Rayleigh or Rician MIMO channels. However, measurements show that signal propagations are subject to rank deficiency caused by insufficient scatterings in certain outdoor [1, 2] as well as indoor environments [3]. This leads to degradation of both multiplexing gain and diversity gain [4]. Motivated by these facts, a double-scattering model or the so-called multi-keyhole model was proposed in [1, 5], which explicitly encompasses the above described scattering structure. For multi-antenna transceivers, the double-scattering channel consists of two stage fading modeled as a product of two MIMO channel matrices.

There are a number of studies concerning the information-theoretic quantities of the double-scattering channels. Shin *et. al.* derived an upper bound to ergodic mutual information for the double scattering channel [6, Th. III.3] and an exact expression for the single keyhole channel [6, Th. III.4]. The authors in [7] investigated the asymptotic Rayleigh-limit when the matrix dimension approaches infinity. In such a limit, the double-scattering model reduces to an equivalent Rayleigh MIMO channel. With all matrices dimensions being large, the ergodic mutual information of the double-scattering channels has been obtained in [5] via numerical integrations. Recently, a closed-form expression for ergodic mutual information was derived in [8] for double scattering channel. Moreover, the authors in [9–11] derived the ergodic mutual information for finite dimensional channel matrices. However, all the above results are valid for ergodic channels, where each codeword has infinite length. In many practical cases, each codeword only sees finitely many channel realizations. In these cases, the ergodic mutual information has no physical significance, whereas the outage probability is a more relevant performance metric [12]. Despite the importance to understand the channel outage probability, it remains a challenging issue in the case of double-scattering channels.

To address this issue, we derive a compact expression for the asymptotic variance of mutual information of the considered channel model. The result is formally valid if dimensions of channel matrices grow to infinity. Yet, the numerical simulations show that the result is served as a good approximation for practical antenna configurations. The presented analysis is enabled by adopting a recent result of higher order freeness in free probability theory [13]. The asymptotic variance is then used to construct a Gaussian approximation to the outage probability.

II. System Model

Consider a MIMO communication channel with T transmit and R receive antennas. The channel output vector $\mathbf{y} \in \mathbb{C}^R$, at a given time instance, reads

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^T$ is the transmit vector and follows the complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix[1] $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\dagger]$. The additive noise $\mathbf{n} \in \mathbb{C}^R$ is modeled as *i.i.d.* complex Gaussian variables with $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_T)$. The channel matrix $\mathbf{H}$ is given by the double-scattering model [5]

$$\mathbf{H} = \frac{1}{\sqrt{RS}} \mathbf{\Psi}^\dagger \mathbf{\Theta}, \tag{2}$$

where S denotes the number of scatterers, $\mathbf{\Psi} \in \mathbb{C}^{S \times R}$ and $\mathbf{\Theta} \in \mathbb{C}^{S \times T}$ describe the propagation between the scattering

[1] $(\cdot)^\dagger$ denotes conjugate transpose.

objects and antenna arrays in the receiver and transmitter, respectively. The entries of $\mathbf{\Psi}$ and $\mathbf{\Theta}$ are *i.i.d.* complex Gaussian distributed with zero-mean and unit variance. Namely, $\mathbf{H}$ is the product of two complex Gaussian random matrices.

We assume that the channel state information is only known by the receiver. In this case, $\mathbb{E}[\mathbf{x}\mathbf{x}^\dagger] = \gamma \mathbf{I}_T$ with γ being the SNR per receive antenna. The mutual information of the channel (1) in nats/s/Hz is defined as

$$\mathcal{I} = \log\det\left(\mathbf{I}_T + \gamma \mathbf{H}^\dagger \mathbf{H}\right) = \sum_{i=1}^{T} g(\lambda_i), \tag{3}$$

where $g(\lambda) = \log(1+\gamma\lambda)$ and λ_i, $i = 1, \ldots, T$, are eigenvalues of $\mathbf{Q} = \mathbf{H}^\dagger\mathbf{H}$. The outage probability, for a given rate c, is

$$P_{\mathrm{out}}(c) = \Pr(\mathcal{I} < c) = F_{\mathcal{I}}(c), \tag{4}$$

where $F_{\mathcal{I}}(c)$ is the cumulative distribution function of $\mathcal{I}$.

III. Asymptotic Variance of Mutual Information

In this section, we consider the limiting variance of mutual information in the asymptotic limit

$$T,\ S,\ \text{and } R \to \infty \text{ such that } \rho = \frac{S}{R} \text{ and } \zeta = \frac{T}{S} \tag{5}$$

are fixed. In literature, the ratio ρ is known as richness of the channel [5]. Under the condition (5), the asymptotic expected value of $\mathcal{I}$ is studied in [5, 8]. In the following, we derive the asymptotic variance of $\mathcal{I}$ with the free probability machinery. The results are then used to construct an approximation to the channel outage probability. In particular, we need the second order Cauchy transform of $\mathbf{Q}$, denoted as $G_{\mathbf{Q}}(x,y)$. Define the first order Cauchy transform of matrix $\mathbf{Q}$ as

$$G_{\mathbf{Q}}(z) = \int \frac{1}{\lambda - z}\,\mathrm{d}\widetilde{F_{\mathbf{Q}}}(\lambda), \tag{6}$$

where $\widetilde{F_{\mathbf{Q}}}(\lambda)$ is the empirical eigenvalue distribution of $\mathbf{Q}$. An explicit expression of $G_{\mathbf{Q}}(x,y)$ is summarized in the following lemma.

Lemma 1. *Let $\mathbf{P} = \mathbf{\Theta}\mathbf{\Theta}^\dagger/S$. Then the second order Cauchy transform of $\mathbf{Q}$ is given by*

$$\begin{aligned} G_{\mathbf{Q}}(x,y) = {} & G'_{\mathbf{Q}}(x)G'_{\mathbf{Q}}(y)H(-G_{\mathbf{Q}}(x), -G_{\mathbf{Q}}(y)) \\ & + \frac{\partial^2}{\partial x \partial y}\log\frac{G_{\mathbf{Q}}(x) - G_{\mathbf{Q}}(y)}{x-y}, \end{aligned} \tag{7}$$

where

$$H(x,y) = \frac{G'_{\mathbf{P}}(1/x)G'_{\mathbf{P}}(1/y)}{x^2y^2(G_{\mathbf{P}}(1/x) - G_{\mathbf{P}}(1/y))^2} - \frac{1}{(x-y)^2}. \tag{8}$$

The proof of Lemma 1, which relies on the results from [13, 14], is omitted due to page limitation.

The limit eigenvalue distribution of $\mathbf{P}$ is the well-known Marchenko-Pastur distribution and its Cauchy transform is given by [5, eq. (14)] as

$$G_{\mathbf{P}}(z) = \sqrt{\frac{1}{4} - \frac{1+\zeta}{2z} + \frac{1-\zeta}{4z^2}} - \frac{1}{2} - \frac{1-\zeta}{2z}. \tag{9}$$

Furthermore, according to [5, Eq. (29)], the Cauchy transform of $\mathbf{Q}$ is uniquely determined by the cubic equation

$$\begin{aligned} & z^2\rho\zeta^2 G^3_{\mathbf{Q}}(z) - z\zeta(1+\rho-2\rho\zeta)G^2_{\mathbf{Q}}(z) \\ & \quad + \big((\rho\zeta-1)(\zeta-1) - z\big)G_{\mathbf{Q}}(z) - 1 = 0. \end{aligned} \tag{10}$$

It is noted that in [5] the implicit eigenvalue distribution function $F_{\mathbf{Q}}(x)$ of $\mathbf{Q}$ is numerically obtained by $G_{\mathbf{Q}}(z)$ using the inverse Cauchy transform.

By following the same procedures as those in [15], the asymptotic variance of $\mathcal{I}$ under the condition (5) is expressed as

$$\sigma^2_{\mathcal{I}} = -\frac{1}{4\pi^2}\oint\!\!\oint_{\mathcal{C}_x,\mathcal{C}_y} g(x)g(y)G_{\mathbf{Q}}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{11}$$

where the contours $\mathcal{C}_x$ and $\mathcal{C}_y$ are closed and taken in the positive direction in the complex plane, each enclosing the support of $F_{\mathbf{Q}}(x)$ but not the point $x = -1/\gamma$. Substituting (7) and (8) into (11) with the change of variables $t_1 = G_{\mathbf{P}}\big(-1/G_{\mathbf{Q}}(x)\big)$, $t_2 = G_{\mathbf{P}}\big(-1/G_{\mathbf{Q}}(y)\big)$, now (11) becomes

$$\sigma^2_{\mathcal{I}} = -\frac{1}{4\pi^2}\oint\!\!\oint_{\mathcal{C}_1,\mathcal{C}_2} \frac{g\big(h(t_1)\big)g\big(h(t_2)\big)}{(t_1-t_2)^2}\,\mathrm{d}t_1\,\mathrm{d}t_2. \tag{12}$$

Here, $h(t) = G^{-1}_{\mathbf{Q}}\big(-1/G^{-1}_{\mathbf{P}}(t)\big)$ and the inverse of $G_{\mathbf{P}}$ and $G_{\mathbf{Q}}$ can be solved using (9) and (10) as

$$G^{-1}_{\mathbf{P}}(t) = -\frac{1}{t} + \frac{\zeta}{1+t}, \tag{13}$$

$$\begin{aligned} G^{-1}_{\mathbf{Q}}(t) = {} & -\frac{\sqrt{1+2\zeta(1+\rho)t+(1-\rho)^2\zeta^2t^2}}{2\rho\zeta^2t^2} \\ & + \frac{1+\zeta(1+\rho-2\rho\zeta)t}{2\rho\zeta^2t^2}. \end{aligned} \tag{14}$$

It is difficult to further simplify the double integral (12). However, when the transmitter and receiver have equal number of antennas, i.e. $\rho = 1/\zeta$ in (5), we obtain an explicit expression for $\sigma^2_{\mathcal{I}}$. The results are summarized in the following proposition.

Proposition 1. *When $\rho = 1/\zeta$, the asymptotic variance of $\mathcal{I}$ is given by*

$$\sigma^2_{\mathcal{I}} = \log\frac{\omega_2\omega_3(\omega_1+1)^2}{(\omega_1-\omega_2)(\omega_1-\omega_3)}, \tag{15}$$

where

$$\omega_1 = -\frac{2}{3} - \frac{-1-3\gamma+3\gamma/\rho}{3u(\gamma)} + \frac{u(\gamma)}{3}, \tag{16}$$

$$\omega_2 = -\frac{2}{3} + \frac{2(-1-3\gamma+3\gamma/\rho)}{3(1-i\sqrt{3})u(\gamma)} - \frac{(1-i\sqrt{3})u(\gamma)}{6}, \tag{17}$$

$$\omega_3 = -\frac{2}{3} + \frac{2(-1-3\gamma+3\gamma/\rho)}{3(1+i\sqrt{3})u(\gamma)} - \frac{(1+i\sqrt{3})u(\gamma)}{6}, \tag{18}$$

and $u(\gamma)$ is given on top of the next page.

Proof: The proof of Proposition 1 is in the Appendix. ■

The asymptotic variance $\sigma^2_{\mathcal{I}}$ is obtained with the assumption that matrices dimensions are large. However, as will be shown in the next section, $\sigma^2_{\mathcal{I}}$ serves as a good approximation for the variance of mutual information even when the matrix

$$u(\gamma) = \left(1 + \frac{9\gamma}{2} + \frac{9\gamma}{\rho} + \sqrt{\left(\frac{3\gamma}{\rho} - 3\gamma - 1\right)^3 + \left(1 + \frac{9\gamma}{2} + \frac{9\gamma}{\rho}\right)^2}\right)^{1/3}$$

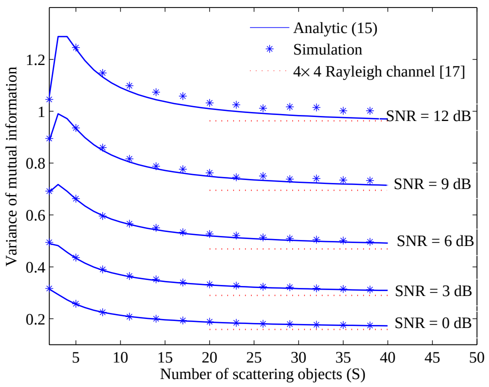


Fig. 1. Variance of mutual information when $T = R = 4$. Solid line: asymptotic variance calculated from (15); markers: simulation; dotted line: variance of the mutual information of a 4×4 Rayleigh channel [16].

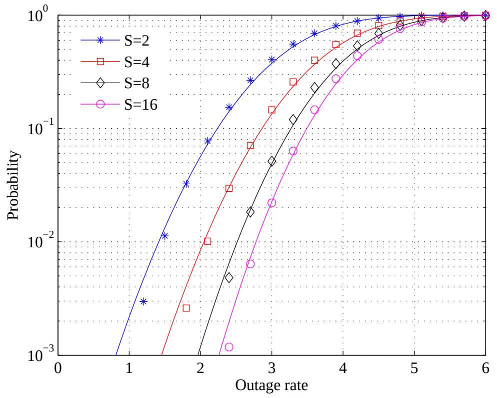


Fig. 2. Cumulative distribution of mutual information with SNR $\gamma = 5$ dB and $T = R = 4$. Solid line: Gaussian approximation; markers: simulation.

dimensions are small. As the matrix dimensions grow to infinity, the mutual information of MIMO Rayleigh channels are shown to be asymptotic Gaussian when the matrix entries are independent [16] and correlated [12, 17, 18]. We emphasize that the channel model considered in [12, 18] is different from (2), where both $\mathbf{\Psi}$ and $\mathbf{\Theta}$ are random. We are recently hinted by [19] that the Gaussian behavior of mutual information may be valid for a wide class of channel matrix ensembles, including the double-scattering channel. Motivated by this, we propose a Gaussian approximation to the distribution of mutual information with mean $\mu_{\mathcal{I}}$ given by [8, Coroll. 2] and variance $\sigma_{\mathcal{I}}^2$ given by (15). Namely, the outage probability can be approximated as $F_{\mathcal{I}}(x) \approx \Phi_{\mu_{\mathcal{I}},\sigma_{\mathcal{I}}^2}(x)$, where $\Phi_{\mu_{\mathcal{I}},\sigma_{\mathcal{I}}^2}(\cdot)$ denotes the distribution function of a Gaussian random variable with mean $\mu_{\mathcal{I}}$ and variance $\sigma_{\mathcal{I}}^2$.

IV. Numerical Results

A. Variance of mutual information

Fig. 1 depicts the variance of mutual information $\mathcal{I}$ as a function of the number of scatterers S. The transmitter and receiver are equipped with equal number of antennas $T = R = 4$. We compare the asymptotic variance (15) with Monte Carlo simulations for various SNR values. For comparison purposes, we also plot the variance of the mutual information for a 4×4 Rayleigh channel derived in [16]. From Fig. 1, we can see that the asymptotic variance achieves a good agreement with the simulations even for the applied small matrix dimensions. As the number of scatterers increases or equivalently the channel richness ρ increases, the variance decreases and approaches the limit set by the Rayleigh MIMO channel. This observation is in line with the result in [7], where a multi-keyhole channel converges to a Rayleigh MIMO channel for large number of scatterers.

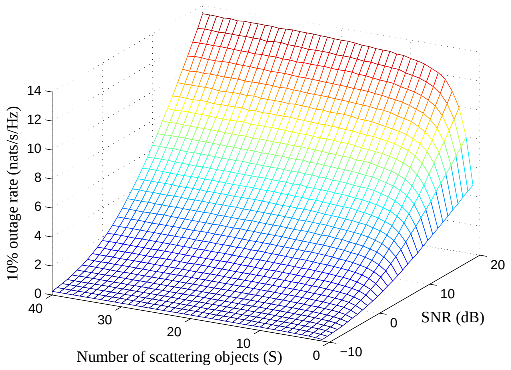


Fig. 3. 10% outage rate when $T = R = 4$.

B. Outage probability

Fig. 2 shows the distribution of mutual information with SNR $\gamma = 5$ dB and $T = R = 4$. We compare the Gaussian approximation against Monte Carlo simulations with $S = 2,\ 4,\ 8,$ and 16 scattering objects. In all cases, approximations show good agreement with simulations for the whole range of outage rate. Particularly, when outage probability is above 1%, approximations accurately predict the outage rate. In Fig. 3, we plot the 10% outage rate as a function of SNR γ and number of scatterers S. As expected, the outage rate is improved as the SNR increases. Meanwhile, when S is smaller than the number of antennas, the outage rate is improved dramatically since the additional scatterers increase the channel diversity. When S is larger than the number of antennas, outage rates are quickly saturated since in this case the channel diversity is limited by the number of antennas.

V. Conclusion

We considered the variance of mutual information in the double-scattering channel, when the channel state information is only available at the receiver. Under the assumption that the dimensions of channel matrices grow infinitely large, we derived the asymptotic variance of the mutual information. In particular, when the transmitter and receiver are equipped with equal number of antennas, we obtained an explicit closed-form expression. Based on the asymptotic variance, we proposed a Gaussian approximation to the channel outage probability. Numerical results show that the approximations are accurate even when the dimensions of channel matrices are small.

Appendix
Proof of Proposition 1

We choose the contour $\mathcal{C}_x$ located inside $\mathcal{C}_y$ such that both cross real-axis in the intervals $(-1/\gamma, 0)$ and (r, ∞). Here, r denotes the right end-point of the support of $F_{\mathbf{Q}}(x)$. Therefore, the transformed contours $\mathcal{C}_1$ and $\mathcal{C}_2$ in (12) cross the real-axis in the intervals $(h^{-1}(-1/\gamma), h^{-1}(0^-)) = (h^{-1}(-1/\gamma), \infty)$ and $(h^{-1}(r), h^{-1}(\infty)) = (h^{-1}(r), 0)$, where

$$h^{-1}(0^-) = \lim_{x\to 0-} h^{-1}(x), \quad h^{-1}(\infty) = \lim_{x\to\infty} h^{-1}(x).$$

Now we can rewrite the integration over $\mathcal{C}_1$ in (12) as

$$\begin{aligned}\mathcal{K}_{\text{inner}}(t_2) &= \frac{1}{2\pi i}\oint_{\mathcal{C}_1} \frac{\log(1+\gamma h(t))}{(t-t_2)^2}\,\mathrm{d}t \\ &= \frac{1}{2\pi i}\oint_{\mathcal{C}_1} \frac{\gamma h'(t)}{(t-t_2)(1+\gamma h(t))}\,\mathrm{d}t \\ &= \frac{\gamma}{2\pi i}\oint_{\mathcal{C}_1} \frac{(t+1)^{-1}\big(2(1-1/\rho)t^2+3t+1\big)}{t(t-t_2)(t-\omega_1)(t-\omega_2)(t-\omega_3)}\,\mathrm{d}t,\end{aligned} \tag{19}$$

where ω_i, $i = 1, 2, 3$, are the three roots of the cubic equation $t^3 + 2t^2 + (\gamma/\rho - \gamma + 1)t - \gamma = 0$ and solved in (16)-(18) via Cardano's formula.

Observing that $\omega_1 = h^{-1}(-1/\gamma)$, then the integrand of (19) has simple poles at $t = 0$ and $t = \omega_1$ within $\mathcal{C}_1$. After applying the residue theorem, the integral $\mathcal{K}_{\text{inner}}(t_2)$ becomes

$$\mathcal{K}_{\text{inner}}(t_2) = \frac{1}{t_2} - \frac{1}{t_2-\omega_1}. \tag{20}$$

Substituting (20) into (12), the variance $\sigma^2_{\mathcal{I}}$ can be therefore expressed as

$$\begin{aligned}\sigma^2_{\mathcal{I}} &= \frac{1}{2\pi i}\oint_{\mathcal{C}_2} \log(1+\gamma h(t))\left(\frac{1}{t} - \frac{1}{t-\omega_1}\right)\mathrm{d}t \\ &= \frac{1}{2\pi i}\oint_{\mathcal{C}_2} \log\frac{(t-\omega_2)(t-\omega_3)}{(t+1)^2}\left(\frac{1}{t} - \frac{1}{t-\omega_1}\right)\mathrm{d}t \\ &\quad + \frac{1}{2\pi i}\oint_{\mathcal{C}_2} \log\frac{t-\omega_1}{t}\left(\frac{1}{t} - \frac{1}{t-\omega_1}\right)\mathrm{d}t.\end{aligned} \tag{21}$$

The second integral in (21) has anti-derivative $\frac{1}{2}\left(\log\frac{t-\omega_1}{t}\right)^2$, which is single-valued over $\mathcal{C}_2$ and therefore vanishes due to Cauchy's theorem. Applying the residue theorem for the first integral of (21), we obtain (15).

Acknowledgment

We would like to thank O. Arizmendi and A. Soshnikov for helpful discussions. This work was supported by Academy of Finland (Grants 257708, 254299 and 251170), European Research Council (Grant 239784), German Research Foundation (Grants SP 419/8-1 and SP 419/9-1), Finnish Centre of Excellence in Computational Inference Research and Graduate School in Electronics Telecommunications and Automation.

References

[1] D. Gesbert, H. Bölcskei, D. A. Gore, and A. J. Paulraj, "Outdoor MIMO wireless channels: models and performance prediction," *IEEE Trans. Commun.*, vol. 50, no. 12, pp. 1926-1934, Dec. 2002.

[2] D. Chizhik, G. J. Foschini, M. J. Gans, and R. A. Valenzuela, "Keyholes, correlations, and capacities of multielement transmit and receive antennas," *IEEE Trans. Wireless Commun.*, vol. 1, no. 2, pp. 361-368, Feb. 2002.

[3] R. R. Müller and H. Hofstetter, "Confirmation of random matrix model for the antenna array channel by indoor measurements," in *Proc. IEEE Int. Symp. Antennas and Propagation Society*, vol. 1, 2001, pp. 472-475.

[4] S. Yang and J.-C. Belfiore, "Diversity-multiplexing tradeoff of double scattering MIMO channels," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2027-2034, Apr. 2011.

[5] R. R. Müller, "A random matrix model of communication via antenna arrays," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, Sept. 2002.

[6] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2636-2647, Oct. 2003.

[7] G. Levin and S. Loyka, "From multi-keyholes to measure of correlation and power imbalance in MIMO channels: outage capacity analysis," *IEEE Trans. Inform. Theory*, vol. 57, no. 6, pp. 3515-3529, June 2011.

[8] J. Hoydis, R. Couillet, and M. Debbah, "Asymptotic analysis of double-scattering channels," in *Proc. ASILOMAR'11*, Nov. 2011, pp. 1935-1939.

[9] L. Wei, Z. Zheng, O. Tirkkonen, and J. Hämäläinen, "On the ergodic mutual information of multiple cluster scattering MIMO channels," *IEEE Commun. Letters*, vol. 17, no. 9, pp. 1700-1703, Sept. 2013.

[10] G. Akemann, M. Kieburg, and L. Wei, "Singular value correlation functions for products of Wishart random matrices," *J. Phys. A: Math. Theor.*, vol. 46, no. 27, 2013.

[11] G. Akemann, J. Ipsen, and M. Kieburg, "Products of rectangular random matrices: singular values and progressive scattering," *Phys. Rev. E*, vol. 88, Nov. 2013.

[12] M. Debbah and R. R. Müller, "MIMO channel modeling and the principle of maximum entropy," *IEEE Trans. Inform. Theory*, vol. 51, no. 5, pp. 1667-1690, May 2005.

[13] B. Collins, J. Mingo, P. Śniady, and R. Speicher, "Second order freeness and fluctuations of random matrices III. Higher order freeness and free cumulants," *Documenta Math.*, vol. 12, pp. 1-70, 2007.

[14] O. Arizmendi and J. A. Mingo, "Second order even and R-diagonal operators," in preparation.

[15] Z. D. Bai and J. W. Silverstein, "CLT for linear spectral statistics of large-dimensional sample covariance matrices," *Ann. Prob.*, vol. 32, no. 1A, pp. 553-605, 2004.

[16] P. Smith and M. Shafi, "An approximation capacity distribution for MIMO systems," *IEEE Trans. Commu.*, vol. 52, no. 6, pp. 887-890, June 2004.

[17] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and noise: a (not so) large N analysis," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2545-2561, Oct. 2003.

[18] A. M. Tulino and S. Verdú, "Asymptotic outage capacity of multi-antenna channels," in *Proc. ICASSP'05*, Mar. 2005, pp. 825-828.

[19] Private communications with A. Soshnikov.

OFDM vs. Single Carrier Modulation — an Achievable Rate Perspective

Yair Carmon and Shlomo Shamai
Technion, Israel Institute of Technology
Email: {yairc@tx,sshlomo@ee}.technion.ac.il

Tsachy Weissman
Stanford University
Email: tsachy@stanford.edu

***Abstract*—We compare analytically the maximum achievable rates of reliable communication in single-carrier and OFDM modulation schemes, under the practical assumptions of i.i.d. finite alphabet inputs and linear ISI with additive Gaussian noise. Our results indicate that single-carrier schemes tend to offer a superior rate. In particular, it is shown that the Shamai-Laroia approximation for the single-carrier achievable rate is, under general conditions, an upper bound on the OFDM achievable rate. Information-Estimation relations and novel estimation-theoretic bounds are applied in order to rigorously establish these conditions.**

I. Introduction and Preliminaries

We consider a complex-valued, discrete-time intersymbol interference (ISI) channel model,

$$y_k = \sum_i h_i x_{k-i} + n_k \tag{1}$$

where $\{x_k\}$ is the channel input sequence, $\{h_0, ..., h_{L-1}\}$ are arbitrary complex-valued ISI taps and $\{n_k\}$ is a circularly symmetric white Gaussian process independent on the input, with $E|n_i|^2 = 1$. Let $H(\theta) = \sum_k h_k e^{-jk\theta}$ be the ISI channel transfer function. Throughout this paper we assume $Ex_i = 0$ and $E|x_i|^2 = 1$.

This model is relevant for a large variety of communication and data storage scenarios of practical importance [1]. Techniques of information transmission over ISI channels can be roughly divided into two types: single-carrier (SC) modulation and orthogonal frequency-division multiplexing (OFDM) modulation.

In SC modulation, every channel input x_k is a symbol drawn from a complex-valued alphabet also known as a signal constellation. Conventional constellations, such as BPSK and 16-QAM, are composed of 2^m regularly spaced values, each representing m data bits [1]. In this paper we assume that the input symbols form an i.i.d. process. This assumption tends to hold in practice, as channel coding schemes are typically designed for memoryless channels, and therefore induce i.i.d. input distributions. As discussed in [2], the assumption is further justified by the fact that our results carry over to the case where linear precoding of the input is allowed.

The maximum achievable rate for reliable communication under these assumptions is given by the input-output Average Mutual Information:

$$\mathcal{I}_{\text{SC}} \triangleq \lim_{k\to\infty} \frac{I(\{x_{-k}, ..., x_k\}\ ;\ \{y_{-k}, ..., y_k\})}{2k+1} \tag{2}$$

For general (non-Gaussian) input distributions, no closed-form expression for $\mathcal{I}_{\text{SC}}$ is known and it must be approximated either analytically or by Monte-Carlo simulations, see [3] and references therein.

Given a unit-variance RV ξ, let

$$I_\xi(\gamma) \triangleq I(\xi\ ;\ \sqrt{\gamma}\xi + \nu) \tag{3}$$

where ν is circularly symmetric Gaussian RV with $E|\nu|^2 = 1$ and independent of ξ, and γ stand for the SNR. A simple and often-used approximation for $\mathcal{I}_{\text{SC}}$ was first proposed by Shamai and Laroia [4],

$$\mathcal{I}_{\text{SC}} \approx \mathcal{I}_{\text{SL}} \triangleq I_x(\mathsf{SNR}_{\text{DFE}}) \tag{4}$$

where x is distributed as a single channel input x_k and

$$\mathsf{SNR}_{\text{DFE}} = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(1+|H(\theta)|^2\right)d\theta} - 1 \tag{5}$$

is the output SNR of the minimum mean-square error (MMSE) *unbiased* linear estimator of x_0 given $x_{-\infty}^{-1}$ and $y_{-\infty}^{\infty}$ [5]. The results of [3] indicate that $\mathcal{I}_{\text{SL}}$ is not always a lower bound on $\mathcal{I}_{\text{SC}}$. Nonetheless, extensive experimentation has shown that $\mathcal{I}_{\text{SL}}$ tightly lower bounds $\mathcal{I}_{\text{SC}}$ for essentially any ISI channel and SNR, as long as conventional input distributions are used [4].

In OFDM [6], information is transmitted in independent blocks of $N + N_{CP}$ channel inputs, where the first N_{CP} elements of each block constitute a "cyclic prefix" (CP) identical to the last N_{CP} elements of the block. The last N channel inputs in a block are given by the elements of $\mathbf{x} = \mathbf{W}^{-1}\tilde{\mathbf{x}}$, where $W_{m,k} = e^{-2\pi jmk/N}$ is the DFT matrix of order N and $\tilde{\mathbf{x}}$ is a column vector of N data symbols, commonly referred to as "subcarriers", as they each correspond to a different orthogonal carrier frequency.

The OFDM receiver discards the first N_{CP} channel outputs in each block. If the cyclic prefix is longer than the channel memory ($L < N_{CP}$), the remaining N channel outputs are described by $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ with $\mathbf{H}$ a square circulant matrix of order N, and $\mathbf{n}$ a white Gaussian noise vector. Letting $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$ yields the equivalent channel $\tilde{\mathbf{y}} = \mathbf{H}^{\mathbf{d}}\tilde{\mathbf{x}} + \tilde{\mathbf{n}}$, with $\mathbf{H}^{\mathbf{d}} = \mathbf{W}\mathbf{H}\mathbf{W}^{-1}$ a diagonal matrix and $\tilde{\mathbf{n}}$ distributed identically to $\mathbf{n}$. Thus, the ISI channel is transformed into N parallel memoryless channels.

As in SC modulation, we assume that all subcarriers (i.e. elements of $\tilde{\mathbf{x}}$) are i.i.d., zero-mean and unit power. This is usually the case in wireless links, where the communication overhead of coordinating different powers and constellations for different subcarriers often makes doing so undesirable. Under the scheme and assumptions described above, the maximum achievable rate is $\mathcal{I}^{(N)}_{\text{OFDM}} \triangleq \frac{1}{N+N_{CP}} \sum_{i=1}^{N} I_x\left(|H^d_{i,i}|^2\right)$, with $I_x(\cdot)$ as defined in (3) and x distributed as a single subcarrier $\tilde{x}_k$. Applying the Toeplitz Distribution Theorem (assuming $N_{CP} = o(N)$) we find that in the limit of large block size:

$$\mathcal{I}_{\text{OFDM}} \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} I_x\left(|H(\theta)|^2\right) d\theta \qquad (6)$$

Today, OFDM is the predominant modulation technique in high-bandwidth communications over channels with significant ISI, and is featured in a large number of standards. This is mainly due to the fact that the OFDM optimal receiver admits a low-complexity implementation using the FFT algorithm. However, since SC waveforms have a lower peak to average power ratio than OFDM, and due to the introduction of efficient frequency-domain decision feedback equalization techniques [7], SC schemes have become a viable alternative to OFDM in certain settings. It is therefore interesting to determine which of the two methods offers the higher rate of reliable communication, given optimal receivers.

This paper aims to answer the above question, under the assumption of a fixed i.i.d. input distribution. Our findings indicate that generally, $\mathcal{I}_{\text{OFDM}} \leq \mathcal{I}_{\text{SC}}$. In particular we show that for some common constellations, $\mathcal{I}_{\text{OFDM}} \leq \mathcal{I}_{\text{SL}}$ regardless of the ISI channel. For general inputs, we prove the same result in the low- and high-SNR regimes. We provide an exact characterization of the maximum value of $\mathcal{I}_{\text{OFDM}} - \mathcal{I}_{\text{SL}}$, and demonstrate numerically that it tends to be very small.

Our results stem from the concavity properties of the input-output mutual information in a scalar Gaussian channel, as a function of a rescaled SNR variable that we call the "log-SNR". In order to investigate these properties, we combine Information-Estimation results [8] with bounds on optimal estimation in the scalar Gaussian channel [2]. These bounds may be of independent interest.

There is previous work on the comparison of SC and OFDM from a fundamental limits perspective. In [9] the cut-off rates are compared analytically, and the SC rate is shown to exceed the OFDM rate in several scenarios. In [10] the achievable rates are compared numerically for particular inputs and ISI channels, and SC is found superior. In a recent report [11], the authors independently found that concavity with respect to log-SNR yields an inequality between the OFDM achievable rate and the Shamai-Laroia approximation. However, our central results remain exclusive to this work, including the analytic proofs of concavity, the study of the concave envelope and the application of Information-Estimation tools. Additional relevant literature appears in [2].

The rest of this paper is organized as follows. Section II introduces the key concavity concepts at the root of our results. Section III states our results precisely, and section IV briefly outlines their proofs.

II. Log-SNR and concavity properties

Central to our results is the study of the function

$$I^{\log}_x(\zeta) \triangleq I_x(e^{\zeta} - 1) \qquad (7)$$

with $I_x(\cdot)$ defined in (3), and x a zero-mean RV satisfying $E|x|^2 = 1$. Since setting $\zeta = \log(1+\gamma)$ yields $I^{\log}_x(\zeta) = I_x(\gamma)$, it is natural to interpret ζ as a "log-SNR" variable and measure it in units of information. For a complex Gaussian input g, we have $I^{\log}_g(\zeta) = \zeta$, and thus $I^{\log}_x(\zeta)$ is sub-linear in ζ for any other input distribution. For finite-entropy inputs, $I^{\log}_x(\zeta)$ is nearly linear for low ζ and nearly constant for high ζ, with the shoulder occurring at ζ values around the input entropy.

Let $\hat{I}_x^{\log}(\zeta)$ denote the concave envelope of $I_x^{\log}$, i.e.

$$\hat{I}_x^{\log}(\zeta) \triangleq \sup_{\substack{\zeta_1,\zeta_2 \text{ s.t.} \\ \zeta_1 \le \zeta \le \zeta_2}} \frac{(\zeta - \zeta_1) I_x^{\log}(\zeta_2) + (\zeta_2 - \zeta) I_x^{\log}(\zeta_1)}{\zeta_2 - \zeta_1} \tag{8}$$

Clearly, $\hat{I}_x^{\log} < \infty$, since $I_x^{\log}$ is increasing and sublinear. As the name implies, $\hat{I}_x^{\log} \ge I_x^{\log}$ is a concave function of ζ. Since $I_x^{\log}$ is real-analytic, $\hat{I}_x^{\log}$ is continuous and has a continuous derivative. We make the following additional definitions,

Definition 1. Let Δ_x denote that maximum difference between $I_x^{\log}$ and its concave envelope.

Definition 2. Let $\underline{\zeta}_0$ ($\bar{\zeta}_0$) be the maximal (minimal) ζ_0 for which $I_x^{\log}(\zeta)$ is concave for every $\zeta < \zeta_0$ ($\zeta \ge \zeta_0$). Similarly, Let $\underline{\zeta}_1$ ($\bar{\zeta}_2$) be the maximal (minimal) ζ_a for which $\hat{I}_x^{\log}(\zeta) = I_x^{\log}(\zeta)$ for every $\zeta < \zeta_a$ ($\zeta \ge \zeta_a$). Let $\underline{\gamma}_1$, $\underline{\gamma}_0$, $\bar{\gamma}_0$ and $\bar{\gamma}_2$ denote the corresponding linear-scale SNR values.

The following propositions characterize the quantities defined above,

Proposition 1. *If $\Delta_x \neq 0$ then $\underline{\gamma}_0 > 0$ and $\bar{\gamma}_0 < \infty$.*

Proof: Setting $\gamma = e^{\zeta} - 1$ and differentiating $I_x^{\log}$ twice, we find that

$$e^{-\zeta} I_x^{\log\prime\prime}(\zeta) = \text{mmse}_x(\gamma) + (1+\gamma)\,\text{mmse}'_x(\gamma) \tag{9}$$

where we have used the Guo-Shamai-Verdú Theorem [8] $I'_x(\gamma) = \text{mmse}_x(\gamma)$, with $\text{mmse}_x(\gamma) \triangleq E\left|x - E\left[x|\sqrt{\gamma}x + n\right]\right|^2$ the MMSE in estimating channel input x from its Gaussian noise corrupted version at SNR γ.

Let $d_{\min}$ denote the minimum distance between any two symbols in the input alphabet. By a standard probability of error upper bound , $\text{mmse}_x(\gamma) \le D^2 e^{-(d_{\min}/2)^2\gamma}$ for some $D > 0$. Moreover, it can be shown that $\text{mmse}'_x(\gamma) \le -Ce^{-(d_{\min}/2)^2\gamma}/\sqrt{\gamma}$ for sufficiently large γ and some $C > 0$. A derivation of this bound is given in [2]. Substituting the above bounds into (9), it is clear that $I_x^{\log\prime\prime}(\zeta) < 0$ for sufficiently large ζ and hence that $\bar{\gamma}_0 < \infty$.

The equality (9) may be rewritten as $e^{-\zeta} I_x^{\log\prime\prime}(\zeta) = r'_x(\gamma)$, where $r_x(\gamma) \triangleq (1+\gamma)\,\text{mmse}_x(\gamma)$ denotes the ratio between the MMSE's of the non-linear and linear optimal estimators of x given $\sqrt{\gamma}x + n$. Clearly, $r_x(\gamma) \le 1$, and $r_x(0) = 1$, so by continuity there must be a neighborhood of 0 in which r_x is decreasing and therefore $I_x^{\log}$ is concave, showing that $\underline{\gamma}_0 > 0$. ■

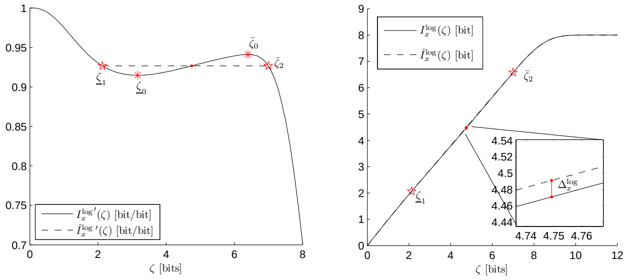


Figure 1. $I_x^{\log}$ and $\hat{I}_x^{\log}$ (right), and their derivatives with respect to ζ (left), for 256-QAM input.

Proposition 2. *If $\Delta_x \neq 0$ then $0 < \underline{\gamma}_1 < \underline{\gamma}_0$ and $\bar{\gamma}_0 < \bar{\gamma}_2 < \infty$.*

Proposition 2 follows straightforwardly from Proposition 1 and the properties of the concave envelope $\hat{I}_x^{\log}$. A detailed proof is provided in [2].

Figure 1 illustrates the quantities discussed in this section, for uniformly distributed 256-QAM input.

III. Statement of results

Our main result provides a connection between $\mathcal{I}_{\text{OFDM}}$ and $\mathcal{I}_{\text{SL}}$,

Theorem 1. *For any ISI channel and any i.i.d. input process with single-letter distribution x,*

$$\mathcal{I}_{\text{OFDM}} \le \mathcal{I}_{\text{SL}} + \Delta_x \tag{10}$$

Where Δ_x is given in Definition 1 . Additionally, $\mathcal{I}_{\text{OFDM}} \le \mathcal{I}_{\text{SL}}$ if the channel satisfies at least one of the following conditions:

1) $\text{SNR}_{DFE} \in [0, \underline{\gamma}_1] \cup [\bar{\gamma}_2, \infty)$
2) $|H(\theta)|^2 \le \underline{\gamma}_0$ *for every* $\theta \in (-\pi, \pi)$
3) $|H(\theta)|^2 \ge \bar{\gamma}_0$ *for every* $\theta \in (-\pi, \pi)$

where $\underline{\gamma}_1 \le \underline{\gamma}_0 \le \bar{\gamma}_0 \le \bar{\gamma}_2$ are given in Definition 2.

The following results sharpen Theorem 1 for specific input distributions,

Theorem 2. *For BPSK and QPSK constellations, $\mathcal{I}_{\text{OFDM}} \le \mathcal{I}_{\text{SL}}$ for every ISI channel.*

Theorem 3. *For M-PAM and square M^2-QAM constellations, $(d_{\min}/2)^2\,\bar{\gamma}_0 \le 1$, where $d_{\min}$ is the minimum distance between input symbols.*

Further numerical study indicates that $\Delta_x = 0$ for 4-PAM, 8-PSK, 16-QAM and 32-QAM inputs, extending Theorem 2. Table I lists the quantities that were presented in Theorem 1 for inputs with $\Delta_x > 0$.

Table I
SUMMARY OF THE QUANTITIES APPEARING IN THEOREM 1

	$(d_{\min}/2)^2 \cdot \{\underline{\gamma}_1, \underline{\gamma}_0, \bar{\gamma}_0, \bar{\gamma}_2\}$ [dB]				Δ_x [bits]
64-QAM	-5.54	-5.25	-4.49	-4.23	$1.86{\cdot}10^{-6}$
256-QAM	-17.0	-13.3	-3.09	-1.27	0.0202
1024-QAM	-24.7	-19.5	-3.17	-0.851	0.0585
4096-QAM	-31.8	-25.6	-3.40	-0.739	0.0987

Note: All values are rounded to three significant digits.

It is seen that Δ_x is quite small even in very high-order constellations such as 4096-QAM. Additionally, Table I indicates that the general bound provided by Theorem 3 is slack by an approximate factor of 2 — i.e., for higher-order QAM, $(d_{\min}/2)^2 \bar{\gamma}_0 \approx 1/2$.

IV. OUTLINE OF PROOFS

Proof of Theorem 1: From the definitions of $\mathcal{I}_{\mathrm{SL}}$, $\mathcal{I}_{\mathrm{OFDM}}$, $I_x^{\log}$, $\hat{I}_x^{\log}$ and Δ_x:

$$\begin{aligned}\mathcal{I}_{\mathrm{OFDM}} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} I_x^{\log}\left(\log\left(1+|H(\theta)|^2\right)\right)d\theta \\ &\leq \hat{I}_x^{\log}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(1+|H(\theta)|^2\right)d\theta\right) \\ &= \hat{I}_x^{\log}\left(\log\left(1+\mathsf{SNR}_{\mathsf{DFE}}\right)\right) \leq \mathcal{I}_{\mathrm{SL}} + \Delta_x \qquad (11)\end{aligned}$$

where in moving to the second we used $I_x^{\log} \leq \hat{I}_x^{\log}$ and then invoked Jensen's inequality based on the concavity of $\hat{I}_x^{\log}$. With $\underline{\gamma}_1$ and $\bar{\gamma}_2$ according to Definition 2, it is clear that if condition 1 holds, then

$$\hat{I}_x^{\log}\left(\log\left(1+\mathsf{SNR}_{\mathsf{DFE}}\right)\right) = \mathcal{I}_{\mathrm{SL}} \qquad (12)$$

and therefore $\mathcal{I}_{\mathrm{OFDM}} \leq \mathcal{I}_{\mathrm{SL}}$. Moreover, if either condition 2 or conditions 3 hold, then $I_x^{\log}$ is a concave function for every value of $|H(\theta)|^2$, and we may therefore exchange $\hat{I}_x^{\log}$ with $I_x^{\log}$ in (11), yielding $\mathcal{I}_{\mathrm{OFDM}} \leq \mathcal{I}_{\mathrm{SL}}$ once more. ■

Proof of Theorem 2: Let b and q be unit-power BPSK and QPSK inputs, respectively. In [2] it is shown that $\mathrm{mmse}_b(\gamma) \leq e^{-\gamma}$ and $\mathrm{mmse}'_b(\gamma) \leq -\frac{2e^{-\gamma}}{\sqrt{1+6\gamma}}$. Substituting these bounds into (9), we find that $I_b^{\log}$ is concave. Using $\mathrm{mmse}_q(\gamma) = \mathrm{mmse}_b(\gamma/2)$ and repeating the derivation used for BPSK proves that $I_q^{\log}(\zeta)$ is concave for $\zeta \geq e-1$. Switching to the linear estimator upper bound $\mathrm{mmse}_q(\gamma) \leq (1+\gamma)^{-1}$ shows that $I_q^{\log}(\zeta)$ is concave for $\zeta < e-1$, completing the proof. ■

Proof of Theorem 3: Let b and m be unit-power BPSK and M-PAM inputs, respectively. In [2] the following bounds are derived,

$$\mathrm{mmse}_m(\gamma) \leq \mu\frac{d_{\min}^2}{4}\left[\mathrm{mmse}_b(\rho) + \bar{B}(\rho)\right] \qquad (13)$$

$$\mathrm{mmse}'_m(\gamma) \leq \mu\frac{d_{\min}^4}{16}\left[\mathrm{mmse}'_b(\rho) + \bar{C}(\rho)\right] \qquad (14)$$

with $\mu = 2(M-1)/M$, $\rho = (d_{\min}/2)^2\gamma$, $\bar{C}(\gamma) = 32e^{8\gamma}Q\left(\sqrt{32\gamma}\right)$ and $\bar{B}(\gamma) = 16Q\left(\sqrt{8\gamma}\right) + 4\sum_{k=2}^{\infty}(2k+1)Q\left(k\sqrt{8\gamma}\right)$, where $Q(\cdot)$ is the error function. These bounds are based on novel "point-wise" bounds for estimation of M-ary PAM inputs in Gaussian noise [2]. Substituting (13) and (14) into (9), it follows that $(d_{\min}/2)^2\underline{\gamma}_0 \leq 1$ for any M-PAM and M^2-QAM input. ■

ACKNOWLEDGMENT

This work was supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM#.

REFERENCES

[1] J.G. Proakis. *Digital communications*, volume 1221. McGraw-hill, 1987.

[2] Yair Carmon, Shlomo Shamai, and Tsachy Weissman. Comparison of the achievable rates in OFDM and single carrier modulation with i.i.d. inputs. *arXiv preprint arXiv:1306.5781*, 2013.

[3] Y. Carmon, S. Shamai, and T. Weissman. Lower bounds and approximations for the I.I.D. achievable rate in the intersymbol interference channel. In preparation.

[4] S. Shamai and R. Laroia. The intersymbol interference channel: Lower bounds on capacity and channel precoding loss. *Information Theory, IEEE Transactions on*, 42(5):1388–1404, 1996.

[5] J.M. Cioffi, G.P. Dudevoir, M. Vedat Eyuboglu, and G.D. Forney Jr. MMSE decision-feedback equalizers and coding I: Equalization results. *Communications, IEEE Transactions on*, 43(10):2582–2594, 1995.

[6] Junyi Li, Xinzhou Wu, and Rajiv Laroia. *OFDMA Mobile Broadband Communications*. Cambdige University Press, 2013.

[7] Nevio Benvenuto, Rui Dinis, David Falconer, and Stefano Tomasin. Single carrier modulation with nonlinear frequency domain equalization: an idea whose time has come — again. *Proceedings of the IEEE*, 98(1):69–96, 2010.

[8] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *Information Theory, IEEE Transactions on*, 51(4):1261–1282, 2005.

[9] Amanda de Paula and Cristiano Panazio. A comparison between OFDM and single-carrier with cyclic prefix using channel coding and frequency-selective block fading channels.

[10] M Franceschini, R Pighi, G Ferrari, and R Raheli. On information theoretic aspects of single-and multi-carrier communications. In *Information Theory and Applications Workshop, 2008*, pages 94–99. IEEE, 2008.

[11] Amanda de Paula and Cristiano Panazio. Comparison of OFDM and SC-DFE capacities without channel knowledge at the transmitter. *arXiv preprint arXiv:1306.3440*, 2013.

Noncoherent Decision-Feedback Equalization in Massive MIMO Systems

Robert F.H. Fischer[1] and Melanie Bense[2]

[1]Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany, Email: robert.fischer@uni-ulm.de
[2]Lehrstuhl für Informationsübertragung, Universität Erlangen-Nürnberg, Erlangen, Germany, Email: bense@lnt.de

Abstract—**In this paper, a noncoherent approach to decision-feedback equalization (DFE) in multi-user massive MIMO systems is presented. Thereby, the contradicting principles of DFE, where interference of already detected symbols is canceled using actual channel knowledge, and noncoherent reception, where the symbols are detected without any channel-state information, are combined. Based on an analysis of the statistics of the interference terms in autocorrelation-based noncoherent receivers, DFE is proposed and optimized. In combination with decision-feedback differential detection of the individual users, a low-complexity high-performance scheme is established.**

I. Introduction

Multiple-input/multiple-output (MIMO) systems, where the base station is equipped with a very large number of receive antennas, so-called *massive MIMO*, gain more and more attention, e.g., [8], [9]. As the number of channel coefficients is extremely large, one of the main challenges is to acquire accurate channel estimates. One solution to overcome the problem of requiring a large amount of pilot symbols and the need to perform channel estimation is to resort to *noncoherent detection*.

In [11], based on the similarities between the huge number of *temporal* echos in ultra-wideband (UWB) systems and the huge number of *spatial* “echos” in the present setting, noncoherent detection schemes for massive MIMO have been proposed and optimized. In order to overcome the poor performance of conventional differential detection, block-wise joint processing based on the idea of *multiple-symbol differential detection (MSDD)* [2] may be applied. Of special interest are low-complexity but well-performing methods, in particular *decision-feedback differential detection (DFDD)* [12]. Here, the decision-feedback principle—use already available decisions for the equalization/detection of the other symbols—is applied over a temporal block for one particular user.

In multi-user systems, *(sorted) decision-feedback equalization (DFE)* over the users (aka successive interference cancellation) is very attractive; in the context of MIMO systems sorted DFE is known as *Bell Labs Layered Space-Time (BLAST)* [4]. However, on first glance, DFE—cancel interference of already detected symbols using actual channel knowledge—and noncoherent reception—detect symbols without actual, but only based on statistical channel-state information—contradict each other.

In this paper, we present a *noncoherent* approach to DFE over the users for use in massive MIMO systems. We analyze the statistics of the interference and noise terms when using autocorrelation receivers. Employing these insights, DFE based on *statistical channel knowledge* is proposed and optimized. As in H-BLAST [4], on the one hand, the users are successively detected; an optimum decision order is derived. On the other hand, the detection of each user is done block-wise over a temporal block (transmission burst) utilizing DFDD. Hence, two DFE procedures (temporal/spatial) are combined to establish low-complexity high-performance noncoherent multi-user detection schemes.

The paper is organized as follows. In Sec. II, the system model is introduced and noncoherent (single-user) detection is briefly reviewed. Sec. III studies the interference in autocorrelation-based detectors and derives noncoherent DFE. Numerical results on the performance of noncoherent DFE are presented in Sec. IV. Sec. V gives some conclusions.

II. System Model and Noncoherent Detection

Throughout the paper we consider the *multi-user uplink scenario* depicted in Fig. 1. N_u users, each equipped with a single transmit antenna, transmit to a central base station equipped with a very large number, $N_\mathsf{rx} \gg 1$, of receive antennas.

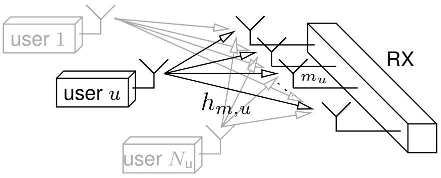


Fig. 1. Illustration of the multi-user massive MIMO uplink system.

In each time step (discrete index k), user u transmits symbols $b_{k,u}$ drawn from an M-ary PSK constellation $\mathcal{M} \overset{\text{def}}{=} \left\{\mathrm{e}^{\mathrm{j}2\pi\cdot i/M} \mid i = 0,1,\ldots,M-1\right\}$. In view of the noncoherent reception based on the *autocorrelation principle*, *differential encoding* is performed at the transmitter, i.e., the transmit symbols are computed via $b_{u,k} = a_{u,k} b_{u,k-1}$, $b_{u,0} = 1$, where the information symbols $a_{k,u}$ to be transmitted are drawn from the M-ary PSK constellation, too.

A. Massive MIMO Channel Model

We assume flat-fading channels—coefficients $h_{m,u}$ in complex baseband notation—from user u to receive antenna m. They are characterized by the user-specific *power-space profile (PSP)*, which describes the average receive power induced by user u at receive antenna m. If the users are placed in front of a uniform linear antenna array (antenna spacing d_a) and a pure path loss model (exponent γ) is studied, the PSP is given by (for details see [11])

$$P_{m,u} \overset{\text{def}}{=} \mathrm{E}\{|h_{m,u}|^2\} = \text{const.} \cdot \mathrm{e}^{-\frac{|m-m_u|^2}{2\varsigma_u^2}} \,, \qquad (1)$$

where m_u is the antenna element closest (distance d_u) to user u and $\zeta_u^2 = d_u^2/(d_{\mathsf{a}}^2\gamma)$. The channel coefficients $h_{m,u}$ are then randomly drawn from a zero-mean, circular-symmetric complex Gaussian distribution, independently of each other, with a variance according to the power-space profile and are expected to be constant during a transmission burst.

Since the application of *multiple-symbol differential detection (MSDD)* [2] improves the performance of noncoherent receivers significantly, we consider the received signal over the entire burst of N time steps.[1] In vector/matrix notation, the block of all N_{rx} receive signals over N time steps ($N_{\mathsf{rx}} \times N$ matrix) can be written as[2]

$$\boldsymbol{R} = \sum\nolimits_{u=1}^{N_{\mathsf{u}}} \boldsymbol{h}_u \boldsymbol{b}_u + \boldsymbol{N} \;, \tag{2}$$

where $\boldsymbol{h}_u \overset{\text{def}}{=} [h_{1,u}, \ldots, h_{N_{\mathsf{rx}},u}]^{\mathsf{T}}$ collects the channel coefficients for user u and $\boldsymbol{b}_u \overset{\text{def}}{=} \big[b_{0,u},\, b_{1,u},\, \ldots,\, b_{N-1,u}\big]$ contains the transmit symbols of user u. The noise matrix $\boldsymbol{N} \overset{\text{def}}{=} [\, n_{m,k}\,]$ gathers the circular-symmetric (zero-mean) complex Gaussian noise $n_{m,k}$ with variance σ_{n}^2.

B. Noncoherent Detection

It is well known, e.g., [5], [10], that the differential detection of the symbols of user u can be based on the $N \times N$ *correlation matrix*

$$\boldsymbol{Z}_u \overset{\text{def}}{=} \boldsymbol{R}^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{R} \;, \tag{3}$$

with a user-specific diagonal weighting matrix $\boldsymbol{W}_u$. In view of the PSP (1), a suited choice is [11]

$$\boldsymbol{W}_u \overset{\text{def}}{=} \mathbf{diag}(w_{1,u}, \ldots, w_{N_{\mathsf{rx}},u}) \;, \quad w_{m,u} = \mathrm{e}^{-\frac{|m-m_u|^2}{2\zeta_{\mathsf{w},u}^2}} \;, \tag{4}$$

where the parameter $\zeta_{\mathsf{w},u}$ may be optimized for each user u individually.

Based on the correlation matrix, the optimum (block-wise, MSDD) detection scheme for user u calculates [6], [11]

$$\hat{\boldsymbol{b}}_u^{\mathsf{MSDD}} = \underset{\bar{\boldsymbol{b}} \in \mathcal{M}^N,\, \bar{b}_0 = 1}{\operatorname{argmax}} \; \bar{\boldsymbol{b}} \boldsymbol{Z}_u \bar{\boldsymbol{b}}^{\mathsf{H}} \;. \tag{5}$$

For $N = 2$ the simplest noncoherent detection scheme—*symbol-wise differential detection*—results.

The high computational complexity of MSDD can significantly be reduced with only a marginal loss by applying the principle of *decision-feedback differential detection* [1], [3], [12]. In [11] this strategy has been adapted to the massive MIMO setting and an optimum decision order within the temporal block (based on the actual receive symbols) has been derived. A pseudo-code description of sorted DFDD (operating over the temporal dimension) is given in [11, Fig. 3].

In the final step, estimates for the information-carrying symbols are calculated via $\hat{a}_k = \hat{b}_k \cdot \hat{b}_{k-1}^*$, $k = 1, \ldots, N$.

III. Noncoherent Decision-Feedback Equalization

Employing DFDD, the detection is done for each user individually (in parallel) over a temporal block of N symbols. Thereby, the *decision-feedback* principle is utilized in *temporal direction*. Since a central receiver is assumed, the feedback of decisions may also be applied *over the users*, cf. Fig. 2.

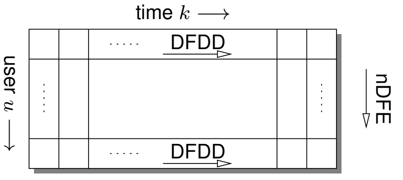


Fig. 2. Illustration of the dimensions over which DFDD and nDFE operate.

However, classical DFE requires channel knowledge for the coherent subtraction of the interference of the already decided symbols. Subsequently, we discuss how known symbols of already detected users may be used in the present correlation-based differential detection of not yet detected users. An optimized sorting over the users for this *noncoherent DFE (nDFE)* is derived.

A. Correlation Matrix and Interference Terms

Assuming a system with N_{u} users, the receive matrix is given in (2). Since $\boldsymbol{h}_\nu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\mu$ is scalar, the correlation matrix (3) for the detection of a particular user u calculates to

$$\begin{aligned}
\boldsymbol{Z}_u &= \big(\sum\nolimits_{\nu=1}^{N_{\mathsf{u}}} \boldsymbol{b}_\nu^{\mathsf{H}} \boldsymbol{h}_\nu^{\mathsf{H}} + \boldsymbol{N}^{\mathsf{H}}\big) \boldsymbol{W}_u \big(\sum\nolimits_{\nu=1}^{N_{\mathsf{u}}} \boldsymbol{h}_\nu \boldsymbol{b}_\nu + \boldsymbol{N}\big) && (6)\\
&= \boldsymbol{h}_u^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_u \, \boldsymbol{b}_u^{\mathsf{H}} \boldsymbol{b}_u && \text{(i)}\\
&\quad + \sum\nolimits_{\substack{\nu=1\\ \nu \neq u}}^{N_{\mathsf{u}}} \boldsymbol{h}_\nu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\nu \, \boldsymbol{b}_\nu^{\mathsf{H}} \boldsymbol{b}_\nu && \text{(ii)}\\
&\quad + \sum\nolimits_{\substack{\nu,\mu=1\\ \nu < \mu}}^{N_{\mathsf{u}}} \Big(\boldsymbol{h}_\nu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\mu \, \boldsymbol{b}_\mu \boldsymbol{b}_\nu^{\mathsf{H}} + \boldsymbol{h}_\mu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\nu \, \boldsymbol{b}_\nu \boldsymbol{b}_\mu^{\mathsf{H}} \Big) && \text{(iii)}\\
&\quad + \sum\nolimits_{\nu=1}^{N_{\mathsf{u}}} \boldsymbol{b}_\nu^{\mathsf{H}} \boldsymbol{h}_\nu^{\mathsf{H}} \, \boldsymbol{W}_u \boldsymbol{N} + \boldsymbol{N}^{\mathsf{H}} \boldsymbol{W}_u \sum\nolimits_{\nu=1}^{N_{\mathsf{u}}} \boldsymbol{b}_\nu^{\mathsf{H}} \boldsymbol{h}_\nu^{\mathsf{H}} && \text{(iv)}\\
&\quad + \boldsymbol{N}^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{N} \;.
\end{aligned}$$

Term (i) contains the desired correlation coefficients for the detection of user u; (ii) and (iii) are interferences due to the other users, and (iv) are "noise × signal" and "noise × noise" terms.

For the further analysis, the statistics of the interference and noise terms have to be known (cf. also [10]). The quadratic form in (i), (ii) calculates to

$$\xi_{u,\nu} \overset{\text{def}}{=} \boldsymbol{h}_\nu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\nu = \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u} \, |h_{m,\nu}|^2 \;. \tag{7}$$

Since $h_{m,\nu} \sim \mathcal{CN}(0, P_{m,\nu})$, the sum $\xi_{u,\nu}$ is approximately real-valued Gaussian[3] with mean and variance (cf. [7])

$$\eta_{u,\nu} \overset{\text{def}}{=} \mathrm{E}\{\xi_{u,\nu}\} = \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u} P_{m,\nu} \;, \tag{8}$$

$$\sigma_{u,\nu}^2 \overset{\text{def}}{=} \mathrm{E}\{(\xi_{u,\nu} - \eta_{u,\nu})^2\} = \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u}^2 P_{m,\nu}^2 \;. \tag{9}$$

The terms in (iii) are given by $\boldsymbol{\Xi}_{u,\nu,\mu} \overset{\text{def}}{=} \xi_{u,\nu,\mu} \, \boldsymbol{b}_\mu \boldsymbol{b}_\nu^{\mathsf{H}} + \xi_{u,\nu,\mu}^* \, \boldsymbol{b}_\nu \boldsymbol{b}_\mu^{\mathsf{H}}$, where the definition

$$\xi_{u,\nu,\mu} \overset{\text{def}}{=} \boldsymbol{h}_\nu^{\mathsf{H}} \boldsymbol{W}_u \boldsymbol{h}_\mu = \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u} \, h_{m,\nu}^* h_{m,\mu} \tag{10}$$

[1] W.l.o.g. we consider the block with time indices $k = 0, \ldots, N-1$.

[2] Note: $\boldsymbol{h}_u$ is a *column* vector over the receive antennas, whereas $\boldsymbol{b}_u$ is a *row* vector over the time.

[3] Due to the individual scaling and the different powers of $h_{m,\nu}$, the quantity $\xi_{u,\nu}$ is *not* χ^2 distributed with $2N_{\mathsf{rx}}$ degrees of freedom.

has been used. $\xi_{u,\nu,\mu}$ is the sum over products of independent, zero-mean complex Gaussians and is well approximated by a zero-mean complex Gaussian distribution. Moreover, since the elements of $\boldsymbol{b}_\nu^{\mathsf{H}}\boldsymbol{b}_\mu$ are drawn from the M-PSK set $\mathcal{M}$, the entries of the hermitian matrix $\Xi_{u,\nu,\mu}$ have the form $\mathrm{e}^{\mathrm{j}\frac{2\pi}{M}l'}\big(\xi + \xi^* \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}l}\big)$. The term in brackets lies on a line in the complex plane with direction $\mathrm{e}^{\mathrm{j}\frac{\pi}{M}l}$, hence the elements of $\Xi_{u,\nu,\mu}$ are real-valued Gaussian distributed along the lines with direction $\mathrm{e}^{\mathrm{j}\frac{\pi}{M}l}$, $l = 0, \ldots, M-1$, and variance

$$\sigma^2_{u,\nu,\mu} \overset{\text{def}}{=} 2 \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w^2_{m,u} P_{m,\nu} P_{m,\mu} \,. \tag{11}$$

Finally, the elements of the sum of noise terms (iv) are zero-mean complex Gaussian distributed with variance

$$\sigma^2_{\mathsf{n},u} \overset{\text{def}}{=} 2\sigma^2_{\mathsf{n}} \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u} \sum\nolimits_{\nu=1}^{N_{\mathsf{u}}} P_{m,\nu} + \sigma^4_{\mathsf{n}} \sum\nolimits_{m=1}^{N_{\mathsf{rx}}} w_{m,u} \,. \tag{12}$$

B. Decision-Feedback Equalization

The main principle of decision-feedback equalization (successive interference cancellation) is the subtraction of the interference caused by already detected users. To this end, both, the data symbols of the users and the channel coefficients, via which the respective data symbols interfere, have to be known.

In the present situation of noncoherent detection, the channel coefficients are not known—only their statistics is available via the PSP. However, with this knowledge and assuming a given receive weighting $\boldsymbol{W}_u$, the statistics of the interference terms (ii) and (iii) can be calculated. All terms in (iii) are zero-mean Gaussian and no knowledge can be exploited. However, the means $\eta_{u,\nu}$ of the terms in (ii) contributed by the already detected users can be utilized.

When $\mathcal{D}$ denotes the index set of the already detected users and $\hat{\boldsymbol{b}}_\nu$ is the estimated vector of differentially encoded symbols of user ν, the correlation matrix for detection of user u hence should be calculated according to

$$\boldsymbol{Z}'_u = \boldsymbol{Z}_u - \sum\nolimits_{\nu\in\mathcal{D}} \eta_{u,\nu} \cdot \hat{\boldsymbol{b}}_\nu^{\mathsf{H}} \hat{\boldsymbol{b}}_\nu \,, \tag{13}$$

where $\boldsymbol{Z}_u$ is the conventional correlation matrix (3).

C. Optimization and Sorting

We assume that the data vectors $\boldsymbol{b}_\nu$ of users $\nu \in \mathcal{D}$ are already (error-freely) detected and that the receiver knows the PSPs of all user.[4] Then it is able to calculate the *signal-to-noise-plus-interference ratio (SINR)* of user $u \notin \mathcal{D}$ for correlation-based detection, when the mean interference term in the correlation matrix is canceled according to (13). Combining the above results it reads

$$\mathrm{SINR}_u = \frac{\eta^2_{u,u} + \sigma^2_{u,u}}{\sum\limits_{\nu\neq u} \sigma^2_{u,\nu} + \sum\limits_{\substack{\nu\notin\mathcal{D}\\ \nu\neq u}} \eta^2_{u,\nu} + \sum\limits_{\nu<\mu} \sigma^2_{u,\nu,\mu} + \sigma^2_{\mathsf{n},u}} \,. \tag{14}$$

Based on this analytic solutions two optimization tasks are carried out: On the one hand, as in the BLAST sorting strategy [4], in each step the user with the highest SINR should be detected (greedy approach). In a successive way—calculating the SINR of all not yet detected users, thereby taking the interference reduction due to the already detected ones into account and deciding for the best one—the optimum decision order is found.

Algorithm 1 Pseudocode of sorted noncoherent DFE.

```
$\hat{\boldsymbol{B}} = \mathtt{nDFE}(\boldsymbol{R}, \boldsymbol{P}, \sigma^2_{\mathsf{n}})$
1: $\mathcal{D} := \{\}$,  $\overline{\mathcal{D}} := \{1, \ldots, N_{\mathsf{u}}\}$
2: while $|\overline{\mathcal{D}}| > 0$ {
3:     $[\mathrm{SINR}_u,\ \zeta_{\mathsf{w},u}] = \mathtt{optSINR}(u, \boldsymbol{P}, \sigma^2_{\mathsf{n}}, \mathcal{D})$,  $u \in \bar{\mathcal{D}}$
4:     $\check{u} = \mathrm{argmax}_{u\in\bar{\mathcal{D}}}\, \mathrm{SINR}_u$
5:     set $\boldsymbol{W}_{\check{u}}$ acc. to (4) with $\zeta_{\mathsf{w},\check{u}}$
6:     $\boldsymbol{Z}'_{\check{u}} := \boldsymbol{R}^{\mathsf{H}} \boldsymbol{W}_{\check{u}} \boldsymbol{R} - \sum_{\nu\in\mathcal{D}} \eta_{\check{u},\nu} \cdot \hat{\boldsymbol{b}}_\nu^{\mathsf{H}} \hat{\boldsymbol{b}}_\nu$
7:     $\hat{\boldsymbol{b}}_{\check{u}} := \mathtt{DFDD}(\boldsymbol{Z}'_{\check{u}})$
8:     $\mathcal{D} := \mathcal{D} \cup \{\check{u}\}$, $\overline{\mathcal{D}} := \overline{\mathcal{D}} \setminus \{\check{u}\}$
9: }

$[\mathrm{SINR}_u,\ \zeta_{\mathsf{w},u}] = \mathtt{optSINR}(u, \boldsymbol{P}, \sigma^2_{\mathsf{n}}, \mathcal{D})$
1: $[\mathrm{SINR}_u,\ \zeta_{\mathsf{w},u}] = \max / \mathrm{argmax}_{\zeta_{\mathsf{w},u}}\, \mathrm{SINR}_u$,  acc. to (14)
```

Please note, as in H-BLAST [4], the users are successively detected; the sorting is based on the SINR (14) which is derived from statistical channel knowledge. For each user, block-wise detection over the temporal dimension is used (the entire block is decided). When, particularly, applying DFDD, for each user the detection order over this temporal block is optimized individually (based on the actual receive symbols). Hence, two DFE procedures (temporal DFDD; nDFE over the users) with two different (and separated) sorting strategies are present.

On the other hand, as all interference terms in (14) depend on the weighting matrix $\boldsymbol{W}_u$, this matrix can be optimized for each user and each detection step individually. Choosing $\boldsymbol{W}_u$ according to (4), the only free parameter is $\zeta_{\mathsf{w},u}$. Via numerical optimization the SINR-maximizing value can be found.

In summary, based on (13), noncoherent DFE can be performed. The optimum receive weighting and the optimum decision order can be determined based on the SINRs (14). Noteworthy, this optimization depends only on the PSPs of the users (statistical channel knowledge) and has only to be carried out when the configuration changes. Given the interference-reduced correlation matrix $\boldsymbol{Z}'_u$ (13), estimates on the data symbols of user u are obtained by DFDD. However, in principle, any noncoherent detection scheme utilizing the correlation matrix can be employed at this step.

Alg. 1 summarizes this procedure—$\boldsymbol{P}$ is a matrix containing all PSPs; $\hat{\boldsymbol{B}}$ row-wise contains the estimated transmit symbol vectors $\hat{\boldsymbol{b}}_u$. In Line 7 the subroutine for detecting the respective user is denoted as $\hat{\boldsymbol{b}}_u := \mathtt{DFDD}(\boldsymbol{Z}'_u)$. A pseudocode description of this DFDD algorithm (with inherent optimal sorting over the temporal dimension, cf. Fig 2) can be found in [11, Fig. 3]. In Line 4, based on the SINRs of the not yet detected users, the optimum decision order over the users (vertical dimension in Fig 2) is determined.

IV. Numerical Results

The performance of the proposed scheme is now assessed by means of numerical simulations. To this end, a uniform linear array with $N_{\mathsf{rx}} = 100$ antennas is expected, e.g., covering a

[4] At the receiver side, only knowledge of the PSP is expected—not actual channel knowledge, $h_{m,u}$, but only statistical one is assumed. The estimation of the PSP is much easier than that of the actual channel coefficients.

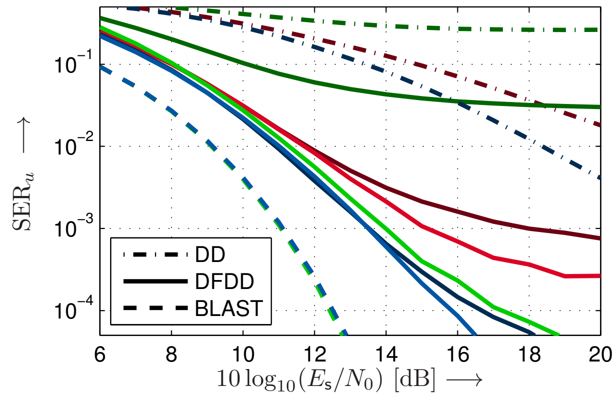


Fig. 3. Symbol error rate vs. E_s/N_0 (in dB). $N_\mathsf{u} = 3$; $m_1 = 20$ (🟥), $m_2 = 50$ (🟩), $m_3 = 85$ (🟦); colors corresponding to users. Uniform linear array with $N_\mathsf{rx} = 100$ antenna elements. Power-space profile according to (1), $\zeta_u = 20$, $\forall u$. Burst length $N = 200$. $M = 4$-ary DPSK.

long hallway. Transmission bursts of length $N = 200$ (equal to the DFDD block) are used; the block-fading channel changes after each burst. The results are averaged over 250 000 channel realizations. $N_\mathsf{u} = 3$ users are active—one is located at $m_1 = 20$ (marked 🟥), the next at $m_2 = 50$ (🟩), and the last at $m_3 = 85$ (🟦). The power-space profile (1) with $\zeta_u = 20$, $\forall u$, is used; it is normalized to $\sum_{m=1}^{N_\mathsf{rx}} P_{m,u} = 1$, $\forall u$. All users employ quaternary ($M = 4$) DPSK.

In Fig. 3, the symbol error rates (uncoded symbols $a_{k,u}$) of the three users (via the respective color) are depicted over the signal-to-noise ratio $E_\mathsf{s}/N_0 = 1/\sigma_\mathsf{n}^2$; the darker curves correspond to individual detection of the users (fixed receive windowing with $\zeta_{\mathsf{w},u} = 15, \forall u$). Conventional DD, which does not perform well, is shown for comparison. Using nDFE, for the "middle" user (🟩) significant gains can be achieved. The other users (🟦, 🟥) slightly profit from the optimization of $\zeta_{\mathsf{w},u}$.

For reference, the performance of coherent BLAST detection with perfect channel knowledge is shown; here all users perform nearly the same. When taking some degradation due to non-perfect actual channel knowledge into account, noncoherent reception with nDFE (over the users) and DFDD (over the temporal dimension) will perform almost comparable.

The variation of the optimum parameters in nDFE when user 2 (🟩) moves along the hallway is shown in the top plot of Fig. 4 ($E_\mathsf{s}/N_0 \,\widehat{=}\, 14$ dB). In the bottom plot, the optimized receive window width $\zeta_{\mathsf{w},u}$ is depicted. For $m_2 \leq 53$ user 3 (🟦) is detected first, then user 1 (🟥), and finally user 2 (🟩). For $m_2 > 53$ users 1 and 3 exchange their role. If the users are close to each other, they cannot be separated based only on the PSPs. Here, coherent detection, where the phase of the channel coefficients is utilized, too, is clearly superior. However, if a sufficient spacing of the users with moderate overlap of the PSPs is present, nDFE/DFDD is attractive and is able to lower the SER of the middle user significantly.

V. Conclusions

In conclusion, it can be stated that noncoherent detection in massive MIMO systems is an attractive alternative to coherent schemes. By combining two DFE procedures low-complexity high-performance noncoherent multi-user detection schemes are enabled. First, DFE over the users utilizing statistical channel knowledge is present. Optimized parameters (sorting, weighting) can be derived purely from the statistical knowledge (PSP). Second, DFE over the temporal direction (blockwise processing) employing DFDD for each user individually is active. Here, an optimized processing order is obtained from the actual receive signal [11]. A joint spatial/temporal DFE with optimum global ordering would be possible, too.

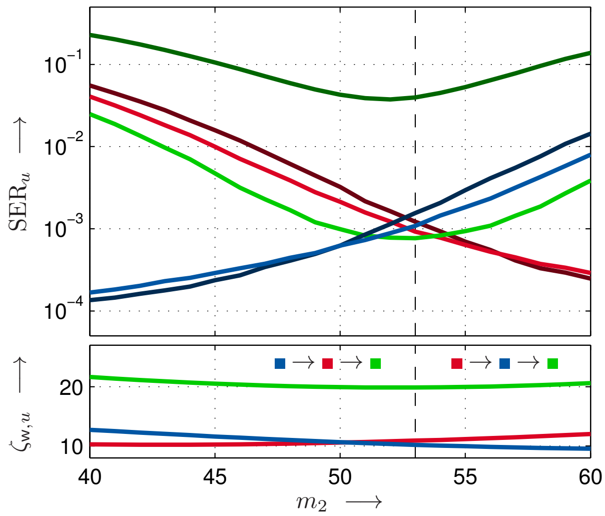


Fig. 4. Top: symbol error rate vs. the position of user 2 (🟩). Bottom: optimized receive window width $\zeta_{\mathsf{w},u}$. The decision order is indicated. $E_\mathsf{s}/N_0 \,\widehat{=}\, 14$ dB. $N_\mathsf{u} = 3$; $m_1 = 20$ (🟥), $m_3 = 85$ (🟦); colors corresponding to users. Uniform linear array with $N_\mathsf{rx} = 100$ antenna elements. Power-space profile according to (1), $\zeta_u = 20$, $\forall u$. Burst length $N = 200$. $M = 4$-ary DPSK.

References

[1] F. Adachi, M. Sawahashi. Decision Feedback Multiple-Symbol Differential Detection for Mary DPSK. *Electronics Letters*, pp. 1385–1387, July 1993.

[2] D. Divsalar, M.K. Simon. Multiple-Symbol Differential Detection of MPSK. *IEEE Tr. Comm.*, pp. 300–308, Mar. 1990.

[3] F. Edbauer. Bit Error Rate of binary and Quaternary DPSK Signals with Multiple Differential Feedback Detection. *IEEE Tr. Comm.*, pp. 457–460, Mar. 1992.

[4] G.J. Foschini, D. Chizhik, M.J. Gans, C.Papadias, R.A. Valenzuela. Analysis and Performance of Some Basic Space-Time Architectures. *IEEE J. Sel. Areas in Comm.*, pp. 303–320, Apr. 2003.

[5] N. Guo, R.C. Qiu. Improved Autocorrelation Demodulation Receivers based on Multiple-Symbol Detection for UWB Communications. *IEEE Tr. Wireless Comm.*, pp. 2026–2031, Aug. 2006.

[6] V. Lottici, Z. Tian. Multiple Symbol Differential Detection for UWB Communications. *IEEE Tr. Wireless Comm.*, pp. 1656–1666, May 2008.

[7] R.K. Mallik, N.C. Sagias. Distribution of Inner Product of Complex Gaussian Random Vectors and its Applications. *IEEE Tr. Comm.*, pp. 3353–3362, Dec. 2011.

[8] T.L. Marzetta. Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas. *IEEE Tr. Wireless Comm.*, pp. 3590–3600, Nov. 2010.

[9] F. Rusek, D. Persson, B.K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, F. Tufvesson. Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays. *IEEE Signal Proc. Magazine*, pp. 40–60, Jan. 2013.

[10] A. Schenk. *Coding, Modulation, and Detection in Impulse-Radio Ultra-Wideband Comm.*, PhD Thesis. University Erlangen-Nürnberg, 2012.

[11] A. Schenk, R.F.H. Fischer. Noncoherent Detection in Massive MIMO Systems. In *Proc. Int. ITG/IEEE Workshop on Smart Antennas*, Stuttgart, Germany, March 2013.

[12] R. Schober, W.H. Gerstacker, J.B. Huber. Decision-Feedback Differential Detection of MDPSK for Flat Rayleigh Fading Channels. *IEEE Tr. Comm.*, pp. 1025–1035, July 1999.

The Anti-Diversity Concept for Secure Communication on a Two-Link Compound Channel

Joseph J. Boutros
Texas A&M University
Electrical Engineering Dept.
23874 Doha, Qatar
Email: boutros@tamu.edu

Volkan Dedeoglu
Texas A&M University
Electrical Engineering Dept.
23874 Doha, Qatar
Email: volkan.dedeoglu@tamu.edu

Matthieu Bloch
Georgia Institute of Technology
School of Elec. and Comp. Eng.
Atlanta GA 30332, USA
Email: matthieu.bloch@ece.gatech.edu

***Abstract*—We propose new coding schemes for secrecy over a two-link compound channel. Firstly, a non-stochastic scheme is developed based on diversity-deficient LDPC ensembles and a source splitter. Secondly, a stochastic scheme is built from the same splitter with the adjunction of a random sequence. These coding structures achieve perfect secrecy in the algebraic and the information-theoretic sense respectively.**

I. Introduction and Notations

While the applications of Wyner's wiretap channel model [1] to physical-layer security have attracted much interest, see for instance [2][3] and references therein, few constructive and low complexity coding schemes have been developed [4]. Recent efforts exploiting powerful families of error-control codes have nevertheless met some success in certain cases. For instance, low-density parity check codes (LDPC) have been shown to provide secrecy over erasure channels [5][6][7], while Polar codes [8] and invertible extractors [9] have been proven to ensure secrecy over some symmetric channels. Several results also suggest the usefulness of LDPC codes and lattice codes over the Gaussian wiretap channel [10][11][12]. However, all the aforementioned constructions only apply to memoryless wiretap channels with full statistical knowledge of the eavesdropper's channel, which limits their scope of applications. In this paper, we provide a first step towards more robust designs by developing a coding scheme that provides secrecy over a compound wiretap channel [13] in which the eavesdropper gets to observe one of two channels.

Our compound channel has two identical links defined by their transition probabilities $p_{Y|X}(y_1|v)$ and $p_{Y|X}(y_2|w)$ respectively, as depicted in Figure 1. These two links can be any binary memoryless symmetric (BMS) channel, $v, w \in \mathbb{F}_2^{N/2}$ and $y_1, y_2 \in \mathcal{Y}^{N/2}$, where $\mathcal{Y}$ is the output alphabet as observed by the legitimate receiver. It is assumed that a uniform binary source produces K binary elements. The length-N codeword generated by a rate-K/N binary encoder is divided into parts v and w to be transmitted in parallel. For the sake of simplifying the notations, we decided to use a unique letter to denote a random variable and any given value taken by that random variable. We apologize for not keeping the notation rigour as Grimmett and Stirzaker. The reader should figure out easily from the context whether we are referring to a random variable or to a given value.

Regarding the eavesdropper, our study considers the worst case scenario. Let the channel between Alice and Eve have output $z \in \mathbb{F}_2^{N/2}$. Eve is reading a noiseless copy of one of the two links, i.e. $z = v$ or $z = w$. While assuring that Bob has excellent performance, our aim is to prevent Eve from determining the source bits, part of the source bits, or any information derived from the source bits. Let $M = (a_1, a_2, \ldots, a_K) \in \mathbb{F}_2^K$ be the source message. In the upcoming sections, two types of security are studied:

- **Algebraic security**. Given z, Eve must not be able to find the value of an individual binary element a_i, $\forall i = 1 \ldots K$. This algebraic security is achieved by the means of a non-stochastic LDPC encoder and a weight-2 splitter as described in Section II.
- **Information theoretic security**. The system design must guarantee a zero leakage, i.e. zero mutual information $I(M; z) = 0$ or equivalently $H(M|z) = H(M)$. This perfect secrecy is achieved via a stochastic encoding including a random sequence as described in Section IV.

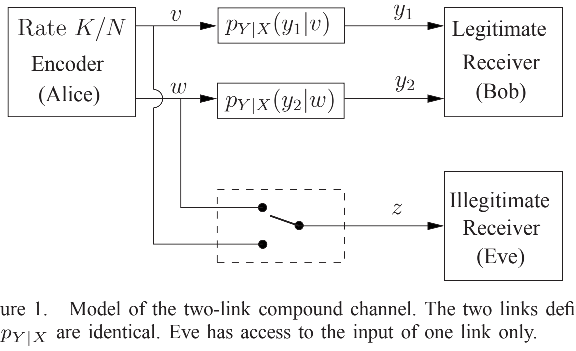


Figure 1. Model of the two-link compound channel. The two links defined by $p_{Y|X}$ are identical. Eve has access to the input of one link only.

II. Anti-Diversity LDPC Encoding

The reader is assumed to be familiar with LDPC codes [14] and diversity methods for fading channels [15][16]. The

The work of Joseph J. Boutros and Volkan Dedeoglu was supported by QNRF, a member of Qatar Foundation, NPRP project 5-603-2-243 at Texas A&M University. The work of Matthieu Bloch was supported by the National Science Foundation, award CC 1320298.

compound channel with two parallel links is very similar to block fading channels considered in [17][18]. On a double diversity fading channel, channel coding is supposed to exhibit an error rate performance P_e proportional to $1/\gamma^2$ at high signal-to-noise ratio γ. In such a case, the channel code is said to be a full-diversity code. An example of full-diversity code ensemble is the Root-LDPC ensemble [17][18].

In coding for double diversity on block fading channels, three fundamental rules should be satisfied [17]:

- The coding rate R must satisfy $R \leq \frac{1}{2}$.
- Let the parity-check matrix be divided into two equal size sub-matrices, $H = [H_1|H_2]$. Under Maximum-Likelihood decoding, H_1 and H_2 must have full rank.
- Under iterative message-passing decoding, information bits must be connected to root checknodes of order one or higher.

Security cannot be achieved with a full-diversity code on the two-link compound channel. Double diversity would let Eve determine the missing link and hence all source bits will be revealed. The LDPC code design for security should not satisfy the rules listed above.

Definition 1: The anti-diversity concept refers to a code design where the three fundamental diversity rules are intentionally violated.

The LDPC code constructed via an anti-diversity concept will be called an anti-root LDPC. We briefly describe the structure of an anti-root LDPC. Let $\frac{K}{N}$ be the design rate (R is the effective rate), then $\frac{1}{2} \leq \frac{K}{N} \leq R < 1$. The N binary digits of a codeword are divided into four families. A family of $K/2$ information digits $1i$ and a family of $(N-K)/2$ parity digits $1p$ to be sent on the first link. Similarly, the two families $2i$ and $2p$ are to be sent on the second link. The design rate $\frac{K}{N}$ is taken in the range $[\frac{1}{2}, 1)$. In the special case $\frac{K}{N} = \frac{1}{2}$, a deficient diversity is assured by the last two rules.
Let H_1, the left half part of H, be written as a block matrix

$$H_1 = \begin{bmatrix} A_1 & B_1 \\ C_1 & S_1 \end{bmatrix}. \tag{1}$$

The submatrix A_1 of size $(N-K)/2 \times K/2$ corresponds to edges connecting bitnodes $1i$ to a first type of checknodes $1c$. The submatrix B_1 of square size $(N-K)/2 \times (N-K)/2$ corresponds to edges connecting bitnodes $1p$ to the first type of checknodes $1c$. In a similar fashion, C_1 and S_1 have the same size as A_1 and B_1 respectively. C_1 and S_1 define edges from $2i$ and $2p$ to the second type of checknodes $2c$. Now, the third rule is violated by taking $B_1 = I$, where I is the identity matrix of size $(N-K)/2$. In the special case $\frac{K}{N} = \frac{1}{2}$, A_1 and I commute, then

$$\det(H_1) = det(C_1 + S_1 A_1). \tag{2}$$

Forcing the equality $C_1 = S_1 A_1$ makes H_1 rank deficient, i.e. now the second fundamental rule is violated. The general structure of the anti-root LDPC ensemble is shown in Figure 2. The algebraic security is proved for a scrambler S_1 of any column and row weight greater than or equal to 1.

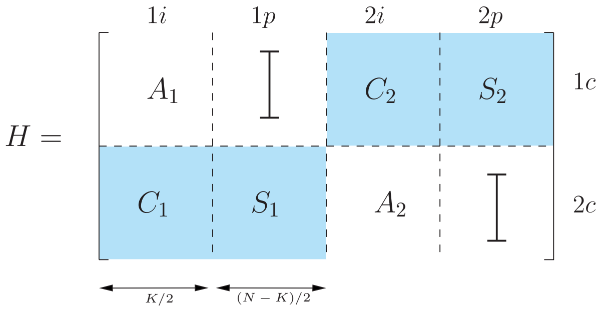


Figure 2. Parity-check matrix of an anti-root LDPC code for violating double diversity. Design rate is $\frac{K}{N}$, where $\frac{1}{2} \leq \frac{K}{N} < 1$.

Our system model is symmetric with respect to the LDPC code and to the two-link channel. Thus, the right half part H_2 has a structure identical to H_1 after switching checknodes $1c$ and $2c$. The expressions of C_1 and C_2 are maintained for any design rate $\frac{K}{N}$.

Definition 2: The anti-root LDPC ensemble is defined by its low-density parity-check matrix in Figure 2 where

$$C_1 = S_1 A_1 \quad \text{and} \quad C_2 = S_2 A_2. \tag{3}$$

The anti-root LDPC is systematic. Eve should not have direct access to source digits [10]. Hence, a source splitter S is placed between the source and the LDPC encoder. The matrix S is $K \times K$, non-singular, and sparse. Suppose that S is regular with degree d_s, i.e. the Hamming weight of all rows and columns is d_s. Let $u = (u_1, u_2, \ldots, u_K) \in \mathbb{F}_2^K$ be the LDPC encoder input. Then $u = MS^{-1}$, or equivalently $M = uS$. The latter is an operation that splits each source digit into d_s digits [19][20]. In this paper, we restrict the splitter to have a degree $d_s = 2$, except for one row and one column in S that have a degree equal to 1.

Lemma 1: Consider a quasi-regular weight-2 non-singular splitter S, except for one row and one column whose degree is 1. Then, S is equivalent to a double diagonal splitter S_0,

$$S = \Pi \cdot S_0 \cdot \Pi', \tag{4}$$

where Π and Π' are $K \times K$ permutation matrices.
In the sequel, we assume that $S = S_0$, i.e. we have

$$a_i = u_i + u_{i+1}, \tag{5}$$

for $i = 1 \ldots K-1$ and $a_K = u_K$. We force to zero the last source bit, $a_K = u_K = 0$. The exact message entropy is $H(M) = K - 1$ bits instead of K. The structure of the non-stochastic coding scheme is shown in Figure 3. The splitter K-bit output $u = (1i \,\&\, 2i)$ is dispatched at the LDPC encoder input such that the $K/2$ bits at odd positions go to $1i$ and the $K/2$ bits at even positions go to $2i$. Thus, Eve must know both family of bits $1i$ and $2i$ in order to find the source message M. When $z = v$, Eve knows all bits $1i$ but the bits $2i$ are all missing. The anti-root LDPC does not allow Eve to find any of the missing bits $2i$. Similarly, when $z = w$, the anti-root LDPC does not allow Eve to find $1i$. The proof of the following theorem is based on (3) and (5).

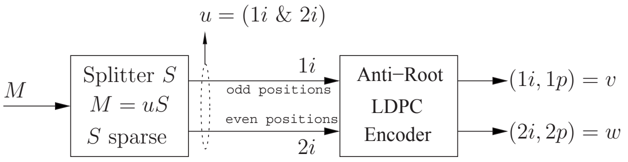


Figure 3. The non-stochastic encoder converts the source message M into half codewords v and w to be sent on each link. $M \in \mathbb{F}_2^K$, $v, w \in \mathbb{F}_2^{N/2}$.

Theorem 2: The anti-root LDPC ensemble of design rate $\frac{K}{N} \in [\frac{1}{2}, 1)$ guarantees the algebraic security of the communication system between Alice and Bob.

Given the non-stochastic scheme which is algebraically secure for any block length N, the next section studies the asymptotic performance of Bob for N sufficiently large.

III. Legitimate Receiver's Performance

Many anti-root LDPC ensembles can be defined, a given ensemble depends on how the submatrices A_1, S_1, A_2, and S_2 are constructed. Due to the lack of space, we restrict this section to a design rate $\frac{K}{N} = \frac{1}{2}$ and to $A_1 = \Pi_1$ and $A_2 = \Pi_2$, where Π_1 and Π_2 are uniformly chosen in the set of $\frac{K}{2} \times \frac{K}{2}$ binary permutation matrices.

The asymptotic performance of Bob under iterative message passing is found via density evolution (DE) [14]. The anti-root LDPC defined by its parity-check matrix H in Figure 2 and by (3) is a multi-edge type code on graphs. As in [17][18], an extra difficulty arises because only the performance on information bits is relevant. Hence, we define the following polynomials to be used by DE at bitnodes and checknodes. The global degree distribution of H from an edge perspective, at bitnodes and checknodes respectively, is [14]:

$$\lambda(x) = \sum_{i=2}^{d_b} \lambda_i x^{i-1}, \quad \text{and} \quad \rho(x) = \sum_{j=2}^{d_c} \rho_j x^{j-1}. \qquad (6)$$

In this section, it is assumed that $\rho_j = 0$ for j odd. We introduce an edge-perspective polynomial $\tilde{\lambda}(x)$ when one edge is missing [18] and a node-perspective polynomial $\mathring{\lambda}(x)$,

$$\tilde{\lambda}(x) = \sum_{i=1}^{d_b-1} \tilde{\lambda}_i x^{i-1} = \frac{\bar{d}_b}{\bar{d}_b - 1} \sum_{i=1}^{d_b-1} i\ \lambda_{i+1}/(i+1)\ x^{i-1}, \quad (7)$$

$$\mathring{\lambda}(x) = \sum_{i=2}^{d_b} \mathring{\lambda}_i x^{i-1} = \bar{d}_b \sum_{i=2}^{d_b} \lambda_i / i\ x^{i-1}, \qquad (8)$$

where $\bar{d}_b$ is the average degree of bitnodes. The polynomials $\tilde{\rho}(x)$ and $\mathring{\rho}(x)$ are defined in a similar manner for checknodes. Finally, two bivariate polynomials are necessary due to the separation of a checknode into two parts for information and parity bits on the same side of H,

$$\mathring{\rho}(x, y) = \sum_{j=2}^{d_c} \mathring{\rho}_j x^{(j-2)/2} y^{(j-2)/2}, \qquad (9)$$

$$\hat{\rho}(x, y) = \sum_{j=1}^{(d_c-2)/2} \hat{\rho}_j x^{j-1} y^j = \sum_{j=1}^{(d_c-2)/2} \frac{2j\mathring{\rho}_j}{\bar{d}_c - 2} x^{j-1} y^j. \quad (10)$$

For a general anti-root ensemble with two distinct links in the compound channel, density evolution may involve up to eight message densities. In this section, due to the identical links and to the LDPC code symmetry, DE equations require two densities only: a- f is the probability density function of log-ratio messages from bitnode $1i$ to checknode $2c$, from $1p$ to $2c$, from $2i$ to $1c$, and from $2p$ to $1c$. b- q is the probability density function of log-ratio messages from bitnode $1i$ to checknode $1c$, from $1p$ to $1c$, from $2i$ to $2c$, and from $2p$ to $2c$. After drawing the local neighborhood of each type of bitnodes (tree representations omitted due to lack of space), we find the following DE equations at decoding iteration $m + 1$:

$$q^{m+1} = \mu \otimes \mathring{\lambda}\left(\hat{\rho}(f^m, f^m) \odot (q^m)^{\odot 2}\right),$$

$$f^{m+1} = \mu \otimes (q^m \odot \mathring{\rho}(f^m, f^m)) \otimes \tilde{\lambda}\left(\hat{\rho}(f^m, f^m) \odot (q^m)^{\odot 2}\right),$$

where μ is the density at the channel output, $\otimes$ and $\odot$ denote convolution at bitnode and checknode levels.

Theorem 3: Consider a rate-$1/2$ anti-root LDPC ensemble. If the ensemble is regular then DE reduces to one equation $f^{m+1} = \mu \otimes \lambda(\rho(f^m))$, i.e. the anti-root LDPC has the same decoding threshold as a regular fully-random LDPC ensemble.

In the regular case, the constraint in (3) did not weaken the LDPC code. For irregular ensembles, thresholds can be optimized by a judicious choice of $\lambda(x)$ and $\rho(x)$.

IV. Stochastic Encoding for Two Links

A non-stochastic algebraically-secure encoding scheme has been described in the previous sections and its performance analyzed via density evolution. Now, we would like to replace algebraic security by perfect secrecy in the information-theoretic sense. A perfectly secure stochastic encoding structure is proposed in this section.

In the non-stochastic case, we had $H(M) = K$ (we omit $a_K = 0$ in order to simplify the notations). The conditional message entropy was given by

$$H(M|z = v) = H(1i|z = v) + H(2i|z = v, 1i) = H(2i|v).$$

The information leakage between v and $2i$ is unknown and may depend on the particular choice of submatrices inside H. Nevertheless, we always have $0 < H(2i|v) \leq K/2$. Similar arguments can be made for $z = w$ and $H(1i|w)$. In summary, the non-stochastic coding scheme satisfies

$$H(M|z) \leq \frac{K}{2} < K = H(M). \qquad (11)$$

Our stochastic scheme will sacrifice $K/2$ bits in the message by reducing the entropy of the message to $H(M) = K/2$ to achieve perfect secrecy in the information theoretic sense after satisfying $H(M|z) = H(M) = K/2$.

The splitter input is modified to include both $M = (a_1, a_2, \ldots, a_{K/2})$ and a zero sequence of length $K/2$. Let $r = (r_1, r_2, \ldots, r_{K/2})$ be a random sequence of $K/2$ independent uniform binary digits. r is added to both splitter

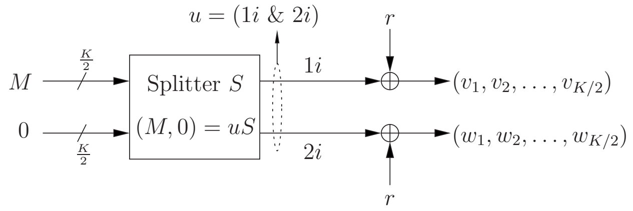


Figure 4. The $K \times K$ splitter in the stochastic scheme reads a message M of $K/2$ bits and a zero sequence of $K/2$ bits. A random sequence of $K/2$ bits is applied at the splitter output before channel transmission.

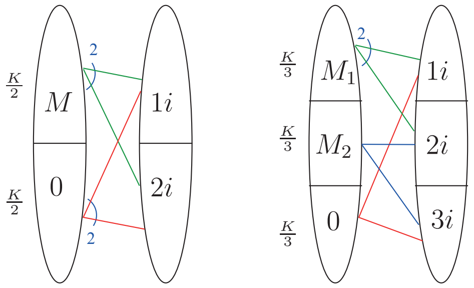


Figure 5. Splitter structure for stochasting encoding for two and three links respectively. The sparse graph represents the expression $(M, 0) = uS$ where S is sparse with degree 2.

outputs. The stochastic structure is shown in Figure 4 where the splitter output fills $K/2$ bits in v and $K/2$ bits in w. The remaining $N - K$ bits in v and w will be equal to parity bits of an LDPC encoder. The analysis below is valid for a two-link anti-root LDPC and for two separate length-$N/2$ LDPC codes. The splitter structure is also illustrated in Figure 5. In a straightforward manner, Theorem 4 can be generalized to an eavesdropper reading one link out of L links, for any $L \geq 2$.

Theorem 4: The stochastic encoding scheme yields $H(M|z) = \frac{K}{2} = H(M)$ on a two-link compound channel, i.e. it is perfectly secure in the information-theoretic sense.

Proof. A sketch of the proof is given. Notice that the zero sequence at the splitter input makes $2i$ a permuted version of $1i$. So $H(2i|z, 1i) = 0$. The equivocation is $H(M|z) = H(1i, 2i|z) = H(1i|z) + H(2i|z, 1i) = H(1i|z)$. Consider $z = v$. For the case of two separate length-$N/2$ codes, we have $H(1i|v) = H(1i|1i + r) = H(1i) = K/2$. For an anti-root LDPC, $H(1i|v) = H(1i|1i + r, 1p)$, the latter is equal to $H(1i|1i + r) = K/2$ because $1p$ is a function of $1i + r$ only thanks to the splitter. Similar proof is made for $z = w$. □.

V. Conclusion

We proposed two original coding schemes for secure communication over a two-link compound channel. A non-stochastic scheme has been developed based on diversity-deficient LDPC ensembles and a source splitter. The anti-root LDPC code guarantees perfect algebraic security. Its joint structure makes it twice longer than two separate LDPC codes for each link and forbids Eve from correcting channel errors when $z = y_1$ or $z = y_2$. The second scheme is stochastic and attains perfect information-theoretic secrecy. It is built from a splitter with the adjunction of a random sequence.

Our work is related to methods in secret sharing such as the material found in [21][22][23], but our channel model does not include feedback and our aim is to increase the information rate rather than finding the worst channel conditions.

References

[1] A. D. Wyner, “The wire-tap channel,” Bell Syst. Tech. J., vol. 54, no. 8, pp. 13551387, Oct. 1975.

[2] Y. Liang, H. V. Poor, and S. Shamai, *Information-Theoretic Security*, Now Publishers, 5 (1–5), pp. 355-580, 2009.

[3] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[4] W.K. Harrison, J. Almeida, M.R. Bloch, S.W. McLaughlin, and J. Barros, “Coding for Secrecy: An Overview of Error-Control Coding Techniques for Physical-Layer Security,” *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 41-50, 2013.

[5] A. Thangaraj, S. Dihidar, A. R. Calderbank, S. W. McLaughlin, and J.-M. Merolla, “Applications of LDPC codes to the wiretap channel,” *IEEE Trans. Inform. Theory*, vol. 53, no. 8, pp. 2933-2945, Aug. 2007.

[6] A. Subramanian, A. Thangaraj, M. Bloch, and S. W. McLaughlin, “Strong secrecy on the binary erasure wiretap channel using large-girth LDPC codes,” *IEEE Trans. Inform. Forensics Sec.*, vol. 6, no. 3, pp. 585-594, Sept. 2011.

[7] V. Rathi, R. Urbanke, M. Andersson, and M. Skoglund, “Rate-equivocation optimal spatially coupled LDPC codes for the BEC wiretap channel,” *in Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, St. Petersburg, Russia, Aug. 2011, pp. 2393-2397.

[8] H. Mahdavifar and A. Vardy, “Achieving the secrecy capacity of wire-tap channels using polar codes,” *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6428-6443, 2011.

[9] M. Bellare, S. Tessaro, and A. Vardy, “Semantic security for the wiretap channel,” *in Advances in Cryptology, CRYPTO 2012* (Lecture Notes in Computer Science, vol. 7417). Berlin, Heidelberg, Germany: Springer-Verlag, pp. 294-311.

[10] M. Baldi, M. Bianchi, and Franco Chiaraluce, “Non-Systematic Codes for Physical Layer Security,” *in Proc. IEEE Information Theory Workshop (ITW 2010)*, pp. 1-5, Dublin, Ireland, Aug. 30 - Sept. 3, 2010.

[11] D. Klinc, J. Ha, S. W. McLaughlin, J. Barros, and B.-J. Kwak, “LDPC codes for the Gaussian wiretap channel,” *IEEE Trans. Inform. Forensics Sec.*, vol. 6, no. 3, pp. 532-540, Sept. 2011.

[12] C. Ling, L. Luzzi, J.-C. Belfiore, and D. Stehl, “Semantically secure lattice codes for the Gaussian wiretap channel,” *Computing Research Repository*, Oct. 2012, pp. 1-19.

[13] Y. Liang, G. Kramer, H. V. Poor and S. Shamai, “Compound Wiretap Channels,” *EURASIP Journal on Wireless Communications and Networking*, 2009, 142374, 1-12.

[14] T.J. Richardson and R.L. Urbanke, *Modern Coding Theory*, Cambridge University Press, 2008.

[15] D.N.C. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[16] E. Biglieri, *Coding for Wireless Channels*, New York, Springer, 2005.

[17] J.J. Boutros, A. Guillén i Fàbregas, E. Biglieri, and G. Zémor, “Low-Density Parity-Check Codes for Nonergodic Block-Fading Channels,” *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4286-4300, Sept. 2010.

[18] J.J. Boutros, “Diversity and coding gain evolution in graph codes,” *in Proc. Information Theory and Appl. (ITA’2009)*, pp. 34-43, UCSD, San Diego, Feb. 2009.

[19] G. Shamir and J. Boutros, “Non-systematic low-density parity-check codes for nonuniform sources,” *in Proc. International Symp. on Information Theory (ISIT 2005)*, Adelaide, Australia, pp. 18981902, Sept. 2005.

[20] A. Alloum, J. Boutros, G. Shamir, and L. Wang, “Non-systematic LDPC codes via scrambling and splitting,” *in Proc. Allertons Conference on Comm. and Control*, Monticello, Illinois, pp. 1879-1888, Sept. 2005.

[21] D. Dolev, C. Dwork, O. Waarts, and M. Yung, “Perfectly secure message transmission,” *Journal of the ACM*, vol. 40, no. 1, pp. 1747, Jan. 1993.

[22] T. Rabin, “Robust sharing of secrets when the dealer is honest or cheating,” *Journal of the ACM*, vol. 41, no. 6, pp. 1089-1109, Nov. 1994.

[23] Q. Yang and Y. Desmedt, “General Perfectly Secure Message Transmission Using Linear Codes,” *Advances in Cryptology - ASIACRYPT 2010*, vol. 6477, pp. 448-465, Dec. 2010.

Almost Linear Complexity Methods for Delay-Doppler Channel Estimation

Alexander Fish and Shamgar Gurevich

***Abstract*—A fundamental task in wireless communication is *channel estimation*: Compute the channel parameters of a medium between a transmitter and a receiver. In the case of delay-Doppler channel, i.e., a signal undergoes only delay and Doppler shifts, a widely used method to compute delay-Doppler parameters is the *pseudo-random* method. It uses a pseudo-random sequence of length N, and, in case of non-trivial relative velocity between transmitter and receiver, its computational complexity is $O(N^2 \log N)$ arithmetic operations. In [1] the flag method was introduced to provide a faster algorithm for delay-Doppler channel estimation. It uses specially designed flag sequences and its complexity is $O(rN \log N)$ for channels of *sparsity* r. In these notes, we introduce the *incidence* and *cross* methods for channel estimation. They use triple-chirp and double-chirp sequences of length N, correspondingly. These sequences are closely related to chirp sequences widely used in radar systems. The arithmetic complexity of the incidence and cross methods is $O(N \log N + r^3)$, and $O(N \log N + r^2)$, respectively.**

I. Introduction

A basic building block in many wireless communication protocols is *channel estimation*: learning the channel parameters of the medium between a transmitter and a receiver [6]. In these notes we develop efficient algorithms for delay-Doppler (also called time-frequency) channel estimation. Throughout these notes we denote by $\mathbb{Z}_N$ the set of integers $\{0, 1, ..., N-1\}$ equipped with addition and multiplication modulo N. We will assume, for simplicity, that N is an odd prime. We denote by $\mathcal{H} = \mathbb{C}(\mathbb{Z}_N)$ the vector space of complex valued functions on $\mathbb{Z}_N$, and refer to it as the *Hilbert space of sequences*.

A. Channel Model

We describe the discrete channel model which was derived in [1]. We assume that a transmitter uses a sequence $S \in \mathcal{H}$ to generate an analog waveform $S_A \in L^2(\mathbb{R})$ with bandwidth W and a carrier frequency $f_c \gg W$. Transmitting S_A, the receiver obtains the analog waveform $R_A \in L^2(\mathbb{R})$. We make the sparsity assumption on the number of paths for propagation of the waveform S_A. As a result, we have[1]

$$R_A(t) = \sum_{k=1}^{r} \beta_k \cdot \exp(2\pi i f_k t) \cdot S_A(t - t_k) + \mathcal{W}(t), \qquad \text{(I-A.1)}$$

where r—called the *sparsity* of the channel—denotes the number of paths, $\beta_k \in \mathbb{C}$ is the *attenuation coefficient*, $f_k \in \mathbb{R}$ is the *Doppler shift* along the k-th path, $t_k \in \mathbb{R}_+$ is the *delay* associated with the k-th path, and $\mathcal{W}$ denotes a random white noise. We assume the normalization $\sum_{k=1}^{r} |\beta_k|^2 \leq 1$. The Doppler shift depends on the relative velocity, and the delay encodes the distance along a path, between the transmitter and the receiver. We will call

$$(\beta_k, t_k, f_k), \ k = 1, ..., r, \qquad \text{(I-A.2)}$$

channel parameters, and the main objective of channel detection is to estimate them.

B. Channel Estimation Problem

Sampling the waveform R_A at the receiver side, with sampling rate $1/W$, we obtain a sequence $R \in \mathcal{H}$. It satisfies

$$R[n] = H(S)[n] + \mathcal{W}[n], \qquad \text{(I-B.1)}$$

where H, called the *channel operator*, acts on $S \in \mathcal{H}$ by[2]

$$H(S)[n] = \sum_{k=1}^{r} \alpha_k e(\omega_k n) S[n - \tau_k], \ n \in \mathbb{Z}_N, \qquad \text{(I-B.2)}$$

with α_k's are the complex-valued (digital) attenuation coefficients, $\sum_k |\alpha_k|^2 \leq 1$, $\tau_k \in \mathbb{Z}_N$ is the (digital) delay associated with the path k, $\omega_k \in \mathbb{Z}_N$ is the (digital) Doppler shift associated with path k, and $\mathcal{W}$ denotes the random white noise. We will assume that all the coordinates of $\mathcal{W}$ are independent identically distributed random variables of expectation zero.

Remark I-B.1: The relation between the physical (I-A.2) and the discrete channel parameters is as follows (see Section I.A. in [1] and references therein): If a standard method suggested by sampling theorem is used for the discretization, and S_A has bandwidth W, then $\tau_k = t_k W$ modulo N, and $\omega_k = N f_k / W$ modulo N, provided that $t_k \in \frac{1}{W}\mathbb{Z}$, and $f_k \in \frac{W}{N}\mathbb{Z}$, $k = 1, ..., r$. In particular, we note that the integer N determines the frequency resolution of the channel detection, i.e., the resolution is of order W/N.

The objective of delay-Doppler channel estimation is:

*Problem I-B.2 (**Channel Estimation**):* Design $S \in \mathcal{H}$, and an effective method for extracting the channel parameters $(\alpha_k, \tau_k, \omega_k)$, $k = 1, ..., r$, using S and R satisfying (I-B.1).

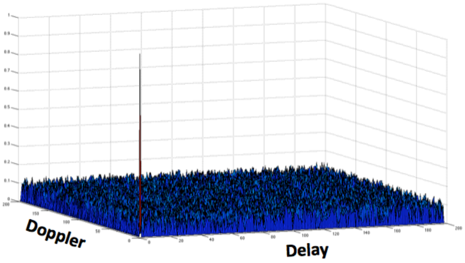


Fig. 1. Profile of $\mathcal{A}(\varphi, \varphi)$ for φ pseudo-random sequence.

C. Ambiguity Function and Pseudo-Random Method

A classical method to estimate the channel parameters in (I-B.1) is the *pseudo-random method* [2], [3], [4], [6], [7]. It uses two ingredients - the ambiguity function, and a pseudo-random sequence.

This material is based upon works supported by DARPA award number N66001-13-1-4052, and NSF Grant DMS-1101660.

A. Fish is with School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia. Email: alexander.fish@sydney.edu.au.

S. Gurevich is with the Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA. Email: shamgar@math.wisc.edu.

[1] In these notes i denotes $\sqrt{-1}$.

[2] We denote $e(t) = \exp(2\pi i t/N)$.

*1) **Ambiguity Function:*** In order to reduce the noise component in (I-B.1), it is common to use the ambiguity function that we are going to describe now. We consider the *Heisenberg operators* $\pi(\tau,\omega)$, $\tau,\omega\in\mathbb{Z}_N$, which act on $f\in\mathcal{H}$ by

$$[\pi(\tau,\omega)f]\,[n]=e(-2^{-1}\tau\omega)\cdot e(\omega n)\cdot f[n-\tau], \tag{I-C.1}$$

where 2^{-1} denotes $(N+1)/2$, the inverse of 2 mod N. Finally, the *ambiguity function* of two sequences $f,g\in\mathcal{H}$ is defined[3] as the $N\times N$ matrix

$$\mathcal{A}(f,g)[\tau,\omega]=\langle\pi(\tau,\omega)f,g\rangle\,,\quad \tau,\omega\in\mathbb{Z}_N, \tag{I-C.2}$$

where $\langle\ ,\ \rangle$ denotes the standard inner product on $\mathcal{H}$.

*Remark I-C.1 (**Fast Computation of Ambiguity Function**):* The restriction of the ambiguity function to a line in the delay-Doppler plane, can be computed in $O(N\log N)$ arithmetic operations using fast Fourier transform [5]. For more details, including explicit formulas, see Section V of [1]. Overall, we can compute the entire ambiguity function in $O(N^2\log N)$ operations.

For R and S satisfying (I-B.1), the law of the iterated logarithm implies that, with probability going to one, as N goes to infinity, we have

$$\mathcal{A}(S,R)[\tau,\omega]=\mathcal{A}(S,H(S))[\tau,\omega]+\varepsilon_N, \tag{I-C.3}$$

where $|\varepsilon_N|\le\sqrt{2\log\log N}/\sqrt{N\cdot SNR}$, with SNR denotes the *signal-to-noise ratio*[4].

*2) **Pseudo-Random Sequences:*** We will say that a norm-one sequence $\varphi\in\mathcal{H}$ is *B-pseudo-random,* $B\in\mathbb{R}$—see Figure 1 for illustration—if for every $(\tau,\omega)\neq(0,0)$ we have

$$|\mathcal{A}(\varphi,\varphi)[\tau,\omega]|\le B/\sqrt{N}. \tag{I-C.4}$$

There are several constructions of families of pseudo-random (PR) sequences in the literature (see [2], [3] and references therein).

*3) **Pseudo-Random Method:*** Consider a pseudo-random sequence φ, and assume for simplicity that $B=1$ in (I-C.4). Then we have

$$\mathcal{A}(\varphi,H(\varphi))[\tau,\omega] \tag{I-C.5}$$
$$=\begin{cases}\widetilde{\alpha}_k+\sum\limits_{j\neq k}\widetilde{\alpha}_j/\sqrt{N}, & \text{if } (\tau,\omega)=(\tau_k,\omega_k)\,,\ 1\le k\le r;\\ \sum\limits_{j}\widehat{\alpha}_j/\sqrt{N}, & \text{otherwise,}\end{cases}$$

where $\widetilde{\alpha}_j$, $\widehat{\alpha}_j$, $1\le j\le r$, are certain multiples of the α_j's by complex numbers of absolute value less or equal to one. In particular, we can compute the delay-Doppler parameter (τ_k,ω_k) if the associated attenuation coefficient α_k is sufficiently large. It appears as a peak of $\mathcal{A}(\varphi,H(\varphi))$. Finding the peaks of $\mathcal{A}(\varphi,H(\varphi))$ constitutes the pseudo-random method. Notice that the arithmetic complexity of the pseudo-random method is $O(N^2\log N)$, using Remark I-C.1. For applications to sensing, that require sufficiently high frequency resolution, we will need to use sequences of large length N. Hence, the following is a natural problem.

*Problem I-C.2 (**Arithmetic Complexity**):* Solve Problem I-B.2, with method for extracting the channel parameters which requires almost linear arithmetic complexity.

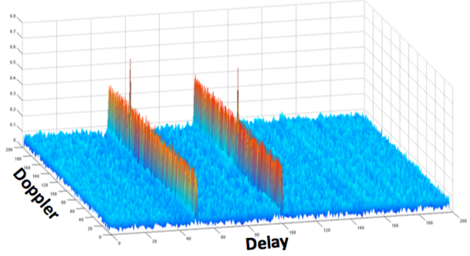


Fig. 2. Profile of $\mathcal{A}(f_L,H(f_L))$ for flag f_L, $L=\{(0,\omega)\}$, $N=199$, and channel parameters $(0.7,50,150)$, $(0.7,100,100)$.

D. Flag Method

In [1] the flag method was introduced in order to deal with the complexity problem. It computes the r channel parameters in $O(rN\log N)$ arithmetic operations. For a given line L in the plane $\mathbb{Z}_N\times\mathbb{Z}_N$, one construct a sequence f_L—called flag—with ambiguity function $\mathcal{A}(f_L,H(f_L))$ having special profile—see Figure 2 for illustration. It is essentially supported on shifted lines parallel to L, that pass through the delay-Doppler shifts of H, and have peaks there. This suggests a simple algorithm to extract the channel parameters. First compute $\mathcal{A}(f_L,H(f_L))$ on a line M transversal to L, and find the shifted lines on which $\mathcal{A}(f_L,H(f_L))$ is supported. Then compute $\mathcal{A}(f_L,H(f_L))$ on each of the shifted lines and find the peaks. The overall complexity of the flag algorithm is therefore $O(rN\log N)$, using Remark I-C.1. If r is large, it might be computationally insufficient.

E. Incidence and Cross Methods

In these notes we suggest two new schemes for channel estimation that have much better arithmetic complexity than previously known methods. The schemes are based on the use of double and triple chirp sequences.

*1) **Incidence Method:*** We propose to use triple-chirp sequences for channel estimation. We associate with three distinct lines L,M, and M° in $\mathbb{Z}_N\times\mathbb{Z}_N$, passing through the origin, a sequence $C_{L,M,M^\circ}\in\mathcal{H}$. This sequence has ambiguity function essentially supported on the union of L, M, and M°. As a consequence—see Figure 3 for illustration—the ambiguity function $\mathcal{A}(C_{L,M,M^\circ},H(C_{L,M,M^\circ}))$ is essentially supported on the shifted lines $\{(\tau_k,\omega_k)+(L\cup M\cup M^\circ)\,|\,k=1,\ldots,r\}$. This observation, which constitutes the bulk of the incidence method, enables a computation in $O(N\log N+r^3)$ arithmetic operations of all the time-frequency shifts (see Section III). In addition, the estimation of the corresponding r attenuation coefficients takes $O(r)$ operations. Hence, the overall complexity of incidence method is $O(N\log N+r^3)$ operations.

*2) **Cross Method:*** We propose to use double-chirp sequences for channel estimation. For two distinct lines L and M in $\mathbb{Z}_N\times\mathbb{Z}_N$, passing through the origin, we introduce a sequence $C_{L,M}\in\mathcal{H}$ with ambiguity function supported on L, and M. Under genericity assumptions—see Figure 4 for illustration—the essential support of $\mathcal{A}(C_{L,M},H(C_{L,M}))$ lies on $r\times r$ grid generated by shifts of the lines L, and M. Denote by $v_{ij}=l_i+m_j,\ l_i\in L, m_j\in M;\ 1\le i,j\le r$, the intersection points of the lines in the grid. Using Remark I-C.1 we find all the points $v_{ij}, 1\le i,j\le r$, in $O(N\log N)$ operations. The following matching problem arises: Find the r points from v_{ij}, $1\le i,j\le r$, which belong to the support of H. To suggest a solution,

[3] For our purposes it will be convenient to use this definition of the ambiguity function. The standard definition appearing in the literature is $A(f,g)[\tau,\omega]=\langle e(\omega n)f[n-\tau],g[n]\rangle$.

[4] We define $SNR=\langle S,S\rangle\,/\,\langle\mathcal{W},\mathcal{W}\rangle$.

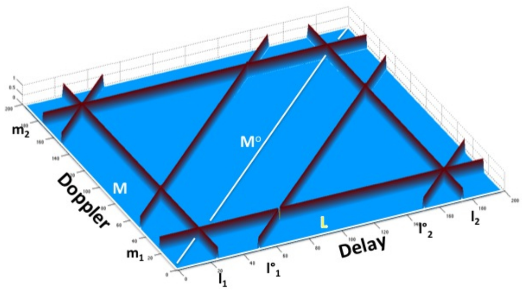


Fig. 3. Essential support of the ambiguity function $\mathcal{A}(C_{L,M,M^\circ}, H(C_{L,M,M^\circ}))$, where L is the delay line, M is the Doppler line, and M° is a diagonal line, and the support of H consists two parameters. Points of $\mathbb{Z}_N \times \mathbb{Z}_N$ through them pass three lines are the true delay-Doppler parameters of H.

we use the values of the ambiguity function to define a certain simple hypothesis function $h : L \times M \to \mathbb{C}$ (see Section IV). We obtain:

*Theorem I-E.1 (**Matching**):* Suppose $v_{ij} = l_i + m_j$ is a delay-Doppler shift of H, then $h(l_i, m_j) = 0$.

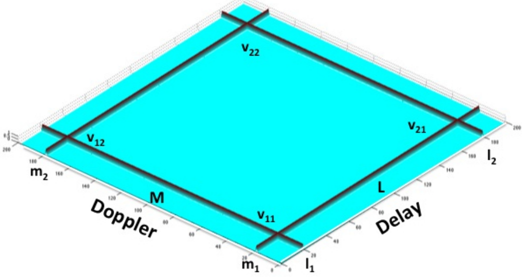


Fig. 4. Essential support of the ambiguity function $\mathcal{A}(C_{L,M}, H(C_{L,M}))$, where L is the delay line, M is the Doppler line, and the support of H consists two parameters.

The cross method makes use of Theorem I-E.1 and checks the values $h(l_i, m_j)$, $1 \leq i, j \leq r$. If a value is less than a priori chosen threshold, then the algorithm returns $v_{ij} = l_i + m_j$ as one of the delay-Doppler parameters. To estimate the attenuation coefficient corresponding to v_{ij} takes $O(1)$ arithmetic operations (see details in Section IV). Overall, the cross method enables channel estimation in $O(N \log N + r^2)$ arithmetic operations.

II. Chirp, Double-Chirp, and Triple-Chirp Sequences

In this section we introduce the chirp, double-chirp, and triple-chirp sequences, and discuss their correlation properties.

A. Definition of the Chirp Sequences

We have $N+1$ lines[5] in the *discrete delay-Doppler plane* $V = \mathbb{Z}_N \times \mathbb{Z}_N$. For each $a \in \mathbb{Z}_N$ we denote by $L_a = \{(\tau, a\tau); \tau \in \mathbb{Z}_N\}$ the line of finite slope a, and we denote by $L_\infty = \{(0,\omega);\ \omega \in \mathbb{Z}_N\}$ the line of infinite slope. To every line L_a, it corresponds the orthonormal basis for $\mathcal{H}$:

$$\mathcal{B}_{L_a} = \left\{ C_{L_{a,b}};\ b\ \in \mathbb{Z}_N \right\},$$

of *chirp sequences* associated with L_a, where

$$C_{L_{a,b}}[n] = e(2^{-1}an^2 - bn)/\sqrt{N}, n \in \mathbb{Z}_N.$$

To the line L_∞ it corresponds the orthonormal basis

$$\mathcal{B}_{L_\infty} = \left\{ C_{L_{\infty,b}}\ ;\ b\ \in \mathbb{Z}_N \right\},$$

of chirp sequences associated with L_∞, where

$$C_{L_{\infty,b}} = \delta_b,$$

denotes the Dirac delta sequence supported at b.

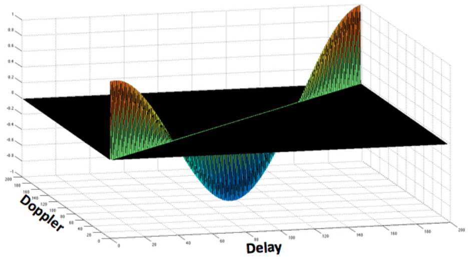


Fig. 5. Plot (real part) of $\mathcal{A}(C_{L_{1,1}}, C_{L_{1,1}})$, for chirp $C_{L_{1,1}}[n] = e[2^{-1}n^2 - n]$, associated with the line $L_1 = \{(\tau,\tau)\}$.

B. Chirps as Eigenfunctions of Heisenberg Operators

The Heisenberg operators (I-C.1) satisfy the commutation relations

$$\pi(\tau,\omega)\pi(\tau',\omega') = e(\omega\tau' - \tau\omega') \cdot \pi(\tau',\omega')\pi(\tau,\omega), \qquad \text{(II-B.1)}$$

for every $(\tau,\omega), (\tau',\omega') \in V$. In particular, for a given line $L \subset V$, we have the family of commuting operators $\pi(l)$, $l \in L$. Hence they admit an orthonormal basis $\mathcal{B}_L$ for $\mathcal{H}$ of common eigenfunctions. Important property of the chirp sequences is that for every chirp sequence $C_L \in \mathcal{B}_L$, there exists a character[6] $\psi_L : L \to \mathbb{C}^*$, i.e. $\psi_L(l + l') = \psi_L(l)\psi_L(l')$, $l, l' \in L$, such that

$$\pi_L(l)C_L = \psi_L(l)C_L, \text{ for every } l \in L.$$

This implies—see Figure 5—that for every $C_L \in \mathcal{B}_L$ we have

$$\mathcal{A}(C_L, C_L)[v] = \begin{cases} \psi_L(v) & \text{if } v \in L; \\ 0 & \text{if } v \notin L. \end{cases} \qquad \text{(II-B.2)}$$

It is not hard to see [4] that for distinct lines L, and M, and two chirps $C_L \in \mathcal{B}_L, C_M \in \mathcal{B}_M$ we have

$$|\mathcal{A}(C_L, C_M)[v]| = 1/\sqrt{N}, \qquad \text{for every } v \in V. \qquad \text{(II-B.3)}$$

C. Double-Chirp Sequences

For any two distinct lines $L, M \in V$, and two characters ψ_L, ψ_M on them, respectively, denote by C_L the chirp corresponding to L and ψ_L, and by C_M the chirp corresponding to M, and ψ_M. We define the *double-chirp* sequence

$$C_{L,M} = (C_L + C_M)/\sqrt{2}.$$

It follows from (II-B.2) and (II-B.3) that for the line $K = L$, or M, we have

$$\mathcal{A}(C_K, C_{L,M})[v] \approx \begin{cases} \psi_K(v)/\sqrt{2} & \text{if } v \in K; \\ 0 & \text{if } v \notin K. \end{cases}$$

[5] In these notes by a *line* $L \subset V$, we mean a line through $(0,0)$.

[6] We denote by $\mathbb{C}^*$ the set of non-zero complex numbers

D. *Triple-Chirp Sequences*

Consider three distinct lines $L, M, M^{\circ} \in V$, and three characters $\psi_L, \psi_M, \psi_{M^{\circ}}$ on them, respectively. Denote by C_L, C_M and $C_{M^{\circ}}$ the chirps corresponding to L, M and M°, and ψ_L, ψ_M, and $\psi_{M^{\circ}}$, respectively. We define the *triple-chirp* sequence

$$C_{L,M,M^{\circ}} = (C_L + C_M + C_{M^{\circ}})/\sqrt{3}.$$

It follows from (II-B.2) and (II-B.3) that for the line $K = L, M$ or M°, we have

$$\mathcal{A}(C_K, C_{L,M,M^{\circ}})[v] \approx \begin{cases} \psi_K(v)/\sqrt{3} & \text{if } v \in K; \\ 0 & \text{if } v \notin K. \end{cases}$$

III. Incidence Method

We describe—see Figure 3 for illustration—the incidence algorithm.

Incidence Algorithm

Input: Randomly chosen lines L, M, and M°, and characters $\psi_L, \psi_M, \psi_{M^{\circ}}$ on them, respectively. Echo $R_{L,M,M^{\circ}}$ of the triple-chirp $C_{L,M,M^{\circ}}$, threshold $T > 0$, and value of SNR.

Output: Channel parameters.

1) Compute $\mathcal{A}(C_M, R_{L,M,M^{\circ}})$ on L, obtain peaks[7] at $l_1, ..., l_{r_1}$.
2) Compute $\mathcal{A}(C_L, R_{L,M,M^{\circ}})$ on M, obtain peaks at $m_1, ..., m_{r_2}$.
3) Compute $\mathcal{A}(C_{M^{\circ}}, R_{L,M,M^{\circ}})$ on L, obtain peaks at $l_1^{\circ}, ..., l_{r_3}^{\circ}$.
4) Find $v_{ij} = l_i + m_j$ which solve $l_i + m_j \in M^{\circ} + l_k^{\circ}$, $1 \leq i \leq r_1$, $1 \leq j \leq r_2$, $1 \leq k \leq r_3$.
5) For every delay-Doppler parameter $v_{ij} = l_i + m_j$ found in the previous step, compute $\alpha_{v_{ij}} = \sqrt{3}\mathcal{A}(C_L, R_{L,M,M^{\circ}})[m_j]\psi_L(l_i)$. Return the parameter $(\alpha_{v_{ij}}, v_{ij})$.

IV. Cross Method

Let $C_{L,M}$ be the double-chirp sequence associated with the lines $L, M \subset V$, and the characters ψ_L, and ψ_M, on L, and M, correspondingly. We define *hypothesis* function $h : L \times M \to \mathbb{C}$ by

$$\begin{aligned} h(l,m) &= \mathcal{A}(C_L, R_{L,M})[m] \cdot \psi_L[l] \qquad \text{(IV-.1)} \\ &\quad - \mathcal{A}(C_M, R_{L,M})[l] \cdot e(\Omega[l,m]) \cdot \psi_M[m], \end{aligned}$$

where[8] $\Omega : V \times V \to \mathbb{Z}_N$ is given by $\Omega[(\tau,\omega),(\tau',\omega')] = \tau\omega' - \omega\tau'$. Below we describe—see Figure 4—the Cross Algorithm.

V. Conclusions

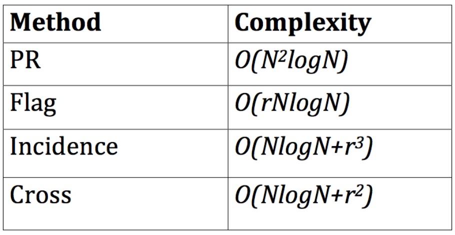
In these notes we present the incidence and cross methods for efficient channel estimation. These methods, in particular, suggest solutions to the arithmetic complexity problem. Low arithmetic complexity enables working with sequences of larger length N, and hence higher velocity resolution of channel parameters is plausible. We summarize these important features in Figure 6, and putting them in comparison with the pseudo-random (PR) and Flag methods.

Cross Algorithm

Input: Randomly chosen lines L, M, and characters ψ_L, ψ_M on them, respectively. Echo $R_{L,M}$ of the double-chirp $C_{L,M}$; thresholds $T_1, T_2 > 0$, and the value of SNR.

Output: Channel parameters.

1) Compute $\mathcal{A}(C_M, R_{L,M})$ on L, and take the r_1 peaks[9] located at points l_i, $1 \leq i \leq r_1$.
2) Compute $\mathcal{A}(C_L, R_{L,M})$ on M, and take the r_2 peaks located at the points $m_j, 1 \leq j \leq r_2$.
3) Find $v_{ij} = l_i + m_j$ which solve $|h(l_i, m_j)| \leq T_2\sqrt{2\log\log(N)}/\sqrt{N \cdot SNR}$, where $1 \leq i \leq r_1$, $1 \leq j \leq r_2$.
4) For every delay-Doppler parameter $v_{ij} = l_i + m_j$ found in the previous step, compute $\alpha_{v_{ij}} = \sqrt{2}\mathcal{A}(C_L, R_{L,M})[m_j]\psi_L(l_i)$. Return the parameter $(\alpha_{v_{ij}}, v_{ij})$.



Method	Complexity
PR	*O(N²logN)*
Flag	*O(rNlogN)*
Incidence	*O(NlogN+r³)*
Cross	*O(NlogN+r²)*

Fig. 6. Comparing methods, with respect to arithmetic complexity, for channels with r parameters.

Remark V-.1: Both new methods are robust to a certain degree of noise since they use the values of the ambiguity functions, which is a sort of averaging.

Acknowledgements. We are grateful to our collaborators A. Sayeed, and O. Schwartz, for many discussions related to the research reported in these notes.

References

[1] Fish A., Gurevich S., Hadani R., Sayeed A., and Schwartz O., Delay-Doppler Channel Estimation with Almost Linear Complexity. *Accepted for publication in IEEE Transaction on Information Theory (2013).*

[2] Golomb, S.W., and Gong G., Signal design for good correlation. For wireless communication, cryptography, and radar. *Cambridge University Press, Cambridge (2005).*

[3] Gurevich S., Hadani R., and Sochen N., The finite harmonic oscillator and its applications to sequences, communication and radar . *IEEE Transactions on Information Theory, vol. 54, no. 9, September 2008.*

[4] Howard S. D., Calderbank, R., and Moran W., The finite Heisenberg–Weyl groups in radar and communications. *EURASIP J. Appl. Signal Process (2006).*

[5] Rader C. M., Discrete Fourier transforms when the number of data samples is prime. *Proc. IEEE 56, 1107–1108 (1968).*

[6] Tse D., and Viswanath P., Fundamentals of Wireless Communication. *Cambridge University Press (2005).*

[7] Verdu S., Multiuser Detection, *Cambridge University Press (1998).*

[7] We say that at $v \in V$ the ambiguity function of f and g has *peak,* if $|\mathcal{A}(f,g)[v]| > T\sqrt{2\log\log N}/\sqrt{N \cdot SNR}$.

[8] In linear algebra Ω is called *symplectic form.*

[9] We say that at $v \in V$ the ambiguity function of f and g has *peak*, if $|\mathcal{A}(f,g)[v]| > T_1\sqrt{2\log\log N}/\sqrt{N \cdot SNR}$.

Simulation of Birth-Death Dynamics in Time-Variant Stochastic Radio Channels

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, Denmark

***Abstract*—We consider an analytically tractable class of time-variant stochastic radio channel models. All models in this class are designed such that individual multipath components emerge and vanish according to a temporal birth-death process $L(\cdot)$. This birth-death process is governed by two facilitating assumptions: $i)$ stationary emergence times, and $ii)$ i.i.d. lifetimes. Multipath channel models with such temporal birth-death dynamics have appeared in the literature several times. More specifically, such channel models have appeared with assumption $i)$ specialized to that of a Poisson point process and with assumption $ii)$ specialized to that of i.i.d. exponentials. So far, these two special-case assumptions have seemingly been invoked by default without justification or clarification for their necessity. Here, we establish and justify their essential necessity from a simulation practical point of view. Specifically, we obtain a tractable and exact (non-approximate) Markovian simulation recipe for drawing realizations of time-variant stochastic radio channels with temporal birth-death dynamics.**

I. Introduction and Preliminaries

A simple, flexible, and commonly used stochastic model for time-variant channel transfer functions is given by [1]

$$H(t,f) = \sum_{\ell=1}^{L(t)} \alpha_\ell(t) \exp\big(-j2\pi f \tau_\ell(t)\big). \tag{1}$$

The integer-valued random variable $L(t)$ gives the instantaneous number of path components at time t, $\alpha_\ell(t)$ is the random complex-valued gain of the ℓ'th path, and $\tau_\ell(t)$ is the associated random propagation delay. Time- and space-varying multipath propagation phenomena, e.g. path components which emerge and vanish, occur partially due to the movements of transmitter, receiver, and surrounding scatterers [1]. Transitions of the random process $L(\cdot)$ reflect when different path components emerge and vanish. The random process $L(\cdot)$ can be generated in numerous ways, for instance according to the following tractable assumptions:

$i)$ **Stationary emergences:** The collection of time instances where new path components emerge forms a stationary point process on the real line.

$ii)$ **i.i.d. lifetimes:** The non-negative lifetimes (or periods) of individual path components are i.i.d.

In [2] we show how the assumptions $i)$ and $ii)$ can be conveniently incorporated using an approach based on *spatial point processes*[1] [3]. Specifically, denote by Y the *one-dimensional* point process from $i)$ and denote by $\{p_y : y \in Y\}$ the collection of *non-negative* periods from $ii)$. The subscript y on each period p_y serves as an identifier for its underlying point. By construction, the random collection $\{(y,p_y) : y \in Y\}$ is a *marked point process* [3] with (stationary) points in $\mathbb{R}$ and (i.i.d.) marks in $\mathbb{R}_+$. Alternatively, this marked point process can be viewed as a *two-dimensional* point process

$$X := \{(y,p_y) : y \in Y\}, \tag{2}$$

i.e. as an *unmarked* point process with points in $\mathbb{R}\times\mathbb{R}_+$. Then, each two-dimensional point $\boldsymbol{x} = (y,p) \in X$ has components y and p interpreted as "birth time" and "lifetime", respectively, and a subscript identifier on p is no longer needed.

In terms of the two-dimensional point process X, the channel model in (1) can now be reformulated as [2]

$$H(t,f) = \sum_{\boldsymbol{x}\in X} \mathbb{1}[\boldsymbol{x} \in B_t]\alpha_{\boldsymbol{x}}(t)\mathrm{e}^{-j2\pi f\tau_{\boldsymbol{x}}(t)}, \tag{3}$$

$$L(t) = |X \cap B_t| = \sum_{\boldsymbol{x}\in X} \mathbb{1}[\boldsymbol{x} \in B_t], \quad t \in \mathbb{R}, \tag{4}$$

where $|\cdot|$ denotes set cardinality while $\mathbb{1}[\cdot]$ denotes a generic *indicator function* taking value one if the logical statement in brackets is fulfilled and zero otherwise. The time-indexed quantity B_t in (3) and (4) is the triangular-shaped region

$$B_t := \big\{(y,p) : y \le t,\ y + p > t\big\} \subset \mathbb{R}\times\mathbb{R}_+. \tag{5}$$

The relationship in (4) states that $L(t)$ is equal to the random number of points from X falling in the region B_t, see Fig. 1. An arbitrary point $\boldsymbol{x} = (y,p) \in X$ contributes to the value of the sum in (4) when it *emerges before* time t (i.e. $y \le t$) and *vanishes after* time t (i.e. $y + p > t$). As a consequence of $i)$ and $ii)$, the (continuous-time) temporal birth-death process $L(\cdot)$ is strict-sense stationary [2], [4].

Notice how the collections $\{\alpha_\ell(\cdot)\}$ and $\{\tau_\ell(\cdot)\}$ of random processes from (1) have been "substituted" in (3) by the collections $\{\alpha_{\boldsymbol{x}}(\cdot)\}$ and $\{\tau_{\boldsymbol{x}}(\cdot)\}$. The previous collections are now conveniently indexed using the points from X as proposed in [2]. The representation in (3) inherits several analytical benefits compared to the traditional representation in (1), especially in terms of the ability to "track" individual path components across time due to the point process-based indexing technique, see Fig. 1.

In the literature [5]–[8], assumptions $i)$ and $ii)$ were originally introduced in more restrictive versions:

$i)^\dagger$ Special-case of $i)$: Poisson point process [9].

$ii)^\dagger$ Special-case of $ii)$: Exponential distribution.

In the rest of this paper we restrict to the special-case assumptions $i)^\dagger$ and $ii)^\dagger$ and treat these in detail. In the earliest channel modeling contribution by S. J. Papantoniou [5], the

[1] The spatial point process approach used in [2] serves to bypass the *enumeration issue* in (1) which occurs since every transition of $L(\cdot)$ leaves the need for a non-trivial reordering or bookkeeping of all ℓ-indexed quantities.

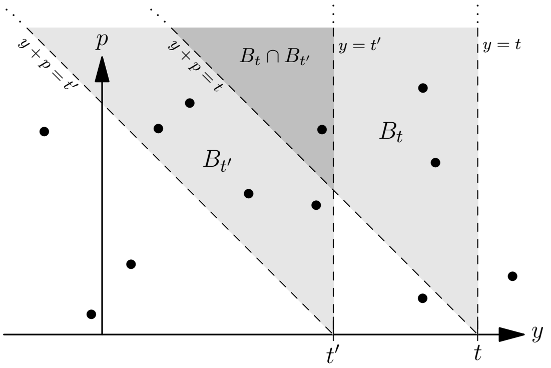


Figure 1. Each black bullet represents a point from the two-dimensional point process X. Points falling within the region B_t indicate path components which are *active* in the radio channel (3) at time t. Those falling in $B_t \cap B_{t'}$ are contributing to (3) at time instances t', t, and everywhere in-between.

construct via $i)^\dagger$ and $ii)^\dagger$ is identified[2] as an $M/M/\infty$ queue [4, Sec. 16-2]. Accordingly, all known properties of this particular queue apply directly to the temporal birth-death process $L(\cdot)$, e.g. that $L(t)$ is Poisson distributed for each fixed $t \in \mathbb{R}$. The original motivation for using $i)^\dagger$ and $ii)^\dagger$ is (quoting Papantoniou [5, Sec. 2.2.6]): "that these assumptions endow the model with *simple mathematics*".

Later contributions such as [7] and [8] have seemingly employed the special-case assumptions $i)^\dagger$ and $ii)^\dagger$ by default. In addition, [8] proposes cumbersome implementation guidelines for computer simulation, e.g. a heuristic channel initialization scheme as well as a procedure for approximating the continuous-time process $L(\cdot)$ on a discrete sampling grid. Yet, both of these approximate simulation guidelines can be circumvented and substituted by exact procedures upon taking direct advantage of the facilitating aspects of the point process perspective in (4). As opposed to [8], the earlier contributions [5]–[7] entirely omit discussions related to computer simulation of their proposed (birth-death) channel models.

This paper presents three main contributions. Firstly, we justify and clarify Papantoniou's claim from the point of view of computer simulation. As shown in [2], $i)^\dagger$ and $ii)^\dagger$ are indeed analytically tractable in general, but as shown here they prove as well highly facilitating for simulation purposes. Secondly, we show the utility of the point process view in (4) with respect to the actual implementation of the temporal birth-death process $L(\cdot)$, especially in the context of radio channel modeling. Knowing that $L(\cdot)$ can be seen from a queuing theory perspective does not straightforwardly aid in being able to track the individual path components in (1). Thirdly, we show how the memoryless property of the exponential distribution can be exploited to represent the continuous-time birth-death process $L(\cdot)$ on an arbitrary discrete sampling grid (which is needed in practice). Compared to [8, Sec. III-C] our representation is exact, i.e. it does not rely on approximations such as disregarding "tiny" probabilities of multiple jump events within "tiny" intervals of time.

II. Simulation of the Temporal Birth-Death Process $L(\cdot)$ in Time-Varying Radio Channels

Under the special-case assumptions $i)^\dagger$ and $ii)^\dagger$ there are several (equivalent) ways to view the birth-death process $L(\cdot)$. The process can be seen as an $M/M/\infty$ queue, as a continuous-time Markov chain [10, Sec. 7.4], as generalized shot-noise, but it can also be seen via the point process perspective in (4). Specifically, it can be seen as the "time-sliding" region count displayed in Fig. 1. All of the aforementioned views have their individual advantages and drawbacks.

In queuing theory it is usually the queue itself which is of primary interest, not the individual customers (they just temporally alter the length of the queue). When considering the channel models in (1) and (3) the situation is different. It makes a crucial difference when we need to be able to track the individual path components across time. In radio channel characterization we often wish to "correlate" the channel with itself at different time-frequency instances. Hence, it is important to be able to identify and track if path components are still present, if new ones have emerged, or if some have vanished in-between any two time instances t' and t. In Fig. 1 the instantaneous counts are $L(t') = 5$ and $L(t) = 3$ but only one path component is shared. Due to this readily accessible (graphical) insight, the point process-based representation in (3) is beneficial for channel modeling purposes [2].

A. Notation and Properties of Poisson Point Processes

Under assumption $i)^\dagger$ the random collection Y in (2) is a stationary Poisson point process with constant *intensity function* $\varrho_Y(\cdot)$. Thus, $\varrho_Y(y) = \lambda_{\mathrm{e}}$ for all $y \in \mathbb{R}$, for some positive constant λ_{e} (subscript abbreviating "emerge").

By $ii)^\dagger$ the collection of periods $\{p_y : y \in Y\}$ is such that

$$p_y \overset{\text{i.i.d.}}{\sim} f_{\text{period}}(\cdot), \quad f_{\text{period}}(p) = \mathbb{1}[p \geq 0]\lambda_{\mathrm{v}} \exp(-\lambda_{\mathrm{v}} p),$$

for some positive rate parameter λ_{v} (subscript abbreviating "vanish"). Since the periods are mutually independent it follows that the two-dimensional point process X in (2) is also a Poisson point process (by the *Marking Theorem* for Poisson point processes [9, Sec. 5.2]). The Poisson point process X is *inhomogeneous* with intensity function given by [9, Sec. 5.2]

$$\varrho_X(\boldsymbol{x}) = \varrho_X(y,p) = \varrho_Y(y) f_{\text{period}}(p) = \lambda_{\mathrm{e}}\lambda_{\mathrm{v}} \exp(-\lambda_{\mathrm{v}} p),$$

i.e. this intensity function is constant with y and decays exponentially with p. By the equality in (4) and the fact that X is a Poisson point process, it follows immediately that $L(t)$ is a Poisson distributed random variable for any fixed $t \in \mathbb{R}$. The mean of $L(t)$ is obtained by integrating the intensity function $\varrho_X(\cdot)$ across the region B_t, i.e.

$$\mathbb{E}[L(t)] = \mathbb{E}\Bigg[\sum_{\boldsymbol{x} \in X} \mathbb{1}[\boldsymbol{x} \in B_t]\Bigg] = \int_{B_t} \varrho_X(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \frac{\lambda_{\mathrm{e}}}{\lambda_{\mathrm{v}}},$$

which does not depend on time t, in accordance with $L(\cdot)$ being strict-sense stationary. The property of $L(t)$ being Poisson distributed with mean $\lambda_{\mathrm{e}}/\lambda_{\mathrm{v}}$ was obtained by Papantoniou [5] using arguments from queuing theory. In [2], this property is readily obtained as a result of the point process perspective.

[2] More precisely, [5] presents a *space-varying* approach which simplifies to a time-varying model like (1) and (3) upon assuming a receiver trajectory with constant velocity vector.

B. Initialization of $L(\cdot)$ at Time t'

Suppose we wish to initialize the channel model in (3) at some time instance $t' \in \mathbb{R}$, e.g. $t' = 0$. In particular, we then have to initialize the temporal birth-death process $L(\cdot)$ at this particular time instance. Since $L(t')$ follows a Poisson distribution with mean $\lambda_\mathrm{e}/\lambda_\mathrm{v}$ we can indeed generate $L(t')$ accordingly. Conditioned on $L(t')$, the points

$$\left\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{L(t')}\right\} = X \cap B_{t'}, \quad \boldsymbol{x}_\ell = (y_\ell, p_\ell), \tag{6}$$

are to be drawn i.i.d. according to the (conditional) joint pdf

$$\begin{aligned} f(y,p;t') &= \frac{\mathbb{1}[(y,p) \in B_{t'}]\varrho_X(y,p)}{\displaystyle\iint_{B_{t'}} \varrho_X(\tilde{y},\tilde{p})\mathrm{d}\tilde{y}\mathrm{d}\tilde{p}} \\ &= \mathbb{1}[0 \le t' - y < p]\lambda_\mathrm{v}^2 \exp(-\lambda_\mathrm{v} p), \end{aligned} \tag{7}$$

i.e. according to a *truncated* and *normalized* version of the intensity function $\varrho_X(\cdot)$, see [3] and [9]. We stress that the integer-labeling employed in (6) does not indicate any ordering of the points whatsoever. Notice also that the set $X \cap B_{t'}$ in (6) can potentially be empty since $L(t')$ was drawn from a Poisson distribution. The individual components y_ℓ and p_ℓ of each two-dimensional point $\boldsymbol{x}_\ell$ are obviously dependent due to the triangular shape of the region $B_{t'}$. The intuitive argument is of course that we have conditioned on the points in (6) to be "active" at time t' (i.e. they have not yet vanished).

By a change of variables and by marginalizing the (conditional) joint pdf in (7) we readily find that

$$t' - y_\ell \sim \mathrm{Exp}(\lambda_\mathrm{v}), \quad y_\ell + p_\ell - t' \sim \mathrm{Exp}(\lambda_\mathrm{v}), \tag{8}$$

i.e. the lifespan which has already elapsed (from the past) and the lifespan which remains (in the future) are both exponentially distributed, and in fact they are *independent*! Thus, the (conditional) marginal distribution of the lifetime p_ℓ is not exponential, rather its distribution is that of the sum of the two independent exponentials in (8) and hence $p_\ell \sim \Gamma(2, \lambda_\mathrm{v})$, namely a gamma distribution[3]. This fact can also be verified by direct marginalization in (7). Conditioned on the lifetime p_ℓ, the corresponding emergence time y_ℓ has a uniform distribution (into the past), i.e. $y_\ell | p_\ell \sim \mathcal{U}(t' - p_\ell, t')$. This is not surprising since by $i)^\dagger$, the collection of emergence times originates from a stationary Poisson point process.

The above procedure describes how to correctly initialize the non-negative integer $L(t')$ together with the individual components in (6). For comparison, [8, Sec. III-C] always initializes $L(t') = 0$ followed by a temporal "burn-in/forerun" to allow the birth-death process to evolve and stabilize before running the actual simulation. The procedure outlined in this paper allows for instantaneous and exact initialization of the channel in (3).

C. Temporal Evolution of $L(\cdot)$ in the Interval $[t', t'']$

Suppose that the birth-death process $L(\cdot)$ has been initialized at time t' in a state of equilibrium as described in the previous subsection. We can now arbitrarily select a stopping time $t'' > t'$ and then generate a realization of the point process X restricted to the unbounded rectangular strip $[t', t''] \times \mathbb{R}_+$, see Fig. 2. To generate this restricted realization of X we draw a Poisson distributed number with mean $\mathbb{E}\big[\big|X \cap ([t', t''] \times \mathbb{R}_+)\big|\big] = \lambda_\mathrm{e}(t'' - t')$, and then we distribute this amount of points inside $[t', t''] \times \mathbb{R}_+$ according to i.i.d. draws from a *truncated* and *normalized* version of the intensity function $\varrho_X(\cdot)$. Essentially, this means that we need to generate pairs of uniformly and exponentially distributed random variables (all mutually independent). Then, to calculate the corresponding realization of the temporal birth-death process $L(\cdot)$, we simply count points while "sliding" the triangular region B_t in (5) from $t = t'$ until $t = t''$, see Fig. 2.

The procedure above is suitable if we know in advance the necessary duration of our simulation. Lengthy simulations require in general a vast amount of numbers to be stored. However, due to Markovian properties of $L(\cdot)$, there is a practically useful alternative to the above simulation procedure.

D. Markovian Temporal Evolution of $L(\cdot)$ in $[t', \infty)$

Suppose that the birth-death process $L(\cdot)$ has been initialized at time t' such that $L(t')$ has been drawn from a Poisson distribution. Now we do not explicitly generate the individual points in (6) anymore. Instead, we make use of the property in (8), namely that the remaining lifespan $y_\ell + p_\ell - t'$ of each path component has an exponential distribution. What occurred in the past is no longer relevant, i.e. we are not interested in knowing when individual path components emerged. In fact, we are now concerned only with the next transition of $L(\cdot)$ which occurs sometime in the future. *We then maintain this concern one single transition at a time*. There are only two possibilities for the next transition since the point process X in (2) has *almost surely* no repetitions of points. Either a new path component emerges or a single of the existing ones vanishes. Thus, the birth-death process $L(\cdot)$ experiences a random "*increment*" from the set $\{-1, +1\}$ at a random time instance in the future. Once this increment has been assigned we wait yet another random time instance until the next increment from $\{-1, +1\}$ arrives, and so on.

If the random "*waiting times*" between consecutive transitions are exclusively exponentially distributed, it means that $L(\cdot)$ forms a continuous-time Markov chain [10, Sec. 7.4]. This is indeed the case (and we already know that). After initialization at time t', the first transition of $L(\cdot)$ occurs either when a new path component emerges (after a random time E), or when one of the existing components vanish (their *remaining lifetimes* can conveniently be denoted by $V_1, V_2, \ldots, V_{L(t')}$). Accordingly, after initialization at time t', the first transition of $L(\cdot)$ occurs at time $t' + \min\{V_1, V_2, \ldots, V_{L(t')}, E\}$, where

$$V_\ell \sim \mathrm{Exp}(\lambda_\mathrm{v}), \quad \ell = 1, 2, \ldots, L(t'), \quad E \sim \mathrm{Exp}(\lambda_\mathrm{e}). \tag{9}$$

The $L(t')+1$ random variables in (9) are mutually independent and it is well-known that the minimum of a fixed number of independent exponentials again has an exponential distribution [10, Sec. 3.10.1]. Thus, conditioned on $L(t)$ and now for an arbitrary time instance t, we conveniently define a waiting-time random variable

$$T_{L(t)} := \min\{V_1, V_2, \ldots, V_{L(t)}, E\} \sim \mathrm{Exp}\big(L(t)\lambda_\mathrm{v} + \lambda_\mathrm{e}\big)$$

[3] Compare this "conditional property" to that of the lifetime of a newly emerged path component. By $ii)^\dagger$, the lifetime of a newly emerged path component should be assigned from an exponential distribution.

as well as the indicator $I_{L(t)} := \mathbb{1}\big[E < \min\{V_1, \ldots, V_{L(t)}\}\big]$. Then one can readily verify that

$$\Pr\big(I_{L(t)} = 1\big) = \Pr\big(E < \min\{V_\ell\}\big) = \frac{\lambda_\text{e}}{L(t)\lambda_\text{v} + \lambda_\text{e}},$$

and in fact, $T_{L(t)}$ and $I_{L(t)}$ are *independent*! Hence, we generate the random variable $T_{L(t)}$ with its distribution depending on $L(t)$. The associated increment from $\{-1, 1\}$ is independently determined from the realization of $I_{L(t)}$. The larger the value of $L(t)$ the smaller the expected waiting-time $\mathbb{E}[T_{L(t)}]$ until transition. Additionally, the more components currently attending in the channel, the more likely it is that one of the existing path components will soon vanish. In case the increment is minus one, i.e. $I_{L(t)} = 0$, then we simply remove an arbitrary component uniformly at random (since $V_1, V_2, \ldots, V_{L(t)}$ are i.i.d.). Overall, instead of simulating $L(t) + 1$ random variables we can always do with simulating only two random variables. Thus, we sequentially generate the transition times and increments of the process $L(\cdot)$.

In practical simulation studies we often need to represent the channel in (3) on a grid discretized in time and in frequency. Suppose we want to simulate the temporal birth-death process $L(\cdot)$ on some *regular* or *irregular* sampling grid $\big(t_n : n \in \mathbb{N}_0\big)$ with $t_0 = t'$ (initialization time) and $t_n < t_{n+1}$ for all n. Doing so yields the sequence

$$L(t_0),\ L(t_1), \ldots, L(t_n), \ldots, \tag{10}$$

and in case of a regular sampling grid, the fixed time-step parameter $t_{n+1} - t_n = \Delta t > 0$ could for instance be dictated by the *signalling* or the *sampling* period of a particular communication system (e.g. OFDM-based). In any case, the transitions of $L(\cdot)$ occur in continuous time and not on the discrete sampling grid $\big(t_n : n \in \mathbb{N}_0\big)$. Yet, by the memoryless property of the exponential distribution we can repeatedly "reset the clock" and sequentially (step-by-step) simulate the sequence in (10). Pseudo-code instructions read as follows:

```
Initialization:
Define $\lambda_\text{e} > 0$, $\lambda_\text{v} > 0$, and $t_n$ for all $n$;        (parameters)
$L \sim$ Poisson$(\lambda_\text{e}/\lambda_\text{v})$;                         (count variable)
$t = t_0$; $L(t) = L$;                                   (assign initial count)
Temporal evolution across the sampling grid:
for $n = 0, 1, 2, \ldots$
  $T_\text{cumulate} = 0$;                                    (reset the clock)
  while ($T_\text{cumulate} < t_{n+1} - t_n$){                (trivially satisfied at first)
    $T \sim \text{Exp}\big(L\lambda_\text{v} + \lambda_\text{e}\big)$;                       (time until next event)
    $T_\text{cumulate} = T_\text{cumulate} + T$;                        (time accumulator)
    if ($T_\text{cumulate} > t_{n+1} - t_n$){break;}           (exit while-loop)
    else{                                            (revisit while-loop)
      $U \sim \mathcal{U}(0, 1)$;                                  (a probability)
      if $\big(U < \frac{\lambda_\text{e}}{L\lambda_\text{v} + \lambda_\text{e}}\big)$\{$L = L + 1$;\}              (new emergence)
      else{$L = L - 1$;}                        (one component vanished)
    }
  end while;
  $t = t_{n+1}$;                                (move one time-step ahead)
  $L(t) = L$;                                          (assign count)
end for;
```

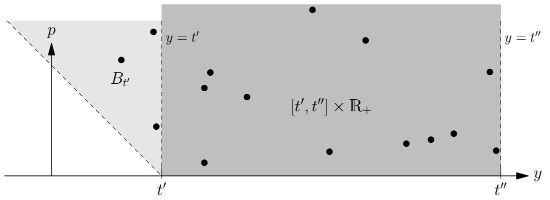


Figure 2. The unbounded rectangular strip $[t', t''] \times \mathbb{R}_+$ from Sec. II-C.

The while-loop in the pseudo-code is present to account for the fact that multiple transition events can occur between two consecutive sampling points t_n and t_{n+1}. The approximate approach in [8, Sec. III-C] is based on a regular sampling grid with time-step parameter $\Delta t > 0$ sufficiently small so that in each step the probability of multiple transition events can be neglected in practice. The approach in this paper is exact (non-approximate) and valid for both regular and irregular sampling grids. Hence, the simulation procedure described here can also be used for block-wise burst communications with long empty (or silent) gaps in-between consecutive transmission bursts.

E. Remarks on Generalization Attempts

The key feature of assumption $i)^\dagger$ is that the distribution of $L(t) = |X \cap B_t|$ is known to be Poisson, even without assumption $ii)^\dagger$ (this is the so-called $M/G/\infty$ queue). In this case we still know how to generate the points in (6) since these belong to the Poisson point process X, see [2]. Hence, assumption $i)^\dagger$ is effectively indispensable since the distribution of $L(t)$ is (in general) intractable for more sophisticated types of point processes. The key feature of assumption $ii)^\dagger$ is expressed in (8), namely that each remaining lifespan has an exponential distribution no matter the current age of the considered path component (inherited from the memoryless property of the exponential distribution). If a different lifetime distribution is employed the remaining lifespan of each "active" path component will inevitably depend on age.

Acknowledgment

This work was funded by the FP7-STREP Wireless Hybrid Enhanced Mobile Radio Estimators - Phase 2 (WHERE2, ICT-248894), the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (ICT-318306), and the 4GMCT cooperative research project co-financed by Intel Mobile Communications, Agilent Technologies, Aalborg University and the Danish National Advanced Technology Foundation.

References

[1] H.Hashemi, "The Indoor Radio Propagation Channel", Proc. IEEE, '93.

[2] M. L. Jakobsen, T. Pedersen, and B. H. Fleury, "Analysis of Stochastic Radio Channels with Temporal Birth-Death Dynamics: A Marked Spatial Point Process Perspective", IEEE Trans. on Anten. and Prop., Submitted July 2013.

[3] A. J. Baddeley, "Spatial Point Processes and their Applications" (in "Stochastic Geometry - Lecture Notes in Mathematics"), Springer, 2007.

[4] A. Papoulis, "Probability, Random Variables and Stochastic Processes", McGraw-Hill, 3rd ed., 1991.

[5] S. J. Papantoniou, "Modelling the Mobile Radio Channel", PhD thesis, ETHZ, Switzerland, 1990.

[6] H. Iwai and Y. Karasawa, "Wideband Propagation Model for the Analysis of the Effect of the Multipath Fading on the Near-Far Problem in CDMA Mobile Radio Systems", IEICE Trans. Commun., 1993.

[7] B. H. Fleury, U. P. Bernhard and R. Heddergott, "Advanced Radio Channel Model for Magic WAND", ACTS Mobile Telecomm. Summit, 1996.

[8] T. Zwick, C. Fischer and W. Wiesbeck, "A Stochastic Multipath Channel Model Including Path Directions for Indoor Environments", IEEE J. Selec. Areas in Commun., 2002.

[9] J. F. C. Kingman, "Poisson Processes", Oxford University Press, 1993.

[10] P. Olofsson, "Probability, Statistics, and Stochastic Processes", John Wiley & Sons, 2005.

Asymmetric Compute-and-Forward with CSIT

Jingge Zhu and Michael Gastpar
School of Computer and Communication Sciences, EPFL
Lausanne, Switzerland
Email: {jingge.zhu, michael.gastpar}@epfl.ch

***Abstract*—We present a modified compute-and-forward scheme which utilizes Channel State Information at the Transmitters (CSIT) in a natural way. The modified scheme allows different users to have different coding rates, and use CSIT to achieve larger rate region. This idea is applicable to all systems which use the compute-and-forward technique and can be arbitrarily better than the regular scheme in some settings.**

I. Introduction

The compute-and-forward scheme [1] is a novel coding scheme for Gaussian networks which takes advantage of the linear structure of lattice codes and the additive nature of Gaussian interference networks. The main idea of compute-and-forward is to decode linear combinations of messages rather than the messages themselves at the receivers. One nice feature of this scheme is that the Channel State Information is not explicitly required at the transmitters, making this scheme attractive to practical considerations. But on the other hand, it is not clear how CSI can be used at the transmitters if the compute-and-forward idea is to be applied. Furthermore, the same lattice code is used for every user, preventing the scheme from exploiting asymmetries of the networks.

In this work we present a modified compute-and-forward scheme with asymmetric message rates which makes CSIT useful. This scheme also extends the concept of the computation rate to a more general definition of the computation rate tuple, which allows flexibility in controlling the individual message rates of different users. The main idea relies on the observation, roughly speaking, that the transmitted lattice codeword does not have to lie in the lattice which is used for lattice coding at the transmitters.

We use the notation $[a:b]$ to denote a set of increasing integers $\{a, a+1, \ldots, b\}$, log to denote $\log_2$ and $\log^+(x)$to denote the function $\max\{\log(x), 0\}$. We also use $x_{1:K}$ to denote a set of numbers $\{x_1, x_2, \ldots, x_K\}$.

II. problem statement

We consider a interference network with K transmitters and M relays. The discrete-time real Gaussian channel has the following vector representation

$$\mathbf{y}_m = \sum_{k=1}^{K} h_{mk}\mathbf{x}_k + \mathbf{z}_m, \quad m \in [1:M]$$

with $\mathbf{y}_m \in \mathbb{R}^n, \mathbf{x}_k \in \mathbb{R}^n, h_{mk} \in \mathbb{R}$ denoting the channel output of relay m, channel input of transmitter k and the channel gain, respectively. The Gaussian white noise with unit variance is denoted by $\mathbf{z}_m \in \mathbb{R}^n$. We impose the same power constraint $\mathbb{E}\{||\mathbf{x}_k||^2\} \leq nP$ on all the transmitters.

The message of user k is represented by a point in $\mathbb{R}^n$ denoted by $\mathbf{t}_k$, which is an element of the codebook $\mathcal{C}_k$ of user k with *message rate* $R_k := \frac{1}{n}\log|\mathcal{C}_k|$.

Each transmitter is equipped with an encoder $\mathcal{E}_k$ which maps its message into the channel input as $\mathbf{x}_k = \mathcal{E}_k(\mathbf{t}_k)$. Each relay m has a decoder $\mathcal{D}_m$ which uses the channel output $\mathbf{y}_m$ to decode a function of all the messages $\mathbf{t}_k, k \in [1:K]$ as $f_m(\mathbf{t}_{[1:K]}) = \mathcal{D}_m(\mathbf{y}_m)$. Here we only consider the function of the form $f_m(\mathbf{t}_{[1:K]}) = \sum_{k=1}^{K} a_{mk}\mathbf{t}_k$ with integer a_{mk} for all $m \in [1:M], k \in [1:K]$. We use $\mathbf{a}_m$ to denote the column vector $[a_{m1}, \ldots, a_{mK}]^T$.

We say a *computation rate tuple* $(R_1, \ldots, R_K)$ *with respect to the function* f_m is achievable, if the relay m can decode the function f_m reliably, namely $\Pr\left(\mathcal{D}_m(\mathbf{y}_m) \neq f_m(\mathbf{t}_{[1:K]})\right) < \delta$ for any $\delta > 0$, with R_k being the message rate of the user k. In the network, we require all the relays to decode their desired functions. We say *a computation rate tuple* $(R_1, \ldots, R_K)$ *w. r. t. the set of functions* $f_m, m \in [1:M]$ is achievable, if $\Pr\left(\mathcal{D}_m(\mathbf{y}_m) \neq f_m(\mathbf{t}_{[1:K]}),\ \text{for all } m \in [1:M]\right) < \delta$ holds for any $\delta > 0$ with R_k being the message rate of user k. In the following we will study the computation rate tuple achieved by a modified compute-and-forward scheme.

III. Lattice codes construction

A lattice Λ is a discrete subgroup of $\mathbb{R}^n$ with the property that if $\mathbf{t}_1, \mathbf{t}_2 \in \Lambda$, then $\mathbf{t}_1 + \mathbf{t}_2 \in \Lambda$. More details about lattice and lattice codes can be found in [2]. Define the lattice quantizer $Q_\Lambda : \mathbb{R}^n \to \Lambda$ as $Q_\Lambda(\mathbf{x}) = \text{argmin}_{\mathbf{t}\in\Lambda} ||\mathbf{t} - \mathbf{x}||$ and define the fundamental Voronoi region of the lattice to be $\mathcal{V} := \{\mathbf{x} \in \mathbb{R}^n : Q_\Lambda(\mathbf{x}) = \mathbf{0}\}$. The modulo operation gives the quantization error: $[\mathbf{x}]\text{mod } \Lambda = \mathbf{x} - Q_\Lambda(\mathbf{x})$. Two lattices Λ and Λ' are said to be nested if $\Lambda' \subseteq \Lambda$.

Let $\Lambda_1, \ldots, \Lambda_M$ be M nested lattice codes constituting a nested lattice chain in which all lattices are simultaneously good in the sense of [2]. This chain can be constructed as shown in [3] and the order of the chain will be determined later. Relay m will perform the lattice decoding with respect to the lattice Λ_m.

Let $\beta_1, \ldots, \beta_K$ be K positive numbers. We can construct K nested lattices such that $\Lambda_k^s \subseteq \Lambda_c$ for all k where

This work was supported in part by the European ERC Starting Grant 259530-ComCom.

Λ_c denotes the coarsest lattice among $\Lambda_1, \dots, \Lambda_M$. We let Λ_k^s to be simultaneously good and with second moment $\frac{1}{n}\int_{\mathcal{V}_k^s} ||\mathbf{x}||^2 \, d\mathbf{x} = \beta_k^2 P$ where $\mathcal{V}_k^s$ denotes the Voronoi region of the lattice Λ_k^s for $k \in [1:K]$. The lattice Λ_k^s is used as the shaping region for the codebook of user k.

For each transmitter k, we construct the codebook

$$\mathcal{C}_k = \Lambda_{m(k)} \cap \mathcal{V}_k^s \tag{1}$$

where $m(k) \in \{1, \dots, M\}$ hence $\Lambda_{m(k)}$ is the decoding lattice at one of the M relays. We will determine which decoding lattice to choose for transmitter k, i.e., the expression of $m(k)$, in the next section. The message rate of user k is

$$R_k = \frac{1}{n}\log|\mathcal{C}_k| = \frac{1}{n}\log\frac{\text{Vol }(\mathcal{V}_k^s)}{\text{Vol }(\mathcal{V}_{m(k)})} \tag{2}$$

where $\mathcal{V}_{m(k)}$ is the Voronoi region of the fine lattice $\Lambda_{m(k)}$.

IV. A MODIFIED COMPUTE-AND-FORWARD SCHEME

When the message (codeword) $\mathbf{t}_k$ is given to encoder k, it forms its channel input as follows

$$\mathbf{x}_k = [\mathbf{t}_k/\beta_k + \mathbf{d}_k] \bmod \Lambda_k^s/\beta_k$$

where $\mathbf{d}_k$ is called a *dither* which is a random vector uniformly distributed in the scaled Voronoi region $\mathcal{V}_k^s/\beta_k$. As pointed out in [2], $\mathbf{x}_k$ is independent from $\mathbf{t}_k$ and also uniformly in Λ_k^s/β_k hence has average power P for all k.

To demonstrate the proposed approach, we first assume there is only one relay, $m = M = 1$. For now there is only one decoding lattice hence the codebooks of all the users are constructed using the same fine lattice and we denote it as $\Lambda_{m(k)} = \Lambda$ for all k.

Theorem 1: Assume there is only one relay m. For any given set of positive numbers $\beta_1, \dots, \beta_K$, there exists lattice codes $\mathcal{C}_1, \dots, \mathcal{C}_K$ such that the achievable computation rate tuple $(R_1, \dots, R_K)$ with respect to the function $f_m = \sum_k a_{mk}\mathbf{t}_k$ at relay m is given by

$$\begin{aligned} &R_k < r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K}) \\ &:= \left[\frac{1}{2}\log\left(||\tilde{\mathbf{a}}_m||^2 - \frac{P(\mathbf{h}_m^T\tilde{\mathbf{a}}_m)^2}{1+P\,||\mathbf{h}_m||^2}\right)^{-1} + \frac{1}{2}\log\beta_k^2\right]^+ \end{aligned} \tag{3}$$

for all k with $\tilde{\mathbf{a}}_m := [\beta_1 a_{m1}, ..., \beta_K a_{mK}]$ and $a_{mk} \in \mathbb{Z}$ for all $k \in [1:K]$.

Proof: At the decoder we form

$$\begin{aligned} \tilde{\mathbf{y}}_m &:= \alpha_m\mathbf{y}_m - \sum_k a_{mk}\beta_k\mathbf{d}_k \\ &= \sum_k a_{mk}\left(\beta_k(\mathbf{t}_k/\beta_k + \mathbf{d}_k) - \beta_k Q_{\Lambda_k^s/\beta_k}(\mathbf{t}_k/\beta_k + \mathbf{d}_k)\right) \\ &\quad - \sum_k a_{mk}\beta_k\mathbf{d}_k + \tilde{\mathbf{z}}_m \\ &\overset{(a)}{=} \tilde{\mathbf{z}}_m + \sum_k a_{mk}(\mathbf{t}_k - Q_{\Lambda_k^s}(\mathbf{t}_k + \beta_k\mathbf{d}_k)) \\ &:= \tilde{\mathbf{z}}_m + \sum_k a_{mk}\tilde{\mathbf{t}}_k \end{aligned}$$

with $\tilde{\mathbf{t}}_k := \mathbf{t}_k - Q_{\Lambda_k^s}(\mathbf{t}_k + \beta_k\mathbf{d}_k)$ and the equivalent noise

$$\tilde{\mathbf{z}}_m := \sum_k(\alpha_m h_{mk} - a_{mk}\beta_k)\mathbf{x}_k + \alpha_m\mathbf{z}_m$$

which is independent of $\sum_k a_{mk}\tilde{\mathbf{t}}_k$ since all $\mathbf{x}_k$ are independent of $\sum_k a_{mk}\tilde{\mathbf{t}}_k$ thanks to the dithers $\mathbf{d}_k$. The step (a) follows because it holds $Q_\Lambda(\beta_X) = \beta Q_{\frac{\Lambda}{\beta}}(X)$ for any $\beta \neq 0$. Notice we have $\tilde{\mathbf{t}}_k \in \Lambda$ since $\mathbf{t}_k \in \Lambda$ and $\Lambda_k^s \subseteq \Lambda$ due to the code construction. Hence the linear combination $\sum_k a_{mk}\tilde{\mathbf{t}}_k$ along belongs to the decoding lattice Λ.

The relay uses lattice decoding to decode $\sum_k a_{mk}\tilde{\mathbf{t}}_k$ with respect to the decoding lattice Λ by quantizing $\tilde{\mathbf{y}}_m$ to its nearest neighbor in Λ. The decoding error probability is equal to the probability that the equivalent noise $\tilde{\mathbf{z}}_m$ leaves the Voronoi region surrounding the lattice point representing $\sum_k a_{mk}\tilde{\mathbf{t}}_k$. If the fine lattice Λ used for decoding is good for AWGN channel, as it is shown in [2], the probability $\Pr(\tilde{\mathbf{z}}_m \notin \mathcal{V})$ goes to zero exponentially if

$$\frac{\text{Vol }(\mathcal{V})^{2/n}}{N_m} > 2\pi e \tag{4}$$

where $N_m := \mathbb{E}\,||\tilde{\mathbf{z}}_m||^2/n = ||\alpha_m\mathbf{h} - \tilde{\mathbf{a}}_m||^2 P + \alpha_m^2$ denotes the average power per dimension of the equivalent noise. Recall that the shaping lattice Λ_k^s is good for quantization hence we have

$$\text{Vol }(\mathcal{V}_k^s) = \left(\frac{\beta_k^2 P}{G(\Lambda_k^s)}\right)^{n/2} \tag{5}$$

with $G(\Lambda_k^s)2\pi e < (1+\delta)$ for any $\delta > 0$ if n is large enough [2]. Together with the message rate expression in (2) (here $\Lambda_{m(k)} = \Lambda$ for all k) we can see that lattice decoding is successful if $\beta_k^2 P 2^{-2R_k}/G(\Lambda_k^s) > 2\pi e N_m$ for every k or equivalently

$$R_k < \frac{1}{2}\log\left(\frac{P}{N_m}\right) + \frac{1}{2}\log\beta_k^2 - \frac{1}{2}\log(1+\delta)$$

By choosing δ arbitrarily small and optimizing over α_m we conclude that under the rate constraints in (3) the lattice decoding of $\sum_k a_k\tilde{\mathbf{t}}_k$ will be successful. Finally, since there is a one-to-one mapping between $\tilde{\mathbf{t}}_k$ and $\mathbf{t}_k$ when the dithers $\mathbf{d}_k$ are known, we can also recover $\sum_k a_k\mathbf{t}_k$. It is easy to see from the expression of the computation rate tuple in (3), that multiplying all β_k with a same constant will not change the result. ■

We see the main difference to the regular compute-and-forward scheme is that here the transmitted signal $\mathbf{x}_k$ contains the scaled version, $\mathbf{t}_k/\beta_k$, of the codeword while the receivers still perform the lattice decoding w. r. t. the lattice Λ in which $\mathbf{t}_k$ lies. We should choose β_k appropriately according to the function $\mathbf{a}$ and the channel $\mathbf{h}$ to obtain the best rate region.

Now we extend the result to allow all relays to be able to decode their desired linear functions.

Theorem 2: For any given set of positive numbers $\beta_1, \dots, \beta_K$, there exist lattice codes $\mathcal{C}_1, \dots, \mathcal{C}_K$ such that the achievable computation rate tuple $(R_1, \dots, R_K)$ with respect to the set of functions $f_m = \sum_k a_{mk}\mathbf{t}_k$, $m \in [1:M]$, where f_m is desired by relay m with $a_{mk} \in \mathbb{Z}$, is given by

$$R_k < \min_{m\in[1:M]} r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K})$$

where $r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K})$ is defined in (3).

Proof: Relay m decodes the function f_m with its decoding lattice Λ_m. The nested structure of fine lattices Λ_m ensures that the sum of codewords seen at relay m lies in the decoding lattice Λ_m. As in Theorem 1, the lattice decoding is successful if the volume-to-noise ratio of the decoding lattice satisfies equation (4). Hence each relay imposes a constraint on the individual message rate, i.e., for all k, we have $R_k \leq r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K})$ for all m. If all relays want to decode successfully, each transmitter has to construct its codebook such that it meets the above constraints at all relays. Therefore when the codebook is constructed as in (1), the fine lattice $\Lambda_{m(k)}$ for $\mathcal{C}_k$ should be such that the message rate R_k does not exceed $\min_{m\in[1:M]} r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K})$, i.e., $m(k) = \text{argmin}_{m\in[1:M]} r_k(\mathbf{h}_m, \mathbf{a}_m, \beta_{1:K})$. The noise variance N_m at each relay determines the order of the lattice chain involving Λ_m: larger N_m corresponds to a coarser lattice. ■

Remark 1: The original scheme in [1] is a special case of this modified scheme with $\beta_k = 1$ for all k. In [1], all message rates are forced to be the same, called *computation rate* at this relay. The modified scheme allows for different message rates among users and leads to the more general definition *computation rate tuple*. We shall see that this asymmetry on message rates can be beneficial in various scenarios.

Remark 2: The modified scheme extends naturally to the case when transmitters have different power constraints, and in general achieves larger computation rate region.

V. Examples

Example 1: The multiple access channel (MAC).

We consider a 2-user Gaussian MAC where the receiver wants to decode a linear function of the two messages. Figure 1 shows the achievable rate regions.

Example 2: Transmitters with different powers.

We consider the Gaussian two-way relay channel shown in Figure 2, which is studied in [3], [4]. Two encoders have different power constraints P_1 and P_2 and the channel gain from both transmitters is 1. The relay has power constraint P_R. All noises are Gaussian with unit variance.

Already shown in [3], [4], it can be beneficial for the relay to decode a linear combination of the two messages rather than decoding the two messages individually. They give the following achievable rate for this network

$$R_1 \leq \min\left\{\frac{1}{2}\log^+\left(\frac{P_1}{P_1+P_2}+P_1\right), \frac{1}{2}\log(1+P_R)\right\} \tag{6a}$$

$$R_2 \leq \min\left\{\frac{1}{2}\log^+\left(\frac{P_2}{P_1+P_2}+P_2\right), \frac{1}{2}\log(1+P_R)\right\} \tag{6b}$$

where the relay decodes the function $\mathbf{t}_1 + \mathbf{t}_2$ and broadcasts it to two decoders. With the modified compute-and-forward scheme we also ask the relay to decode a linear combination of the form $\sum_{k=1}^{2} a_k\mathbf{t}_k$ where $a_1, a_2 \neq 0$, with which each decoder can solve for the desired message. We can show the following achievable rate region:

$$R_1 \leq \min\left\{\frac{1}{2}\log^+\left(\frac{P_1\beta_1^2}{\tilde{N}(\beta_{1:2})}\right), \frac{1}{2}\log(1+P_R)\right\}$$

$$R_2 \leq \min\left\{\frac{1}{2}\log^+\left(\frac{P_2\beta_2^2}{\tilde{N}(\beta_{1:2})}\right), \frac{1}{2}\log(1+P_R)\right\}$$

where

$$\tilde{N}(\beta_{1:2}) := \frac{P_1P_2(a_1\beta_1 - a_2\beta_2)^2 + (a_1\beta_1)^2P_1 + (a_2\beta_2)^2P_2}{P_1+P_2+1}$$

for any positive β_1, β_2. Figure 3 shows the achievable rate region.

Example 3: The MIMO integer-forcing linear receiver.

We now apply the same idea to the MIMO system with an integer-forcing linear receiver [5]. We consider a point-to-point MIMO system with channel matrix $\mathbf{H} \in \mathbb{R}^{M\times K}$ which is full rank. It is shown in [5] that the following rate is achievable using integer-forcing receiver

$$R_{IF} \leq \min_{m\in[1:k]} K(-\frac{1}{2}\log \mathbf{a}_m^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{a}_m)$$

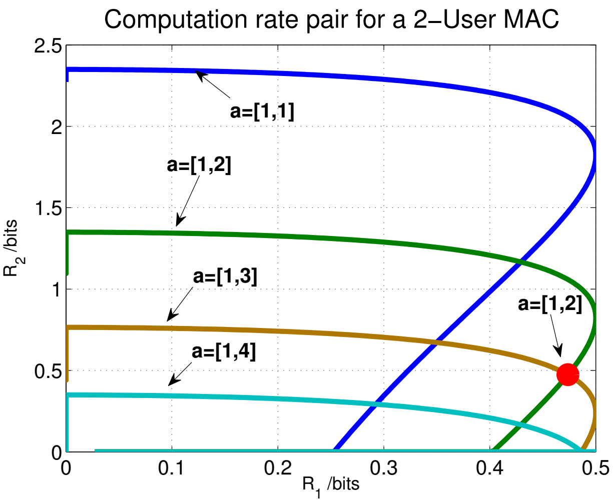


Fig. 1. We consider a 2-user Gaussian MAC with channel coefficients $\mathbf{h} = [1, 5]$ and power $P = 1$ where the receiver decodes one linear function. Here we show four achievable computational rate pair regions (R_1, R_2) of four different linear functions marked in the plot. For each function, by adjusting parameters β_1 and β_2 we can achieve different points on the curve. The red dot indicates the equal rate pair achieved with the best coefficients ($\mathbf{a} = [1, 2]$ in this case) using the regular compute-and-forward, given in [1, Thm. 4].

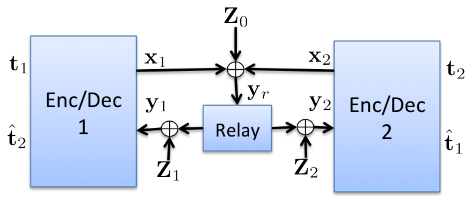


Fig. 2. A Gaussian two-way relay channel.

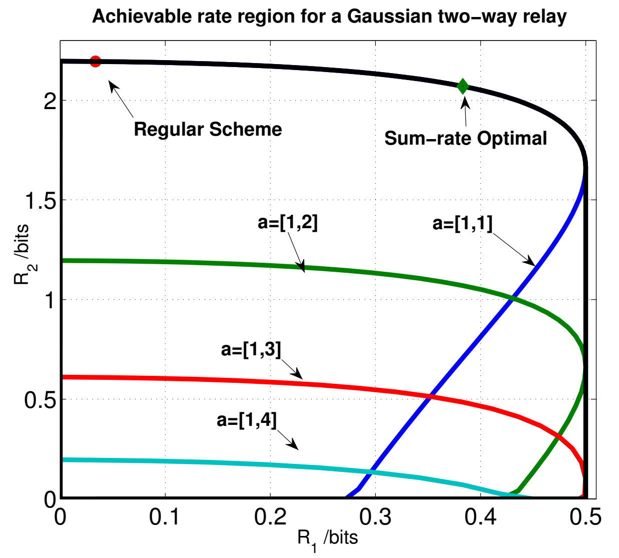


Fig. 3. Achievable rate region for the Gaussian two-way relay in Figure 2 with unequal power constraints $P_1 = 1$, $P_2 = 20$ and equal channel gain $\mathbf{h} = [1, 1]$. The relay has power $P_R = 20$. Color curves show different achievable rate region when the relay decodes different linear functions as marked in the plot. The red dot denotes the achievable rate pair given in (6) when relay decodes $\mathbf{t}_1 + \mathbf{t}_2$ using regular compute-and-forward (other function will give worse rate pair). Notice this point is not sum-rate optimal. The achievable rate region given by the black convex hull is strictly larger than the regular scheme since the CSI can be used at the transmitters.

for any full rank integer matrix $\mathbf{A} \in \mathbb{Z}^{K\times K}$ with its m-th row as $\mathbf{a}_m$ and $\mathbf{V} \in \mathbb{R}^{K\times K}$ is composed of the eigenvectors of $\mathbf{H}^T\mathbf{H}$. The matrix $\mathbf{D} \in \mathbb{R}^{K\times K}$ is diagonal with element $\mathbf{D}_{k,k} = 1/(P\lambda_k^2 + 1)$ and λ_k is the k-th singular value of $\mathbf{H}$.

Applying the modified compute-and-forward to the integer-forcing receiver gives the following result. We note that a similar idea also appears in [6] where a pre-coding matrix is used at the encoder.

Theorem 3: For a $K \times M$ real MIMO system with full rank channel matrix $\mathbf{H} \in \mathbb{R}^{M\times K}$, the following rate is achievable using an integer-forcing linear receiver for any $\beta_1, \ldots, \beta_K$

$$R_{mIF} \leq \sum_{k=1}^{K} \min_{m\in[1:K]} \left(-\frac{1}{2} \log \frac{\tilde{\mathbf{a}}_m^T \mathbf{V}\mathbf{D}\mathbf{V}^T \tilde{\mathbf{a}}_m}{\beta_k^2} \right) \qquad (7)$$

for any full rank $\mathbf{A} \in \mathbb{Z}^{K\times K}$ with its m-th row being $\mathbf{a}_m$. We have $\tilde{\mathbf{a}}_m := [\beta_1 a_{m1}, ..., \beta_K a_{mK}]$ for $m = 1, \ldots, K$ and $\mathbf{V}, \mathbf{D}$ defined as above.

In Figure 4 we compare the achievable rates of two schemes.

We give another example where the modified scheme performs arbitrarily better than the regular scheme. Consider the 2×2 MIMO channel with channel matrix $\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix}$ where $0 < \epsilon < 1$. It has been shown in [5, Section V, C] that the achievable rate of integer forcing is upper bounded as $R_{IF} \leq \log(\epsilon^2 P)$ which is of order $O(1)$ if $\epsilon \sim \frac{1}{\sqrt{P}}$ while the joint ML decoding can achieve a rate at least $\frac{1}{2}\log(1 + 2P)$. With the modified scheme we can show the following result.

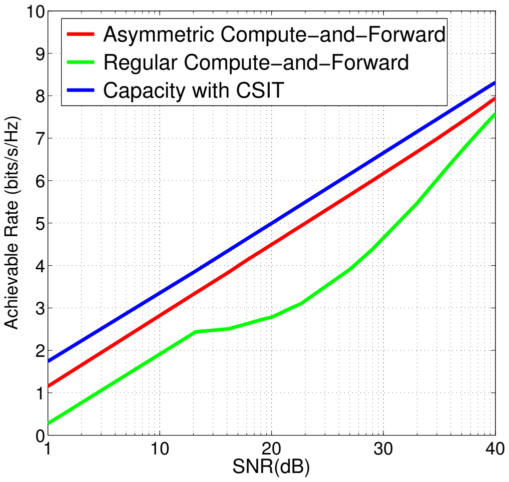


Fig. 4. Achievable rates for a 2×2 MIMO system $\mathbf{H} = [0.7, 1.3; 0.8, 1.5]$. At SNR $=$ 40dB, the best coefficients for regular scheme are $\mathbf{a}_1 = [1, 2]$ and $\mathbf{a}_2 = [7, 13]$, while for the modified scheme we have the best parameters as $\beta_1 = 1, \beta_2 = 4.887, \mathbf{a}_1 = [8, 3]$ and $\mathbf{a}_2 = [13, 5]$.

Lemma 1: For the channel $\mathbf{H}$ above, R_{mIF} in (7) scales as $\log P$ for any $\epsilon > 0$.

To see this, we can show (assuming w. l. o. g. $\beta_1 = 1$)

$$\begin{aligned} R_{mIF} \geq & \min_{m=1,2} \frac{1}{2} \log^+ \left(\frac{P}{a_{m1}^2 + (a_{m2}\beta_2 - a_{m1})^2 \frac{1}{\epsilon^2}} \right) \\ & + \min_{m=1,2} \frac{1}{2} \log^+ \left(\frac{\beta_2^2 P}{a_{m1}^2 + (a_{m2}\beta_2 - a_{m1})^2 \frac{1}{\epsilon^2}} \right) \end{aligned}$$

Based on the standard results on simultaneous Diophantine approximation [7], for any given a_{m2} and $Q > 0$ there exists $\beta_2 < Q$ and a_{m1} such that $|a_{m2}\beta_2 - a_{m1}| < Q^{-1/2}$ for $m = 1, 2$. Hence the we have the achievable rate

$$\min_{m=1,2} \frac{1}{2} \log^+ \left(\frac{P}{a_{m1}^2 + Q^{-1}\frac{1}{\epsilon^2}} \right) + \min_{m=1,2} \frac{1}{2} \log^+ \left(\frac{\beta_2^2 P}{a_{m1}^2 + Q^{-1}\frac{1}{\epsilon^2}} \right)$$

If we choose $Q \sim \epsilon^{-2}$, and notice that we also have $\beta_2, a_{m1} \sim Q$, then the second term above scales as $\frac{1}{2}\log P$ for P large. Consequently R_{mIF} also scales as $\frac{1}{2}\log P$ for any ϵ, hence can be arbitrarily better than the regular scheme.

References

[1] B. Nazer and M. Gastpar, “Compute-and-forward: Harnessing interference through structured codes,” *IEEE Trans. Inf. Theory*, vol. 57, 2011.

[2] U. Erez and R. Zamir, “Achieving 1/2 log (1+ SNR) on the AWGN channel with lattice encoding and decoding,” *IEEE Trans. Inf. Theory*, vol. 50, pp. 2293–2314, 2004.

[3] W. Nam, S.-Y. Chung, and Y. H. Lee, “Capacity of the gaussian two-way relay channel to within 1/2 bit,” *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5488–5494, 2010.

[4] M. Wilson, K. Narayanan, H. Pfister, and A. Sprintson, “Joint physical layer coding and network coding for bidirectional relaying,” *IEEE Trans. Inf. Theory*, vol. 56, no. 11, 2010.

[5] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, “Integer-forcing linear receivers,” arXiv e-print, Mar. 2010.

[6] O. Ordentlich and U. Erez, “Precoded integer-forcing universally achieves the MIMO capacity to within a constant gap,” arXiv e-print, Jan. 2013.

[7] J. W. S. Cassels, *An introduction to Diophantine approximation.* University Press Cambridge, 1957.

Topological Interference Management with Alternating Connectivity: The Wyner-Type Three User Interference Channel

Soheyl Gherekhloo, Anas Chaaban, and Aydin Sezgin
Chair of Communication Systems, RUB, Germany
Email: {soheyl.gherekhloo, anas.chaaban, aydin.sezgin}@rub.de

***Abstract*—Interference management in a three-user interference channel with alternating connectivity with only topological knowledge at the transmitters is considered. The network has a Wyner-type channel flavor, i.e., for each connectivity state the receivers observe at most one interference signal in addition to their desired signal. Degrees of freedom (DoF) upper bounds and lower bounds are derived. The lower bounds are obtained from a scheme based on joint encoding across the alternating states. Given a uniform distribution among the connectivity states, it is shown that the channel has $2+1/9$ DoF. This provides an increase in the DoF as compared to encoding over each state separately, which achieves 2 DoF only.**

I. Introduction

The smart management of interference beyond the classical approaches of avoidance and suppression is nowadays the focus of research on wireless networks. The means to apply smart management depend certainly (among other things) on the information available at the transmitting nodes, such as channel states. Often it is assumed that comprehensive channel state information is available at the transmitters (CSIT). However, providing comprehensive (or perfect) CSIT is a challenging issue in wireless networks, especially for networks with high mobility and size. It is thus of interest to study networks based on the assumption of limited or imperfect CSIT.

The case of completely stale CSIT (using the so-called retrospective interference alignment (IA)) was considered in [1] for the broadcast channel with two antennas at the base station and single-antennas at the users. It was shown that a degrees of freedom (DoF) of $4/3$ are achievable. Note that this is less than the DoF of 2 in the perfect CSIT case, however, more than the DoF of 1 in the case of completely absent CSIT. The approach was generalized to other networks in [2]. Naturally, it might occur that a mixture of CSIT quality is available at the transmitters. This issue was addressed in [3] and [4] in which the DoF is studied under the assumption of delayed as well as imperfect current CSIT. As most wireless networks are rather heterogeneous in terms of node mobility and capability, the CSI quality at the transmitters is not the same for all users. This was considered in [5], in which users have either perfect, delayed, or no CSIT at all. Similarly, the capacity region of the two-user binary fading channel was characterized in [6] for different models of availability of CSIT.

A paradigm shift towards interference management with minimal CSIT has been pursued in [7]. The main assumption of [7] is restricting the CSI feedback to 1 bit only; which provides information about presence or absence of a link. A link is assumed to be absent if its corresponding interference to noise ratio (INR) is lower than 1. Clearly, by this assumption the CSIT cannot exceed the topology of the network. Therefore, this problem is called “topological interference management". It is shown in [7] that the “topological" interference management problem for the linear wired and wireless network reduces to a single problem. In other words, solving one of these problems leads to the solution for the other one, in such a way that the DoF of a linear wireless network leads to the capacity of the corresponding linear wired channel, or vice versa. For more details, the interested reader is referred to [7].

Note that in [7] the channels are assumed to be time-invariant, which leads to a fixed connectivity within the network. In [8], the alternating connectivity was considered for the two-user interference channel (IC). It was shown that a DoF of $4/3$ can only be achieved by jointly encoding across alternating topologies. A natural question which arises is whether this gain is preserved in larger networks.

In this work, we characterize the DoF of a three user interference channel in which each receiving node is either free of interference or is interfered solely by one transmitter. The main motivation for considering such a network is to investigate whether the gain in [8] is preserved in larger networks given that interference is still caused by at most one user as in the two user IC in [8]. The analysis is focused on the corresponding wired network with equiprobable topologies, for which the capacity is characterized. This capacity characterization of the wired network leads then (as mentioned before) to the DoF characterization of the wireless network.

II. Motivation

Consider three adjacent cells in a wireless network. In each cell, a base station wants to send a message to one desired receiver. Suppose that a signal is received under the noise level if the distance between the transmitter (Tx) and the receiver (Rx) is less than the radius of the cell. Therefore, all receivers receive their desired signal over the noise level. However, there are some cases in which the receivers observe one interference signal over the noise level in addition to their desired signal. This can be seen in Fig. 1 which shows the circular coverage area of three adjacent cells. The area which is allocated to a base station is shown as a hexagon inside a circle. Therefore, there are some areas close to the edges of each cell in which

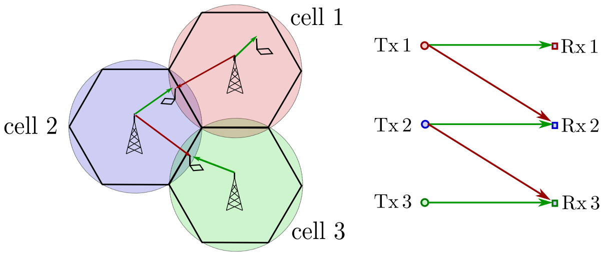


Fig. 1. Each base station serves the users located in its cell. However, its signal can be received over the noise level in some areas of the adjacent cells. For example, the users in cell 2 and 3 receive an interference signal from the base stations in cell 1 and 2, respectively. On the right, the topology of this network is shown. Note that each receiver experiences at most one interference.

the receiver experiences one interference signal in addition to its own desired signal. As an example, Rx 2 and Rx 3 in Fig. 1 observe an interference from undesired base stations Tx 1 and Tx 2, respectively. Since an interferer which is weaker than noise does not have an impact on the DoF of the network; the corresponding link to that interferer is assumed to be absent in the topology of the network (see the topology of the wireless network in Fig. 1).

III. System Model

As it is shown in [7], the capacity of a noiseless wired network normalized by the capacity of a single link gives us the degrees of freedom for the corresponding wireless network with additive white Gaussian noise. For simplicity, and in order to avoid the unnecessary treatment of noise in the wireless network which does not have an impact on the DoF of the network, we study the wired noiseless network. Consider three Tx's which want to communicate with their desired Rx's. Tx i, $i \in \{1,2,3\}$ wants to send a message W_i to Rx i. It encodes this message into a length-n sequence $\boldsymbol{X}_i = (X_i(1), \ldots, X_i(n))$ and sends this sequence. The received symbol at Rx j in kth channel use is given by

$$Y_j(k) = \sum_{i=1}^{3} h_{ji}(k) X_i(k), \quad \forall j \in \{1,2,3\} \tag{1}$$

where $X_i(k)$ and $h_{ji}(k)$ denote the transmitted symbol by Tx i and the channel coefficient corresponding to the link between Tx i and Rx j. All symbols are chosen from a Galois Field $\mathbb{GF}$. Moreover, the linear operations are performed over this $\mathbb{GF}$. The capacity of each channel is $\log|\mathbb{GF}|$, where $|\mathbb{GF}|$ represents the cardinality of $\mathbb{GF}$. Therefore, only one symbol can be transmitted over a link per channel use.

In our model, CSIT is restricted only to the topology of the network. Therefore, the only information available to the transmitters is about the presence or absence of links but not about the channel coefficients. However, both the local channel coefficients and the topology of the network are known at the receivers.

Since the channel coefficients change, the topology of the network varies during the transmission. Following the motivation in Fig. 1, the desired channels always exist and each receiver is disturbed by at most one interferer. Therefore, the network has a total of 27 topologies as shown in Fig. 2.

It is worth to note that the receivers have an infinite memory and they start the decoding after receiving a complete sequence $\boldsymbol{Y}_j$. Therefore, the order of the occurrence of the states is not important. Let $\mathcal{A}$ be a set of states shown in Fig. 2 and $\boldsymbol{X}_{i,\mathcal{A}}$ be the sequence of transmitted symbols by Tx i in all states in $\mathcal{A}$. Assuming a length-n sequence $\boldsymbol{X}$, in which n is sufficiently large, the length of $\boldsymbol{X}_{i,\mathcal{A}}$ is $n\lambda_{\mathcal{A}}$, where $\lambda_{\mathcal{A}}$ denotes the sum of the probabilities of the states in $\mathcal{A}$.

The goal of this work is to study the DoF gain obtained by jointly encoding across the alternating topologies, when all states occur with the same probability.

IV. Main Result

The following theorem provides the main result of this work.

Theorem 1. *The three user interference channel with alternating connectivity and equiprobable states with at most one interferer per receiver has DoF*=$2 + 1/9$.

Proof: We establish Theorem 1 by showing that the sum capacity of the corresponding wired network is $(2 + \frac{1}{9}) \log|\mathbb{GF}|$. In order to do this, we need to find an optimal achievability scheme. The optimality of the scheme is shown by comparing it with a tight upper bound of the sum capacity. We start by proposing an achievability scheme leading to a DoF lower bound denoted $\underline{\text{DoF}}$.

Achievability:

The achievability is based on the joint encoding over the sates [8]. To this end, consider states B_1, C_1, D_1, and H_1 in Fig. 2. It can be seen that all interference links in states B_1, C_1, and D_1 are present in state H_1. Therefore, we can utilize state H_1 to resolve the interferences in these states. As it is shown in Fig. 3, the symbols b_1, c_2, and d_3 cannot be decoded at the desired receivers in states B_1, C_1, and D_1. However, by using the state H_1, the transmitters provide the symbols which cause interference in states B_1, C_1, and D_1 to the receivers. Therefore, in total 9 symbols are decoded correctly at the desired receivers by combining these four states. Similarly, the same joint encoding scheme can be used for B_2, C_2, D_2, and H_2 due to symmetry. The remaining states are encoded individually. In all these states except in state A, we achieve $\underline{\text{DoF}}$=2 by choosing two active transmitters. For instance, in state I_1, $\underline{\text{DoF}}$=2 is achievable when Tx 2 and Tx 3 send while Tx 1 is silent. Overall, the following $\underline{\text{DoF}}$ is achievable

$$\underline{\text{DoF}} = \begin{cases} 9/4 & \text{for } B_1 \cup C_1 \cup D_1 \cup H_1 \\ 9/4 & \text{for } B_2 \cup C_2 \cup D_2 \cup H_2 \\ 3 & \text{for } A \\ 2 & \text{in all remaining 18 states} \end{cases}$$

Since, all states occur with equal probability, we can transmit 57 symbols reliably in 27 channel uses in average. Since every symbol is chosen from $\mathbb{GF}$ with the entropy $\log|\mathbb{GF}|$, the achievable sum rate is

$$R_\Sigma \leq \left(2 + \frac{1}{9}\right) \log|\mathbb{GF}|. \tag{2}$$

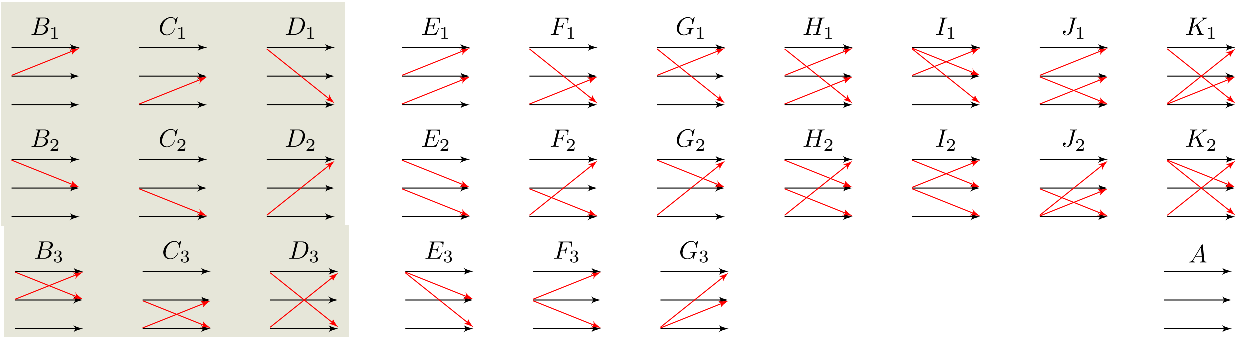


Fig. 2. All possible states for the three users interference channel, when each receiver observes at most one interferer. The desired links are always present.

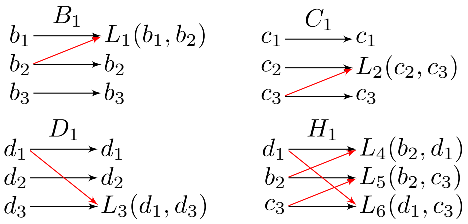


Fig. 3. By combining these four states, we can recover 9 symbols. However, by considering them separately, we cannot exceed $15/2$ symbols.

Upper bound:

We establish the upper bound as follows

$$\begin{aligned} nR_\Sigma &= \sum_{i=1}^{3} H(W_i) \\ &= \sum_{i=1}^{3} H(W_i) + H(W_i|\boldsymbol{Y}_i) - H(W_i|\boldsymbol{Y}_i) \\ &\overset{(a)}{\leq} \sum_{i=1}^{3} I(W_i;\boldsymbol{Y}_i) + 3n\epsilon_n, \end{aligned} \tag{3}$$

where (a) follows from Fano's inequality and $\epsilon_n \to 0$ when $n \to \infty$. By multiplying the inequality in (3) by 2, every mutual information appears twice which corresponds to creating (virtually) three additional receivers. In the next step, we give side information to the actual receivers. The side information equals to the undesired messages at those receivers. Therefore, we write

$$\begin{aligned} 2nR_\Sigma \leq & I(W_1;\boldsymbol{Y}_1,W_2,W_3) + I(W_2;\boldsymbol{Y}_2,W_1,W_3) \\ &+ I(W_3;\boldsymbol{Y}_3,W_1,W_2) + \sum_{i=1}^{3} I(W_i;\boldsymbol{Y}_i) + 6n\epsilon_n. \end{aligned} \tag{4}$$

By using the chain rule and since the messages of three transmitters are independent from each other, we write

$$\begin{aligned} 2nR_\Sigma \leq & I(W_1;\boldsymbol{Y}_1|W_2,W_3) + I(W_2;\boldsymbol{Y}_2|W_1,W_3) \\ &+ I(W_3;\boldsymbol{Y}_3|W_1,W_2) + \sum_{i=1}^{3} I(W_i;\boldsymbol{Y}_i) + 6n\epsilon_n. \end{aligned} \tag{5}$$

By expressing the mutual information as entropy terms, (5) is restated as

$$\begin{aligned} 2nR_\Sigma \leq & H(\boldsymbol{Y}_1|W_2,W_3) - H(\boldsymbol{Y}_1|W_2,W_3,W_1) \\ &+ H(\boldsymbol{Y}_2|W_1,W_3) - H(\boldsymbol{Y}_2|W_1,W_3,W_2) \\ &+ H(\boldsymbol{Y}_3|W_1,W_2) - H(\boldsymbol{Y}_3|W_1,W_2,W_3) \\ &+ \sum_{i=1}^{3} I(W_i;\boldsymbol{Y}_i) + 6n\epsilon_n. \end{aligned} \tag{6}$$

Note that knowing all messages, $\boldsymbol{Y}_i$ can be reconstructed. Therefore, $H(\boldsymbol{Y}_i|W_1,W_2,W_3) = 0$. The first term in (6) reduces to

$$H(\boldsymbol{Y}_1|W_2,W_3) = H(\boldsymbol{X}_1),$$

as $\boldsymbol{X}_1$ is independent of W_2 and W_3 and the fact that scaling a discrete random variable by a constant does not influence entropy [9]. Similar treatment applies to $H(\boldsymbol{Y}_2|W_1,W_3)$ and $H(\boldsymbol{Y}_3|W_1,W_2)$ in (6). Next, we rewrite (6) as shown in (7) on the top of next page. The parameters Δ_i, Γ_i, and Θ_i, $i \in \{1,2,3\}$ are defined as follows

$$\begin{aligned} \Delta_1 &= \{D_1,F_1,G_1,H_1,I_1,K_1,K_2,D_3\} \\ \Gamma_1 &= \{B_2,E_2,G_2,H_2,I_2,B_3\} \\ \Theta_1 &= \overline{\{E_3\}\cup\Delta_1\cup\Gamma_1} \\ \Delta_2 &= \{B_1,E_1,G_1,H_1,I_1,J_1,I_2,B_3\} \\ \Gamma_2 &= \{C_2,E_2,F_2,H_2,J_2,C_3\} \\ \Theta_2 &= \overline{\{F_3\}\cup\Delta_2\cup\Gamma_2} \\ \Delta_3 &= \{C_1,E_1,F_1,H_1,J_1,K_1,J_2,C_3\} \\ \Gamma_3 &= \{D_2,F_2,G_2,H_2,K_2,D_3\} \\ \Theta_3 &= \overline{\{G_3\}\cup\Delta_3\cup\Gamma_3}. \end{aligned}$$

The notation $\overline{\mathcal{A}}$ denotes the complement set of $\mathcal{A}$. By using the chain rule, together with the facts that conditioning does not increase entropy, and that the messages of the users are independent of each other, the individual terms in (7) can be rewritten as in (8)-(16) on the top of next page. We can see that by substituting (8)-(16) into (7) many terms will cancel out and we can rewrite (7) as

$$\begin{aligned} 2nR_\Sigma \leq & \sum_{i=1}^{3} H(\boldsymbol{X}_{i,\Theta_i}) + H(\boldsymbol{Y}_{1,\overline{E_3}}|\boldsymbol{X}_{1,E_3}) \\ &+ H(\boldsymbol{Y}_{2,\overline{F_3}}|\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{Y}_{3,\overline{G_3}}|\boldsymbol{X}_{3,G_3}) + 6n\epsilon_n. \end{aligned} \tag{17}$$

$$\begin{aligned}2nR_\Sigma \leq & H(\boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Delta_1}, \boldsymbol{X}_{1,\Gamma_1}, \boldsymbol{X}_{1,\Theta_1}) + H(\boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Delta_2}, \boldsymbol{X}_{2,\Gamma_2}, \boldsymbol{X}_{2,\Theta_2}) + H(\boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Delta_3}, \boldsymbol{X}_{3,\Gamma_3}, \boldsymbol{X}_{3,\Theta_3})\\ & + H(\boldsymbol{X}_{1,E_3}, \boldsymbol{Y}_{1,\overline{E_3}}) - H(\boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Delta_2}, \boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Gamma_3}, \boldsymbol{X}_{3,K_1}, \boldsymbol{X}_{3,J_2})\\ & + H(\boldsymbol{X}_{2,F_3}, \boldsymbol{Y}_{2,\overline{F_3}}) - H(\boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Delta_3}, \boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Gamma_1}, \boldsymbol{X}_{1,I_1}, \boldsymbol{X}_{1,K_2})\\ & + H(\boldsymbol{X}_{3,G_3}, \boldsymbol{Y}_{3,\overline{G_3}}) - H(\boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Delta_1}, \boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Gamma_2}, \boldsymbol{X}_{2,J_1}, \boldsymbol{X}_{2,I_2}) + 6n\epsilon_n \end{aligned} \tag{7}$$

$$H(\boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Delta_1}, \boldsymbol{X}_{1,\Gamma_1}, \boldsymbol{X}_{1,\Theta_1}) \leq H(\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{1,\Delta_1}|\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{1,\Gamma_1}|\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{1,\Theta_1}) \tag{8}$$

$$H(\boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Delta_2}, \boldsymbol{X}_{2,\Gamma_2}, \boldsymbol{X}_{2,\Theta_2}) \leq H(\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{2,\Delta_2}|\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{2,\Gamma_2}|\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{2,\Theta_2}) \tag{9}$$

$$H(\boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Delta_3}, \boldsymbol{X}_{3,\Gamma_3}, \boldsymbol{X}_{3,\Theta_3}) \leq H(\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{3,\Delta_3}|\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{3,\Gamma_3}|\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{3,\Theta_3}) \tag{10}$$

$$H(\boldsymbol{X}_{1,E_3}, \boldsymbol{Y}_{1,\overline{E_3}}) \leq H(\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{Y}_{1,\overline{E_3}}|\boldsymbol{X}_{1,E_3}) \tag{11}$$

$$H(\boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Delta_2}, \boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Gamma_3}, \boldsymbol{X}_{3,K_1}, \boldsymbol{X}_{3,J_2}) \geq H(\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{2,\Delta_2}|\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{3,\Gamma_3}|\boldsymbol{X}_{3,G_3}) \tag{12}$$

$$H(\boldsymbol{X}_{2,F_3}, \boldsymbol{Y}_{2,\overline{F_3}}) \leq H(\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{Y}_{2,\overline{F_3}}|\boldsymbol{X}_{2,F_3}) \tag{13}$$

$$H(\boldsymbol{X}_{3,G_3}, \boldsymbol{X}_{3,\Delta_3}, \boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Gamma_1}, \boldsymbol{X}_{1,I_1}, \boldsymbol{X}_{1,K_2}) \geq H(\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{3,\Delta_3}|\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{1,\Gamma_1}|\boldsymbol{X}_{1,E_3}) \tag{14}$$

$$H(\boldsymbol{X}_{3,G_3}, \boldsymbol{Y}_{3,\overline{G_3}}) \leq H(\boldsymbol{X}_{3,G_3}) + H(\boldsymbol{Y}_{3,\overline{G_3}}|\boldsymbol{X}_{3,G_3}) \tag{15}$$

$$H(\boldsymbol{X}_{1,E_3}, \boldsymbol{X}_{1,\Delta_1}, \boldsymbol{X}_{2,F_3}, \boldsymbol{X}_{2,\Gamma_2}, \boldsymbol{X}_{2,J_1}, \boldsymbol{X}_{2,I_2}) \geq H(\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{1,\Delta_1}|\boldsymbol{X}_{1,E_3}) + H(\boldsymbol{X}_{2,F_3}) + H(\boldsymbol{X}_{2,\Gamma_2}|\boldsymbol{X}_{2,F_3}) \tag{16}$$

The inequality (17) can be further upper bounded by

$$\begin{aligned}2nR_\Sigma \leq & \log|\mathbb{GF}|[n\lambda_{\Theta_1} + n\lambda_{\Theta_2} + n\lambda_{\Theta_3} + n(1-\lambda_{E_3})\\ & + n(1-\lambda_{F_3}) + n(1-\lambda_{G_3})] + 6n\epsilon_n, \end{aligned} \tag{18}$$

where we used the chain rule, the fact that conditioning does not increase the entropy, and that the entropy of discrete random variable in $\mathbb{GF}$ is upper bounded by $\log|\mathbb{GF}|$ [9].

Since the set Θ_i consists of 12 states, $\lambda_{\Theta_i} = \frac{12}{27}$ if all states are equiprobable. Next, we divide the inequality in (18) by $2n$, and let $n \to \infty$ to obtain

$$R_\Sigma \leq \left(2 + \frac{1}{9}\right) \log|\mathbb{GF}|. \tag{19}$$

This agrees with the lower bound in (2). Normalizing the result by $\log|\mathbb{GF}|$, we get the DoF for the wireless case which proves Theorem 1. ■

We observe from Theorem 1 that no joint processing is necessary for $\overline{\{B_1, C_1, D_1, H_1, B_2, C_2, D_2, H_2\}}$. However, for $\{B_1, C_1, D_1, H_1, B_2, C_2, D_2, H_2\}$, we need joint encoding to achieve the optimal DoF. The alternative approach would be to treat these states separately as well. This would result in a DoF=$3/2$ and DoF=2 for the states $\{H_1, H_2\}$ (as shown in [10]) and $\{B_1, C_1, D_1, B_2, C_2, D_2\}$ (as shown in [11]), respectively. Therefore, the overall DoF=2 for separate encoding while by using joint encoding across the alternating topologies $2 + \frac{1}{9}$ is the optimal achievable DoF. Therefore, the gain of jointly encoding is $\frac{1}{9}$ in a three user IC with at most one interferer. This gain is smaller than the attained gain in a two user IC which is $\frac{1}{4}$ for the equiprobable case [8].

V. Conclusion

We studied the DoF of the three users interference channel with an alternating connectivity with only topological knowledge at the transmitters. To do this, we proposed a new joint encoding across the alternating topologies. Moreover, a new genie aided upper bound is established to verify the optimality of the joint encoding scheme. The upper bound is tight for the equiprobable case. As future work, the non-equiprobable case will be addressed. However, this extension is non-trivial due to the increase in the number of possible combination of states.

References

[1] M. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," in *48th Allerton Conf. on CCC*, 2010, pp. 1188–1195.

[2] H. Maleki, S. A. Jafar, and S. Shamai, "Retrospective Interference Alignment over Interference Networks," *IEEE Journal of Selected Topics in Signal Processing*.

[3] T. Gou and S. A. Jafar, "Optimal use of current and outdated channel state information – degrees of freedom of the MISO BC with mixed CSIT," *IEEE Comm. Letters*, vol. 16, no. 7, pp. 1084–1087, July 2012.

[4] M. Kobayashi, S. Yang, D. Gesbert, and X. Yi, "On the degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," in *IEEE ISIT*, 2012, pp. 2501–2505.

[5] R. Tandon, S. Jafar, S. Shamai Shitz, and H. Poor, "On the synergistic benefits of alternating CSIT for the MISO broadcast channel," *IEEE Trans. on IT*, July 2013.

[6] A. Vahid, M. A. Maddah-Ali, and A. S. Avestimehr, "Capacity results for binary fading interference channels with delayed CSIT," Jan. 2013. [Online]. Available: http://arxiv.org/abs/1301.5309

[7] S. A. Jafar, "Topological interference management through index coding," Jan. 2013. [Online]. Available: http://arxiv.org/pdf/1301.3106.pdf

[8] H. Sun, C. Geng, and S. A. Jafar, "Topological interference management with alternating connectivity," Feb. 2013. [Online]. Available: http://arxiv.org/abs/1302.4020

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, August 1991.

[10] L. Zhou and W. Yu, "On the symmetric capacity of the K-user symmetric cyclic Gaussian interference channel," in *CISS*, Princeton, NJ, 2010.

[11] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. on IT*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

Modify-and-Forward for Securing Cooperative Relay Communications

Sang Wu Kim
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
E-mail: swkim@iastate.edu

Abstract - We proposed a new physical layer technique that can enhance the security of cooperative relay communications. The proposed approach modifies the decoded message at the relay according to the unique channel state between the relay and the destination such that the destination can utilize the modified message to its advantage while the eavesdropper cannot. We present a practical method for securely sharing the modification rule between the legitimate partners and present the secrecy outage probability in a quasi-static fading channel. It is demonstrated that the proposed scheme can provide a significant improvement over other schemes when the relay can successfully decode the source message.

I. Introduction

In recent years, there have been considerable efforts devoted to using the channel to provide security in wireless communications. It is shown in [1] that fading alone guarantees that information-theoretic security is achievable, even when the eavesdropper has a better average SNR than the legitimate receiver. A traditional approach to enhancing the secrecy rate is to introduce interference (jamming) into the channel so as to harm the eavesdropper's ability to eavesdrop while strengthening the ability for legitimate entities to communicate. This idea has appeared in the literature under the name of artificial noise [2], cooperative jamming (CJ) [3], [4], [5], [6], [7], or noise forwarding (NF) [8], [9].

In this paper we propose a new physical layer technique that can enhance the security of cooperative relay communications. Unlike traditional approaches in which no context (message) is sent by the relay, in the proposed scheme the relay decodes the source message X and forwards a *modified* message X' to the destination such that the intended destination can utilize X' to its advantage while the eavesdropper cannot. The basic idea is to exploit the unique physical channel state between the relay and the destination as the inherent shared secret in sharing $X' - X$ without exchanging any information about $X' - X$. Once the difference $X' - X$ is known at the destination, it can be canceled from the modified message X' to get the original message X, while the eavesdropper without knowing the difference[1] cannot extract X from X'. The additional information about X provided by the relay can improve the rate towards the intended destination without improving the rate towards the eavesdropper. Hereafter, the proposed scheme will be referred to as *modify-and-forward* (MF).

We present a practical method for securely sharing the difference $X' - X$ (or modification rule in general) by exploiting the unique physical channel state between the legitimate partners. We characterize the security level in a quasi-static fading environment by computing the secrecy outage probability that provides the fraction of fading realizations for which the wireless channel cannot support a target secure rate. We compare the secrecy outage probability of the proposed scheme with that of direct transmission (DT), decode-and-forward (DF), and CJ under different system setups.

II. System Model

We consider the cooperative relay communication system shown in Fig. 1 in which a source (S) communicates with a destination (D) with the help of a relay (R) in the presence of a eavesdropper (E). We assume that each node carries a single omnidirectional antenna. Channels between all pairs of nodes are modeled as independent quasi-static Rayleigh fading channels: fading coefficients remain constant during the transmission of an entire codeword but they change from one codeword to another according to a complex Gaussian distribution.

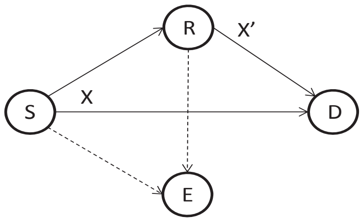


Fig. 1. Cooperative relay communication model for modify-and-forward relaying.

In the first phase, S broadcasts the message X to D and E. In the second phase, the relay decodes the message transmitted by S, modifies the decoder output to X' and broadcasts X' to D and E. We require the relay to fully decode the source message X and the source to remain silent during the second phase. We assume that X and X' are of length n and are independently chosen from a Gaussian random codebook of M codewords. We also assume that each codeword is chosen with equal probability and that $E[X] = E[X'] = 0$ and $E[||X||^2]/n = E[||X'||^2]/n = P$. Thus the total transmission power is $2P$.

[1] The eavesdropper cannot determine the physical channel state between the legitimate nodes as long as the former is more than half of the wavelength away from the latter.

The received signals at the destination that are originated from the source and relay are, respectively, given by

$$\begin{aligned} Y_{sd} &= h_{sd}X + N_{sd} && (1) \\ Y_{rd} &= h_{rd}X' + N_{rd} && (2) \end{aligned}$$

where h_{ij} is the channel gain between the node i and node j, and N_{ij} is white Gaussian noise with mean zero and variance σ_n^2. Once $X' - X$ is known at the destination, it can be removed from Y_{rd} to get

$$\begin{aligned} Y'_{rd} &= Y_{rd} - h_{rd}(X' - X) && (3) \\ &= h_{rd}X + N_{rd} && (4) \end{aligned}$$

and X can be decoded based on Y_{sd} and Y'_{rd}.

We assume that the eavesdropper knows that the message is modified by the relay. However, without knowing the difference $X' - X$, it has to discard the signal received from the relay $Y_{re} = h_{re}X' + N_{re}$ and decode X based on the signal received from the source only:

$$Y_{se} = h_{se}X + N_{se} \qquad (5)$$

where h_{se} is the channel gain between the source and the eavesdropper and N_{se} is the noise. This is because Y_{re} does not provide any information about X unless $X' - X$ is known.

The question is how to achieve the agreements on message modification secretly between the relay and the destination. Only when two nodes share the same modification rule they can achieve high secrecy rate. Our approach is based on the uniqueness and reciprocity of wireless fading channel. The reciprocity theory demonstrates that bidirectional wireless channel states should be identical between two transceivers during the channel's coherence time [10]. We use this unique channel state as the inherent shared secret between the relay and the destination for message modification and restoration. As long as the eavesdropper is more than half of the wavelength away from legitimate communicators, the channel states he observed should be independent to the channel state between the legitimate ones [11]. This means the eavesdropper can never eavesdrop the secret $X' - X$ shared between legitimate communicators. Since the legitimate communicators do not exchange any information about $X' - X$, our approach provides a strong security. The uniqueness of the wireless channel between two locations has also been utilized in authenticating legitimate users [12].

III. Secrecy Outage Probability

In this section we derive the secrecy outage probability which provides the fraction of fading realizations for which the wireless channel cannot support a target secrecy rate of R. It provides a security metric for the situation where the source and destination have no channel state information about the eavesdropper.

A. Modify-and-Forward

The maximum rate at which the relay and the destination can reliably decode the message X is given by [13]

$$\begin{aligned} C_d &= \min\left\{ \frac{1}{2}\log_2\left(1 + |h_{sr}|^2 P/\sigma_n^2\right), \right. \\ &\qquad \left. \frac{1}{2}\log_2\left(1 + (|h_{sd}|^2 + |h_{rd}|^2)P/\sigma_n^2\right) \right\} \qquad (6) \end{aligned}$$

where the factor $1/2$ accounts for the two-phase transmission. Similarly, the maximum rate at which E can reliably decode the message X is

$$C_e = \frac{1}{2}\log_2\left(1 + |h_{se}|^2 P/\sigma_n^2\right) \qquad (7)$$

because the eavesdropper cannot utilize the modified message which is sent by the relay. Then, the instantaneous secrecy capacity between S and D is [14]

$$C_s = \max(C_d - C_e, 0) \qquad (8)$$

Communication is secure if the instantaneous secrecy capacity C_s is higher than the target secrecy rate R (b/s/Hz). If $C_s < R$, then security is compromised and secrecy outage occurs. The secrecy outage probability for the proposed scheme can be shown to be

$$\begin{aligned} P_o(R) &= P(C_s < R) && (9) \\ &= 1 - \frac{1}{\gamma_{rd} - \gamma_{sd}}\left(1 + \frac{\gamma_{rd}}{\gamma_{sr}}\right) e^{-(2^{2R}-1)\left(\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{rd}}\right)} \\ &\quad \times \left[\frac{1}{\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{rd}}} - \frac{1}{\frac{2^{-2R}}{\gamma_{se}} + \frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{rd}}} \right] \\ &\quad + \frac{1}{\gamma_{rd} - \gamma_{sd}}\left(1 + \frac{\gamma_{sd}}{\gamma_{sr}}\right) e^{-(2^{2R}-1)\left(\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{sd}}\right)} \\ &\quad \times \left[\frac{1}{\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{sd}}} - \frac{1}{\frac{2^{-2R}}{\gamma_{se}} + \frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{sd}}} \right] && (10) \end{aligned}$$

where $\gamma_{sd} = E[|h_{sd}|^2]P/\sigma_n^2$, $\gamma_{rd} = E[|h_{rd}|^2]P/\sigma_n^2$, $\gamma_{se} = E[|h_{se}|^2]P/\sigma_n^2$, and $\gamma_{re} = E[|h_{re}|^2]P/\sigma_n^2$. Proof of (10) is provided in Appendix A.

B. Direct Transmission

For the direct transmission (DT), within a transmission slot, the source transmits its n encoded symbols directly to the destination using the available transmit power of $2P$. The secrecy outage probability with the DT is given by [1]

$$P_o(R) = 1 - \frac{\gamma_{sd}}{\gamma_{sd} + 2^R\gamma_{se}} \exp\left(-\frac{2^R - 1}{2\gamma_{sd}}\right) \qquad (11)$$

where the factor 2 in front of γ_{sd} accounts for the total transmit power of $2P$.

C. Decode-and-Forward

Like MF, decode-and-forward (DF) is also a two-phase scheme. The first phase is the same as in the MF scheme. In the second phase, the relay decodes the information transmitted by the source and re-encodes it using the same codeword as the source to transmit the information to D. Thus the total transmission power is $2P$. The secrecy outage probability with the DF is given by [7]

$$\begin{aligned} P_o(R) &= \frac{a(\gamma_{re}) - a(\gamma_{se})}{\gamma_{re} - \gamma_{se}} \\ &\quad + \frac{\gamma_{sr}2^{-2R}a(\gamma_{se})(h(\gamma_{se}, \gamma_{sd}) - h(\gamma_{se}, \gamma_{rd}))}{(\gamma_{re} - \gamma_{se})(\gamma_{rd} - \gamma_{sd})} \\ &\quad - \frac{\gamma_{sr}2^{-2R}a(\gamma_{re})(h(\gamma_{re}, \gamma_{sd}) - h(\gamma_{re}, \gamma_{rd}))}{(\gamma_{re} - \gamma_{se})(\gamma_{rd} - \gamma_{sd})} \qquad (12) \end{aligned}$$

where

$$h(x,y) = \frac{\gamma_{sr}}{x(1+\gamma_{sr}/y)+\gamma_{sr}2^{-2R}} \quad (13)$$

$$a(x) = \frac{x^2}{\gamma_{sr}2^{-2R}+x}\exp\left(-\frac{2^{-2R}-1}{x}\right) \quad (14)$$

D. Cooperative Jamming

Various cooperative jamming (CJ) schemes that involve the transmission of jamming signals from different nodes have been proposed [3], [4], [6]. In this paper we consider the cooperative jamming scheme where, while S transmits, the relay transmits a jamming signal that is independent of the source message with the purpose of confounding E. The jamming signal, white Gaussian noise, causes interference at both D and E. The total transmission power[2] is $2P$ as the source and relay transmits with power P. The secrecy outage probability for the CJ is given by [3]

$$\begin{aligned} P_o(R) &= 1 - \frac{2^{-\kappa}}{\gamma_{rd}\gamma_{re}}\frac{\gamma_{re}}{\left(\kappa+\frac{1}{\gamma_{rd}}-\frac{\beta}{\gamma_{re}}\right)} \\ &\quad + \frac{2^{-\kappa}}{\gamma_{rd}\gamma_{re}}\left(\kappa+\frac{1}{\gamma_{rd}}-\frac{\beta}{\gamma_{re}}\right)^{-2} \\ &\quad \times\left[\beta\left(\kappa+\frac{1}{\gamma_{rd}}-\frac{\beta}{\gamma_{re}}+1\right)\Omega\left(\frac{1+\beta}{\gamma_{re}}\right)\right. \\ &\quad + \left(\kappa+\frac{1}{\gamma_{rd}}-\frac{\beta}{\gamma_{re}}-\beta\right) \\ &\quad \left.\times\,\Omega\left(\frac{1+\beta}{\beta}\left(\kappa+\frac{1}{\gamma_{rd}}\right)\right)\right] \quad (15) \end{aligned}$$

where $\kappa = (2^{2R}-1)/\gamma_{sd}$, $\beta = 2^{2R}\gamma_{se}/\gamma_{sd}$, and $\Omega(x) = e^x E_1(x)$ where $E_1(x) = \int_x^\infty u^{-1}e^{-u}du$.

E. Numerical Results

Fig. 2 shows the secrecy outage probability, $P_o(R)$, versus the average signal-to-noise ratio (SNR) between the source and the eavesdropper, γ_{se}. As expected the secrecy outage probability increases with increasing γ_{se} because the rate at which the eavesdropper can reliably decode the message increases as the channel condition between the source and itself improves. It can also be seen that the improvement provided by MF over DF is more significant at lower γ_{se}. This is because the eavesdropper relies sorely on the channel between the source and eavesdropper in MF, while in DF the eavesdropper can rely on the channel between the relay and itself when γ_{se} is low. Similarly, in DT the eavesdropper relies sorely on the channel between the source and itself and therefore the secrecy outage probability depends heavily on γ_{se}.

Fig. 3 shows the secrecy outage probability, $P_o(R)$, versus the average SNR between the source and the relay, γ_{sr}. For DF and MF schemes, the relay has to decode the source message in order to provide any additional information to the destination. Therefore, if γ_{sr} is low, the secrecy outage probability for DF and MF is high because the relay cannot decode the source message. However, if γ_{sr} is high enough such that the relay can decode the source message, then it can provide additional information to the destination, which increases the secrecy capacity. At sufficiently high γ_{sr}, the secrecy outage probability for DF and MF remains constant because all other channel gains are assumed to be constant.

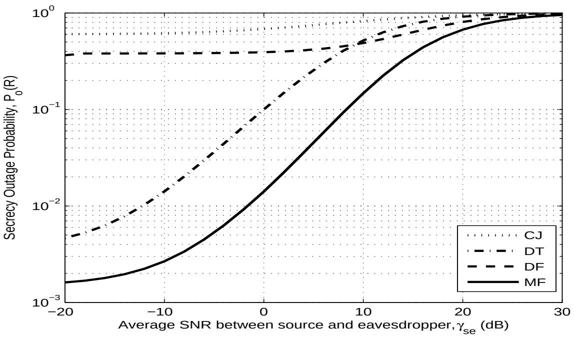


Fig. 2. Secrecy outage probability, $P_o(R)$, versus average SNR between source and eavesdropper, γ_{se} (dB); $R = 0.1$b/s/Hz, $\gamma_{sd} = 10$dB, $\gamma_{sr} = 20$dB, $\gamma_{rd} = 20$dB, $\gamma_{re} = 15$dB.

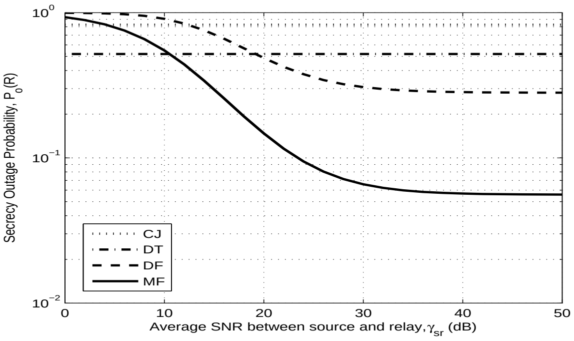


Fig. 3. Secrecy outage probability, $P_o(R)$, versus average SNR between source and relay, γ_{sr} (dB); $R = 0.1$b/s/Hz, $\gamma_{sd} = 10$dB, $\gamma_{se} = 10$dB, $\gamma_{rd} = 20$dB, $\gamma_{re} = 15$dB.

Fig. 4 shows the secrecy outage probability, $P_o(R)$, versus the target secrecy rate R. It can be seen that the improvement that MF provides over the traditional approaches is more significant when the target secrecy rate R is smaller. However, if R is above a threshold, DT provides the smallest secrecy outage probability, although the secrecy outage probability in that rate region is unacceptably high. It can also be seen from Figs. 2-4 that MF can always provide a lower secrecy outage probability than DF under any channel conditions and rates.

IV. Conclusion

We proposed a new physical layer technique that can enhance the security of cooperative relay communications. The proposed approach modifies the decoded message at the relay according to the unique channel state between the relay and the destination such that the destination can utilize it to its advantage while the eavesdropper cannot. We derived the secrecy outage probability in quasi-static fading channel, and compared with direct transmission, decode-and-forward, cooperative jamming under different system setups. Numerical

[2] The total transmission power of CJ schemes in [4], [6] is $3P$ because each of three nodes (source, relay, and destination) transmits with power P.

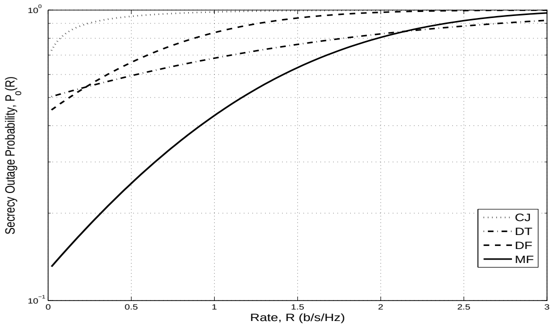


Fig. 4. Secrecy outage probability $P_o(R)$ versus rate R (b/s/Hz); $\gamma_{sd} = 10$dB, $\gamma_{sr} = 20$dB, $\gamma_{rd} = 20$dB, $\gamma_{se} = 10$dB, $\gamma_{re} = 15$dB.

results reveal that each scheme provides an advantage over the others depending on the channel gains and secrecy rates, although the proposed scheme can always provide a lower secrecy outage probability than decode-and-forward scheme. The proposed approach can provide a significant improvement over other schemes when the relay can successfully decode the source message.

Appendix A

In this appendix we provide a proof of (10). Let

$$\begin{aligned} X &= (|h_{sd}|^2 + |h_{rd}|^2)P/\sigma_n^2 & (16)\\ Y &= |h_{se}|^2 P/\sigma_n^2 & (17)\\ Z &= |h_{sr}|^2 P/\sigma_n^2 & (18)\end{aligned}$$

Since h_{ij}'s are complex Gaussian, $i, j \in \{s, r, d\}$, the probability density function of X, Y, and Z are given by

$$\begin{aligned} f_X(x) &= \frac{\exp(-x/\gamma_{rd}) - \exp(-x/\gamma_{sd})}{\gamma_{rd} - \gamma_{sd}} & (19)\\ f_Y(y) &= \frac{\exp(-y/\gamma_{se})}{\gamma_{se}} & (20)\\ f_Z(z) &= \frac{\exp(-z/\gamma_{sr})}{\gamma_{sr}} & (21)\end{aligned}$$

where $\gamma_{ij} = E[|h_{ij}|^2]P/\sigma_n^2$. Then,

$$\begin{aligned} P_o(R) &= P(\min\{\log_2(1+Z), \log_2(1+X)\} \\ &\qquad < \log_2(1+Y) + 2R) & (22)\\ &= P(\log_2(1+\min\{X, Z\}) \\ &\qquad < \log_2(1+Y) + 2R) & (23)\\ &= P(2^{-2R}(1+\min\{X,Z\}) - 1 < Y) & (24)\\ &= P(2^{-2R}(1+X) - 1 < Y)P(Z > X) \\ &\quad + P(2^{-2R}(1+Z) - 1 < Y)P(Z < X) & (25)\end{aligned}$$

If $2^{-2R}(1+X)-1 < 0$ or $X < 2^{2R}-1$, then $P(2^{-2R}(1+X)-1 < Y) = 1$ because $Y > 0$. Similarly, if $2^{-2R}(1+Z)-1 < 0$ or $Z < 2^{2R}-1$, then $P(2^{-2R}(1+Z)-1 < Y) = 1$. Therefore, we get

$$\begin{aligned} P_o(R) &= \int_0^{2^{2R}-1} f_X(x) \int_x^\infty f_Z(z)dzdx \\ &+ \int_{2^{2R}-1}^\infty f_X(x) \int_{2^{-2R}(1+x)-1}^\infty f_Y(y)dy \int_x^\infty f_Z(z)dzdx \\ &+ \int_0^{2^{2R}-1} f_Z(z) \int_z^\infty f_X(x)dxdz \\ &+ \int_{2^{2R}-1}^\infty f_Z(z) \int_{2^{-2R}(1+z)-1}^\infty f_Y(y)dy \int_z^\infty f_X(x)dxdz & (26)\\ &= \frac{\gamma_{sr}(1+\gamma_{sr})}{(\gamma_{sr}+\gamma_{rd})(\gamma_{sr}+\gamma_{sd})} \\ &\quad - \frac{e^{-[(2^{2R}-1)/\gamma_{sr}]}(1+\gamma_{sr})2^{-2R}}{(\gamma_{rd}-\gamma_{sd})\gamma_{sr}\gamma_{se}} \\ &\quad \cdot \left[\frac{e^{-[(2^{2R}-1)/\gamma_{rd}]}}{\left(\frac{2^{-2R}}{\gamma_{se}} + \frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{rd}}\right)\left(\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{rd}}\right)} \right. \\ &\quad \left. - \frac{e^{-[(2^{2R}-1)/\gamma_{sd}]}}{\left(\frac{2^{-2R}}{\gamma_{se}} + \frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{sd}}\right)\left(\frac{1}{\gamma_{sr}} + \frac{1}{\gamma_{sd}}\right)} \right] & (27)\end{aligned}$$

Acknowledgement

This work was supported in part by the U.S. National Science Foundation under grant ECCS 1254086.

References

[1] J. Barros and M. Rodrigues, "Secrecy capacity of wireless channels," in *Proc. IEEE Int. Symp. Information Theory*, pp. 356-360, Seattle, WA, Jul. 2006.

[2] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180-2189, Jun. 2008.

[3] J.P. Vilela, M. Bloch, J. Barros, and S.W. McLaughlin, "Wireless secrecy regions with friendly jamming," *IEEE Tr. Infor. Forensics and Security*, pp.256–266, VOL. 6, No. 2, Jun. 2011.

[4] Z. Ding, Member, K.K. Leung, D.L. Goeckel, and D. Towsley, "Opportunistic relaying for secrecy communications: Cooperative jamming vs. relay chatting," *IEEE Tr. on Wireless Commun.*, pp. 1725–1729, Jun. 2011.

[5] L. Dong, Z. Han, A. P. Petropulu, and H. V. Poor, "Improving wireless physical layer security via cooperating relays," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1875-1888, Mar. 2010.

[6] J. Huang, A. Mukherjee, and A.L. Swindlehurst, "Outage performance of amplify-and-forward channels with an unautheticated relay," in *Proc. of IEEE ICC*, 2012.

[7] F. Gabry, R. Thobaben and M. Skoglund, "Outage performances for amplify-and-forward, decode-and-forward and cooperative jamming strategies for the wiretap channel," in *Proc. of IEEE WCNC*, 2011.

[8] L. Lai and H.E.Gamal, "The relay-eavesdropper channel: Cooperation for secrecy," *IEEE Tr. on Infor. Th.*, pp.4005–4019, Sep. 2008.

[9] R. Bassily and S. Ulukus, "Deaf cooperation and relay selection strategies for secure communication in multiple relay networks," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1544–1554, Mar. 2013.

[10] R. Wilson, D. Tse, and R. A. Scholtz,"Channel identification: Secret sharing using reciprocity in ultrawideband Channels," *IEEE Tr. Information Forensics and Security*, Sep. 2007.

[11] W.C.Jakes Jr., *Microwave Mobile Communications*, Wiley, 1974.

[12] L. Xiao, L. Greenstein, N. Mandayam, and W. Trappe, "Fingerprints in the Ether: Using the physical layer for wireless authentication," in *Proc. of IEEE ICC*, 2007.

[13] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[14] S. Leung-Yan-Cheong and M. Hellman, "The Gaussian wire-tap channel," *IEEE Trans. Inf. Theory*, vol. 24, no. 4, pp. 451-456, Jul. 1978.

Optimized Noisy Network Coding for Gaussian Relay Networks

Ritesh Kolte, Ayfer Özgür, Abbas El Gamal
Stanford University
{rkolte, aozgur}@stanford.edu, abbas@ee.stanford.edu

***Abstract*—In this paper, we provide an improved lower bound on the rate achieved by noisy network coding in arbitrary Gaussian relay networks, whose gap to the cutset upper bound depends on the network not only through the total number of nodes but also through the degrees of freedom of the min cut of the network. We illustrate that for many networks this refined lower bound can lead to a better approximation of the capacity. The improvement is based on a judicious choice of the quantization resolutions at the relays.**

I. Introduction

Characterizing the capacity of Gaussian relay networks has been of interest for long time. Recently, significant progress has been made in [1], [2], [3], [4] which show that compress-forward based strategies (such as noisy network coding) can achieve the capacity of any Gaussian relay network within a gap that is independent of the topology of the network, the SNR and the channel coefficients. However, the gap depends linearly on the number of nodes in the network. This limits the applicability of these results to small networks with few relays.

A natural question is whether the gap to capacity can be made smaller than linear in the number of nodes. In this paper, we provide an improved lower bound on the rate achieved by noisy network coding in arbitrary Gaussian relay networks, which for many networks can lead to an approximation gap which is significantly better than the linear gap in [1], [2], [3], [4]. The improvement is based on the observation that in the compress-and-forward based strategies (such as quantize-map-and-forward in [1] and noisy network coding in [2]) there is a fundamental trade-off involved in the choice of the quantization (or compression) resolutions at the relays. If relays quantize their received signals finely, they introduce less quantization noise to the communication. If they quantize more coarsely however, there is a smaller number of quantization indices that need to be communicated to the destination on top of the desired message. This trade-off is not immediately evident from the development of these strategies in [1] and [2], since the employed decoder does not require the quantization indices of the relays to be uniquely decoded. Therefore it is not clear if the quantization indices are indeed decoded at, and thus communicated to the destination, and therefore whether there is a penalty involved in communicating these indices. Based on the work of [5], we argue that in the optimal distribution for the quantization indices, the quantization indices of all relays can be uniquely decoded at the destination. Moreover, an optimal choice of the quantization indices requires much coarser quantization than the noise level. We then apply the new lower bound to a class of layered networks with fixed channel coefficients of unit magnitude and arbitrary phases (i.e. each channel coefficient is of the form $e^{j\theta}$ for some arbitrary $\theta \in [0, 2\pi]$) and show that it leads to a capacity gap that is logarithmic in the number of nodes rather than linear.

A similar insight was used earlier in [6], [7] and [8] to obtain improved capacity approximations for other classes of Gaussian relay networks. [6] and [7] provide an approximation for the capacity of the diamond network which is logarithmic in the number of nodes, while [8] considers a layered network with i.i.d. fast-fading links and shows that the gap to capacity increases logarithmically in the depth of the network. However, in both settings there are other strategies which can yield similar performance. For the diamond network, [6] shows that a partial-decode-and-forward strategy also achieves the logarithmic dependence on the number of nodes, and for the fast fading layered network, ergodic computation-alignment over independent realizations of the fading distribution [9] achieves a gap that does not increase with the number of layers. (Note that both these alternative schemes require increased CSI at the relays and the source nodes.) However, for the layered network with fixed channel gains considered in this paper, these schemes are not applicable and we know of no scheme other than compress-forward that can give a constant gap capacity approximation.

II. Gap to Capacity with Noisy Network Coding

In this section, we discuss the elements of the gap between the rate achieved by noisy network coding (NNC) and the cutset-upper bound and identify a trade-off between different elements of the gap. Our main result in the next section builds on the understanding of this trade-off.

Consider an arbitrary discrete memoryless network with a set of nodes $\mathcal{N}$ where a source node s wants to communicate to a destination node d with the help of the remaining nodes acting as relays. NNC can achieve a communication rate [2, Theorem 1]:[1]

$$\min_{\Omega \subseteq \mathcal{N}} I(X_\Omega; \hat{Y}_{\Omega^c} | X_{\Omega^c}) - I(Y_\Omega; \hat{Y}_\Omega | X_{\mathcal{N}}, \hat{Y}_{\Omega^c}) \qquad (1)$$

for any distribution of the form $\prod_{k \in \mathcal{N}} p(x_k) p(\hat{y_k} | y_k, x_k)$; where for brevity of expressions of, $\hat{Y}_{\Omega^c}$ is assumed to include Y_d. Comparing this with the information-theoretic

[1] In this paper, we need to consider only $s - d$ cuts, which means $s \in \Omega, d \in \Omega^c$. Hence we do not state this explicitly.

cutset upper bound on the capacity of the network given by [10, Theorem 15.10.1]

$$\overline{C} = \sup_{X_{\mathcal{N}}} \min_{\Omega \subseteq \mathcal{N}} I(X_\Omega; Y_{\Omega^c} \mid X_{\Omega^c}) \tag{2}$$

we observe the following differences. First, while the maximization in (2) is over all possible input distributions, only independent input distributions are admissible in (1). The gap corresponds to a potential beamforming gain that is allowed in the upper bound but not exploited by NNC. Second, the first term in (1) is similar to (2) but with Y_{Ω^c} in (2) replaced by $\hat{Y}_{\Omega^c}$ in (1). The difference corresponds to a rate loss due to the quantization noise introduced by the relays. Third, there is the extra term $I(Y_\Omega; \hat{Y}_\Omega | X_{\mathcal{N}}, \hat{Y}_{\Omega^c})$ reducing the rate in (1). One way to potentially interpret this term would be as the rate penalty for communicating the quantized (compressed) observations $\hat{Y}_\Omega$ to the destination on top of the desired message. Note that this is the rate required to describe the observations Y_Ω at the resolution of $\hat{Y}_\Omega$ to a decoder that already knows (or has decoded) $X_{\mathcal{N}}, \hat{Y}_{\Omega^c}$.

However, it is not completely clear if this interpretation is precise because the non-unique decoder employed by NNC does not require the quantization indices to be explicitly decoded. The non-unique decoder of NNC searches for the unique source codeword that is jointly typical with a (not necessarily unique) set of quantization indices at the relays and the received signal at the destination. The following example in Figure (1) illustrates that in certain cases the decoder can indeed recover the transmitted message even if it can not uniquely recover the quantization index of the relay.[2]

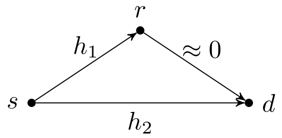


Fig. 1: Example

Consider the classical relay channel with a very weak link from the relay to the destination. Clearly, as long as the source uses a codebook of rate less than the capacity of the direct link, no matter what the operation at the relay is, the destination can always decode the source message by performing a joint typicality test between its received signal and the source codebook (which is subsumed by the non-unique typicality test of NNC). In particular, if the relay quantizes too finely, then there is no way for the destination to recover the relay's quantization index, even though the source message can still be recovered.

On the other hand, this example reveals the following strange property of the expression in (1). While the above discussion reveals that in the setup of Fig. 1, the rate achieved by NNC is equal to the capacity of the direct link independent of the relay's operation (i.e. what $\hat{Y}_r$ is), the rate in (1) is decreasing with increasing resolution for the quantization at the relay (due to the subtractive term $I(Y_\Omega; \hat{Y}_\Omega | X_{\mathcal{N}}, \hat{Y}_{\Omega^c})$). This suggests a more careful analysis of the rate achieved by NNC which leads to the following improved rate:

$$\max_{\mathcal{M} \subseteq \mathcal{N}} \min_{\Omega \subseteq \mathcal{M}} I(X_\Omega; \hat{Y}_{\Omega^c} | X_{\Omega^c}) - I(Y_\Omega; \hat{Y}_\Omega | X_{\mathcal{M}}, \hat{Y}_{\Omega^c}). \tag{3}$$

Here only a subset $\mathcal{M} \subseteq \mathcal{N}$ of the relays is considered in the non-unique typicality decoding, while the other relay transmissions are treated as noise.

It has been shown in [5] that if $\mathcal{M}^*$ is the subset that maximizes (3) for a given $\prod_{i \in \mathcal{N}} p(x_i) p(\hat{y}_i | y_i, x_i)$, then the quantization indices of the relays in $\mathcal{M}^*$ can be uniquely decoded at the destination, while the quantization indices of the relays in $\mathcal{N} \setminus \mathcal{M}^*$ cannot be decoded and in fact, it is optimal to treat the transmissions from these relays as noise. Since the transmissions from $\mathcal{N} \setminus \mathcal{M}^*$ are treated as noise in (3), the rate can be further improved if these relays are shut down. Hence, we can conclude that in the optimal distribution $\prod_{i \in \mathcal{N}} p(x_i) p(\hat{y}_i | y_i, x_i)$, some relays can be off (not utilized or equivalently always quantizing their received signals to zero) and some relays can be active, but the quantization indices of all relays can be uniquely decoded at the destination. Thus, $I(Y_\Omega; \hat{Y}_\Omega | X_{\mathcal{M}}, \hat{Y}_{\Omega^c})$ can indeed be interpreted as the associated rate penalty for communicating these indices.

The above discussion reveals that NNC communicates not only the source message but also the quantization indices to the destination; and while making quantizations finer introduces less quantization noise in the communication, it leads to a larger rate penalty for communicating the quantization indices. This tradeoff is made explicit in the following section.

III. Main Result

Consider a Gaussian relay network where a source node s communicates to a destination node d with the help of a set of relay nodes. The signal received by node i is given by

$$\mathbf{Y}_i = \sum_{j \neq i} H_{ij} \mathbf{X}_j + \mathbf{Z}_i$$

where H_{ij} is the $N_i \times M_j$ channel matrix from node j equipped with M_j transmit antennas to node i equipped with N_i receive antennas. We assume that each node is subject to an average power constraint P per antenna and $\mathbf{Z}_i \sim \mathcal{CN}(0, \sigma^2 I)$, independent across time and across different receive antennas. Let N be the total number of receive antennas and M be the total number of transmit antennas in the network. Also, define

$$C_Q^{i.i.d.}(\Omega) \triangleq \log\det\left(I + \frac{P}{(Q+1)\sigma^2} H_{\Omega \to \Omega^c} H_{\Omega \to \Omega^c}^\dagger\right),$$

which is the mutual information across the cut Ω if the channel input distribution at node j is i.i.d. $\mathcal{CN}(0, PI)$ and the noise is i.i.d. $\mathcal{CN}(0, (Q+1)\sigma^2)$. The matrix $H_{\Omega \to \Omega^c}$ denotes the induced MIMO matrix from Ω to Ω^c and log denotes the natural logarithm. The main result of this paper is given in the following theorem.

Theorem 1. *The rate achieved by noisy network coding in this network can be lower bounded by*

$$C \geq \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - \frac{N}{Q} - d_Q^* \log(Q+1),$$

[2] Even though we focus on the extremal case where the $r - d$ link is zero, the discussion extends to the case where this link is sufficiently weak.

for any non-negative Q where $\overline{C}$ is the cutset upper bound on the capacity of the network given in (2) *and d_Q^* is the degrees-of-freedom (*DOF*) of the* MIMO *channel corresponding to the cut Ω_Q^* that minimizes $C_Q^{i.i.d.}(\Omega)$, denoted succinctly as*

$$d_Q^* = \text{DOF}\left(\arg\min_{\Omega} C_Q^{i.i.d.}(\Omega)\right).$$

The proof of Theorem 1 is presented in Section IV.

Remark 1. *The result can be extended to the case of multiple multicast, i.e. when multiple sources are multicasting their information to a common group of destination nodes.*

Note that Q in the theorem is a free parameter that can be optimized for a given network to minimize the gap between the achieved rate and the cutset upper bound. $Q\sigma^2$ corresponds to the variance of the quantization noise introduced at the relays; larger Q corresponds to coarser quantization. In previous works [1], [2], Q is chosen to be constant independent of the number of nodes (or antennas) N (i.e. $Q \approx 1$ and the quantization noise $Q\sigma^2$ is of the order of the Gaussian noise variance σ^2). This results in an overall gap that is linear in N. Note that both d_0^* and d_Q^* can be trivially upper bounded by N. However, in many cases, the min cut of the network can have much smaller DOF than M and N and in such cases allowing Q to depend on N can result in a much smaller gap.

For example, in the diamond network with single antenna at each node it is clear a priori that any cut of the network has at most two degrees of freedom, regardless of the number of relays, and therefore $d_Q^* \leq 2$ for any Q. It can be seen immediately from the above theorem that choosing $Q = N$ in this case results in a gap logarithmic in N [6], [7], which compares favorably with a gap that is linear in N. Similarly, for the fast-fading layered network with K single antenna nodes per layer considered in [8], it is the case that $d_Q^* \leq K$ for any Q. If there are D layers in the network so that $N = M = KD$, the above expression tells us that choosing Q to be proportional to D gives a gap that is logarithmic in D instead of linear in D. In Section V, we demonstrate another setting in which applying Theorem 1 with Q increasing with the number of layers in the network allows us to obtain an improved gap. This demonstrates that the rule of thumb in the current literature to quantize received signals at the noise level ($Q \approx 1$) can be highly suboptimal.

IV. Proof of Theorem 1

We know that the rate in (1) is achievable in the Gaussian network for any $\prod_{k\in\mathcal{N}} p(x_k)p(\hat{y}_k|y_k, x_k)$ that satisfies the power constraint. We choose the channel input vector at each node j as $\mathbf{X}_j \sim \mathcal{CN}(0, PI)$ and $\hat{Y}_k$ for each receive antenna in the network is chosen such that

$$\hat{Y}_k = Y_k + \hat{Z}_k \text{ where } \hat{Z}_k \sim \mathcal{CN}(0, Q\sigma^2),$$

independent of everything else.

Consider the achievable rate expression in (1). We first show that $\max_{\Omega\subseteq\mathcal{N}} I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) \leq \frac{N}{Q}$. This follows on similar lines as [8, Lemma 1].

$$\begin{aligned} I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) &\leq h(\hat{Y}_\Omega|X_\mathcal{N}) - h(\hat{Y}_\Omega|Y_\Omega, X_\mathcal{N}) \\ &= (\sum_{j\in\Omega} N_j)\log\left(1 + \frac{1}{Q}\right) \leq \frac{N}{Q}. \quad (4) \end{aligned}$$

We now lower bound the first term in (1). Let Ω_Q^* denote $\arg\min C_Q^{i.i.d.}(\Omega)$. Then,

$$\begin{aligned} &\min_{\Omega} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) = \min_{\Omega} C_Q^{i.i.d.}(\Omega) = C_Q^{i.i.d.}(\Omega_Q^*) \\ &\overset{(a)}{\geq} C_0^{i.i.d.}(\Omega_Q^*) - d_Q^* \log(Q+1) \\ &\geq C_0^{i.i.d.}(\Omega_0^*) - d_Q^* \log(Q+1) \\ &\overset{(b)}{\geq} \max_{X_\mathcal{N}} I(X_{\Omega_0^*}; Y_{(\Omega_0^*)^c} \,|\, X_{(\Omega_0^*)^c}) - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - d_Q^* \log(Q+1) \\ &\geq \max_{X_\mathcal{N}} \min_{\Omega} I(X_\Omega; Y_{\Omega^c} \,|\, X_{\Omega^c}) - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - d_Q^* \log(Q+1) \\ &\quad = \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - d_Q^* \log(Q+1), \quad (5) \end{aligned}$$

where

(a) is justified by the following:

$$\begin{aligned} &C_Q^{i.i.d.}(\Omega_Q^*) \\ &\quad = \log\det\left(I + \frac{P}{(Q+1)\sigma^2} H_{\Omega_Q^*\to(\Omega_Q^*)^c} H^\dagger_{\Omega_Q^*\to(\Omega_Q^*)^c}\right) \\ &\quad \geq \log\det\left(I + \frac{P}{\sigma^2} H_{\Omega_Q^*\to(\Omega_Q^*)^c} H^\dagger_{\Omega_Q^*\to(\Omega_Q^*)^c}\right) \\ &\qquad\qquad - d_Q^* \log(Q+1) \\ &\quad = C_0^{i.i.d.}(\Omega_Q^*) - d_Q^* \log(Q+1), \quad \text{and} \end{aligned}$$

(b) follows from [1, Lemma 6.6] equation (144).

The proof of Theorem 1 follows from (4) and (5). ■

Remark 2. *If there exists a class of cuts $\mathcal{A}$ such that*

$$\min_{\Omega} C_Q^{i.i.d.}(\Omega) \geq \min_{\Omega\in\mathcal{A}} C_Q^{i.i.d.}(\Omega) - \kappa$$

for all Q, where κ is a constant, then the gap in Theorem 1 can be possibly improved to

$$\tilde{d}_0^* \log\left(1 + \frac{M}{\tilde{d}_0^*}\right) + \frac{N}{Q} + \tilde{d}_Q^* \log(Q+1) + \kappa, \quad (6)$$

where

$$\tilde{d}_Q^* \triangleq \text{DOF}\left(\arg\min_{\Omega\in\mathcal{A}} C_Q^{i.i.d.}(\Omega)\right). \quad (7)$$

This can be seen by modifying the proof of the lower bound (5) *slightly as:*

$$\begin{aligned} \min_{\Omega} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) &= \min_{\Omega} C_Q^{i.i.d.}(\Omega) \\ &\geq \min_{\Omega\in\mathcal{A}} C_Q^{i.i.d.}(\Omega) - \kappa \\ &\geq \min_{\Omega\in\mathcal{A}} C_0^{i.i.d.}(\Omega) - \tilde{d}_Q^* \log(Q+1) - \kappa \\ &\geq \overline{C} - \tilde{d}_0^* \log\left(1 + \frac{M}{\tilde{d}_0^*}\right) - \tilde{d}_Q^* \log(Q+1) - \kappa. \end{aligned}$$

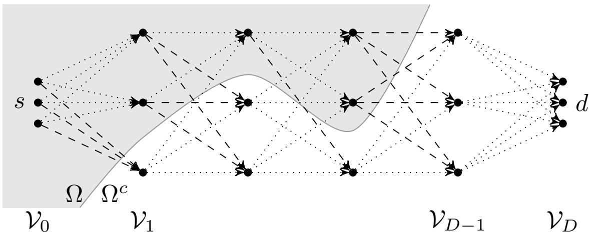


Fig. 2: The cut Ω depicted here $\notin \mathcal{A}$ since the crossing links come from 4 layers, and $4 > K-1 = 2$.

V. Layered Network with Multiple Relays

In this section, we apply Theorem 1 to obtain an improved approximation for the capacity of a layered network in which a source node with K transmit antennas communicates to a destination node with K receive antennas over $D-1$ layers of relays each containing K single-antenna nodes (see Figure 2). Since the network is layered, for $0 \le i \le D-1$, the received signal at nodes in $\mathcal{V}_{i+1}$ (or antennas if $i = D-1$) depends only on the transmit signals of nodes in $\mathcal{V}_i$ and at time t is given by

$$Y_{\mathcal{V}_{i+1}}[t] = H_{\mathcal{V}_i \to \mathcal{V}_{i+1}} X_{\mathcal{V}_i}[t] + Z_{\mathcal{V}_{i+1}}[t],$$

where $Y_{\mathcal{V}_{i+1}}$ and $X_{\mathcal{V}_i}$ are vectors containing the received and transmitted signals at nodes in $\mathcal{V}_{i+1}$ and $\mathcal{V}_i$ respectively; and $Z_{\mathcal{V}_{i+1}} \sim \mathcal{CN}(0, \sigma^2 I)$. The (k,l)'th entry of the matrix $H_{\mathcal{V}_i \to \mathcal{V}_{i+1}}$ denotes the channel coefficient from l'th relay in $\mathcal{V}_i$ to k'th relay in $\mathcal{V}_{i+1}$ at time t and we assume that it is an arbitrary fixed complex number with unit magnitude, i.e., of the form $e^{j\theta_{kl}}$ for some $\theta_{kl} \in [0, 2\pi]$. The phases θ_{kl} are arbitrary for different links. All transmitting nodes are subject to an average power constraint P. We can assume that $Y_{\mathcal{V}_0} = 0$ and $X_d = 0$. Note that $N = M = KD$. We have the following lower bound on the capacity C of this network.

Theorem 2. *For $K \ge 2$ and $D \ge 2$,*

$$C \ge \overline{C} - 2K^2 \log D - K \log K - K. \tag{8}$$

This theorem shows that for a fixed number of nodes K per layer, the gap to capacity grows only logarithmically with the number of layers D. We note that the constants in the gap can be carefully optimized for, however to maintain brevity we do not worry about getting the best constants.

VI. Proof of Theorem 2

We first show that for any Q, $\min_\Omega C_Q^{i.i.d.}(\Omega)$ can be approximated upto an additive constant by restricting the minimization to cuts in a particular class. Then, Theorem 2 follows immediately from Remark 2.

For convenience, we call the K^2 entries in $H_{\mathcal{V}_i \to \mathcal{V}_{i+1}}$ as the links in layer i. With this convention in mind, let $\mathcal{A}$ denote the set of $s-d$ cuts Ω for which the links crossing from Ω to Ω^c come from at most $K-1$ layers, e.g. see Figure 2.

Lemma 1. *We have*

$$\min_{\Omega \in \mathcal{A}} C_Q^{i.i.d.}(\Omega) - K \log K \le \min_{\Omega} C_Q^{i.i.d.}(\Omega) \le \min_{\Omega \in \mathcal{A}} C_Q^{i.i.d.}(\Omega).$$

Proof: The upper bound is immediate. The lower bound can be proved as follows. For any cut $\Omega \notin \mathcal{A}$,

$$\begin{aligned} C_Q^{i.i.d.}(\Omega) &= \sum_{i=1}^{D} C_{Q,MIMO}^{i.i.d.}\left(H_{(\mathcal{V}_i \cap \Omega) \to (\mathcal{V}_{i+1} \cap \Omega^c)}\right) \\ &\overset{(a)}{\ge} K \log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) \\ &\overset{(b)}{\ge} C_Q^{i.i.d.}(\mathcal{V}_0) - K \log K \\ &\ge \min_{\Omega \in \mathcal{A}} C_Q^{i.i.d.}(\Omega) - K \log K, \end{aligned}$$

where (a) follows since for any cut $\notin \mathcal{A}$, at least K terms in the summation are non-zero, each lower-bounded by the point-to-point AWGN capacity; and (b) follows by Lemma 2. This concludes the proof of the lemma. ■

Lemma 2. *We have*

$$C_Q^{i.i.d.}(\mathcal{V}_0) \le K \log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) + K \log K.$$

Proof:
$$\begin{aligned} C_Q^{i.i.d.}(\mathcal{V}_0) = &\ \log\det\left(I + \frac{P}{(Q+1)\sigma^2} H_{\mathcal{V}_0 \to \mathcal{V}_1} H_{\mathcal{V}_0 \to \mathcal{V}_1}^\dagger\right) \\ \overset{(a)}{\le} &\ \sum_{i=1}^{K} \log\left(1 + \frac{P}{(Q+1)\sigma^2} \mathbf{h}_i \mathbf{h}_i^\dagger\right) \\ \le &\ K \log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) + K \log K, \end{aligned}$$

where (a) follows by using Hadamard's inequality; $\mathbf{h}_i$ denotes the ith row of $H_{\mathcal{V}_0 \to \mathcal{V}_1}$. ■

The desired result (8) follows from (6) by setting $Q = D-1$ and noting that the DOF (7) of the MIMO channel created by any cut in $\mathcal{A}$ is trivially upper-bounded by K^2. ■

References

[1] A. Avestimehr, S. Diggavi, and D. Tse, “Wireless network information flow: A deterministic approach,” *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1872–1905, 2011.

[2] S. Lim, Y.-H. Kim, A. El Gamal, and S.-Y. Chung, “Noisy network coding,” *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3132–3152, 2011.

[3] A. Ozgur and S. Diggavi, “Approximately achieving gaussian relay network capacity with lattice codes,” in *IEEE Int. Symp. Inf. Theory*, 2010, pp. 669–673.

[4] G. Kramer and J. Hou, “Short-message quantize-forward network coding,” in *Multi-Carrier Systems Solutions (MC-SS), 2011 8th International Workshop on*, 2011, pp. 1–3.

[5] X. Wu and L.-L. Xie, “On the optimal compressions in the compress-and-forward relay schemes,” *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2613–2628, 2013.

[6] B. Chern and A. Özgür, “Achieving the capacity of the n-relay gaussian diamond network within log n bits,” in *IEEE Information Theory Workshop (ITW) Lausanne*, 2012, pp. 377–380.

[7] A. Sengupta, I.-H. Wang, and C. Fragouli, “Optimizing quantize-map-and-forward relaying for gaussian diamond networks,” in *IEEE Information Theory Workshop (ITW), Lausanne*, 2012, pp. 381–385.

[8] R. Kolte and A. Özgür, “Improved capacity approximations for gaussian relay networks,” in *IEEE Information Theory Workshop (ITW) Seville*, 2013.

[9] U. Niesen, B. Nazer, and P. Whiting, “Computation alignment: Capacity approximation without noise accumulation,” *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3811–3832, 2013.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

Sequential Transmission of Markov Sources over Burst Erasure Channels

Farrokh Etezadi and Ashish Khisti
University of Toronto
10 King's College Road
Toronto, ON, Canada
Email: {fetezadi,akhisti}@comm.utoronto.ca

***Abstract*—We study the sequential coding of a Markov source process under error propagation and decoding delay constraints. The source sequence at each time is sampled i.i.d. along the spatial dimension and from a first order Markov process along the temporal dimension. The channel can introduce a single erasure burst of maximum length B starting at any arbitrary time. The encoder must operate in a causal fashion, whereas the decoder must losslessly reconstruct each source sequence with a delay of T, except those sequences that occur in a window of length $B+W$ following the erasure burst. We study the minimum achievable compression rate as a function of B, W and T. When specialized to $T=0$, we recover our earlier results for the case of zero decoding delay. We also treat an extension where the channel introduces multiple erasure bursts, separated by a guard interval of a certain minimum duration. Finally we present a special source model for which an exact characterization of the minimum rate is obtained.**

I. Introduction

There exists a fundamental tradeoff between the compression rate and error propagation at the receiver in any video compression scheme. At one extreme stands the predictive coding that attains the minimum possible rate when the channel is an ideal bit pipe. However it is very sensitive to packet losses. At the other extreme is the still image coding that is robust to the channel losses but requires high transmission rates. Many practical systems involve a combination of these schemes to strike a balance between the compression rate and error propagation.

In earlier works [1], [2] we introduced the information theoretic framework to characterize the tradeoff between error propagation and compression rate. We studied the sequential transmission of a spatially i.i.d. temporally first order Markov source process, over a burst erasure channel model that introduces a single erasure burst of maximum length B. The encoder must operate in a causal fashion, whereas the decoder must reconstruct each source sequence with zero delay in a lossless fashion. However the decoder is not required to reconstruct those source sequences that belong to a window of length $B+W$ following the start of an erasure burst. We studied the minimum required compression rate in this setup and defined it to be the *rate-recovery function*. In this work we consider the case when the decoder must recover each source sequence within a delay of T in a lossless fashion. Our results reduce to the results of [1, Theorem 1] for the case $T=0$.

Problems involving sequential coding and compression have been studied from many different perspectives in the literature. Our present work builds upon the problem of sequential coding of correlated sources introduced by Viswanathan and Berger [3]. In this setup, a set of correlated sources must be sequentially compressed by the encoder, whereas the decoder at each stage is required to reconstruct the corresponding source sequence, given all the encoder outputs up to that time. It is noted in [3] that the correlated source sequences can model consecutive video frames and each stage at the decoder maps to sequential reconstruction of a particular source frame. However the setup considered in [3] and followup works assumes that the channel is an ideal bit-pipe and does not consider the effect of packet losses over the channel. In this work we consider such a setup when the channel introduces burst erasures.

II. Problem Statement

We consider a semi-infinite stationary vector source process $\{\mathbf{s}_i^n\}_{i\geq 0}$ 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:

$$\begin{aligned}&\Pr(\ \mathbf{s}_i^n = s_i^n \mid \mathbf{s}_{i-1}^n = s_{i-1}^n,\ \mathbf{s}_{i-2}^n = s_{i-2}^n, \ldots)\\ &\quad = \prod_{j=1}^{n} p_{\mathbf{s}_1|\mathbf{s}_0}(s_{i,j}|s_{i-1,j}), \quad \forall i \geq 0. \end{aligned} \tag{1}$$

We assume that the underlying random variables $\{\mathbf{s}_i\}$ constitute a time-invariant, stationary and a first-order Markov chain with a common marginal distribution denoted by $p_{\mathbf{s}}(\cdot)$. We remark that the results may also generalize when the source sequence is a stationary process (not necessarily i.i.d.) in the spatial dimension. The sequence $\mathbf{s}_{-1}^n$, as a synchronization frame, is revealed to both the encoder and decoder before the start of the communication.

A rate-R causal encoder maps the sequence $\{\mathbf{s}_i^n\}_{i\geq 0}$ to an index $f_i \in [1, 2^{nR}]$ according to some function

$$f_i = \mathcal{F}_i\left(\mathbf{s}_0^n, \ldots, \mathbf{s}_i^n, \mathbf{s}_{-1}^n\right) \tag{2}$$

for each $i \geq 0$. The channel introduces an erasure burst of size B, i.e. for some particular $j \geq 0$, it introduces an erasure burst such that $g_i = \star$ for $i \in \{j, j+1, \ldots, j+B-1\}$ and $g_i = f_i$ otherwise.

Upon observing the sequence $\{g_i\}_{i\geq 0}$ the delay-constrained decoder is required to perfectly recover all the

This work was supported by a Natural Science Engineering Research Council (NSERC) discovery grant and the Canada Research Chairs program.

source sequences using decoding functions

$$\hat{s}_i^n = \mathcal{G}_i(s_{-1}^n, g_0, g_1, \dots, g_{i+T}). \tag{3}$$

for $i \notin \{j, \dots, j+B+W-1\}$, where j denotes the time at which the erasure burst starts. It is however not required to produce the source sequences in the window of length $B+W$ following the start of an erasure burst.

A rate $R(B, W, T)$ is feasible if there exists a sequence of encoding and decoding functions and a sequence ϵ_n that approaches zero as $n \to \infty$ such that, $\Pr(s_i^n \neq \hat{s_i}^n) \leq \epsilon_n$ for all $i \notin \{j, ..., j+B+W-1\}$. We seek the minimum feasible rate $R(B, W, T)$, which we define to be the *lossless rate-recovery* function for *delay-constrained* decoder. For compactness throughout the paper, we refer to this function simply as *rate-recovery* function.

III. Main Results

In this paper we consider the case where the channel introduces an isolated erasure burst of length up to B during the transmission duration. The following theorem characterizes the upper and lower bounds on rate-recovery function

Theorem 1. *The rate-recovery function of discrete Markov sources with delay-constrained decoders satisfies*

$$R^{-}(B, W, T) \leq R(B, W, T) \leq R^{+}(B, W, T)$$

where

$$R^{-}(B, W, T) = H(s_1|s_0) + \frac{1}{W+T+1} I(s_B; s_{B+W+1}|s_0) \tag{4}$$

$$R^{+}(B, W, T) = H(s_1|s_0) + \frac{1}{W+T+1} I(s_B; s_{B+1}|s_0) \tag{5}$$

□

It can be observed from Theorem 1 that both the upper and lower bounds consists of an entropy term plus another mutual information term inversely scaled by $(W+T+1)$. We can interpret the term $H(s_1|s_0)$ as the amount of uncertainty in s_i when the past sources are perfectly known. This term is equivalent to the rate associated with ideal predictive coding in absence of any erasures. The second term in both (4) and (5) is the additional penalty that arises due to the recovery constraint following an erasure burst. Note that the mutual information term associated with the lower bound is $I(s_B; s_{B+W+1}|s_0)$ while that in the upper bound is $I(s_B; s_{B+1}|s_0)$. Intuitively this difference arises because in the lower bound we only consider the reconstruction of s_{B+W+1}^n following an erasure bust in $[1, B]$ while, as explained below in Corollary 1 the upper bound involves a binning based scheme that reconstructs all sequences $(s_{B+1}^n, \dots, s_{B+W+1}^n)$ at time $t = B+W+T+1$.

The proof of Theorem 1 is discussed in Sec. IV. The lower bound is based on the idea of considering a periodic burst erasure channel rather than single burst erasure channel. The upper bound is based on random-binning coding scheme. The following proposition provides an alternative expression for the achievable rate. The proof is omitted due to the lack of space.

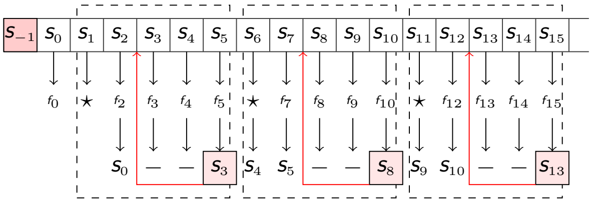


Fig. 1. Periodic burst erasure channel considered in proof of converse.

Corollary 1. *The upper bound in* (5) *can also be expressed as*

$$R^{+}(B, W, T) = \frac{H(s_{B+1}, s_{B+2}, \dots, s_{B+W+T+1}|s_0)}{W+T+1} \tag{6}$$

for the first order Markov source process. □

Note that the upper and lower bounds of Theorem 1 coincide for some special cases which establishes the lossless rate-recovery function.
– When $W = 0$, i.e. the decoder is interested in recovering all the source sequences with non-erased channel outputs within delay of T, the lossless rate-recovery is as follows.

$$R(B, W=0, T) = H(s_1|s_0) + \frac{1}{T+1} I(s_B; s_{B+1}|s_0) \tag{7}$$

$$= \frac{1}{T+1} H(s_{B+1}, s_{B+2}, \dots, s_{B+T+1}|s_0) \tag{8}$$

– When each or both of the variables W and T become very large, i.e. W or $T \to \infty$, the lossless rate-recovery function reduces to the rate required for predictive coding.

Note also that Theorem 1 can be viewed as a generalization of the zero-delay results of [1, Theorem 1] as the upper and lower bounds when $T = 0$ reduce to the results of [1, Theorem 1].

IV. Proof of Theorem 1

A. Achievability

The achievability of the rate expression (5) is based on random binning technique. A Slepian-Wolf codebook is constructed by partitioning the space of all typical sequences $s_i^n \in T_\epsilon^n(s)$ into 2^{nR} bins and the bin index f_i is transmitted at time i. The decoder is required to output $\hat{s}_i^n$ in one of two ways. If it has access to s_{i-1}^n then it finds a sequence jointly typical with $\hat{s}_{i-1}^n$ in the bin index of f_i. This succeeds with high probability if $R \geq H(s_1|s_0)$ which is clearly satisfied in (5).

Next suppose that there is an erasure burst spanning $t \in \{j-B, \dots, j-1\}$. The receiver has access to s_{j-B-1}^n and needs to use $[f]_j^{j+W+T}$ to recover s_{j+W}^n. It simultaneously attempts to decode all of $s_j^n, \dots, s_{j+W+T}^n$ using $f_j, \dots, f_{j+W+T}$ and s_{j-B-1}^n. This succeeds if $(W+T+1)R \geq H(s_j, \dots, s_{j+W+T}|s_{j-B-1})$ which in turn holds via (5) according to Corollary 1. This completes the proof of the achievability.

B. Converse

To derive the lower bound on rate-recovery, we consider a periodic erasure channel with period $P = B + W + T + 1$. The intuition behind considering periodic erasure channel is as follows. Because of the first-order Markov property of the source process, whenever the decoder recovers the source sequence at a particular time, say t, it is expected to become oblivious to the channel erasures before time t. At this point if a new erasure burst is introduced, the decoder can treat it is as the only erasure burst during the whole transmission period. As a result, one simple way to lower bound the rate-recovery function is considering N periods of the periodic erasure channels as follows. To keep the notation simple we only focus on the simple example of $B = W = 1$ and $T = 2$ where the period $P = 5$. Generalization to any B, T and W is analogous and will be treated in the full paper. The periodic erasure channel model is illustrated in Fig. 1. Note that at each period 4 out of 5 channel inputs are observed by the decoder while 3 source sequences are recovered. In particular in the k-th period, $k \geq 0$, 4 channel inputs $\{[\mathsf{f}]_{5k+2}^{5k+5}\}$ are observed and 3 source sequences $\{[\mathbf{s}]_{5k-1}^{5k}, \mathbf{s}_{5k+3}\}$ are recovered[1]. Thus we can write

$$
\begin{aligned}
4nR \geq H([\mathsf{f}]_{5k+2}^{5k+5}) &\geq H([\mathbf{s}]_{5k-1}^{5k}, \mathbf{s}_{5k+3} | \mathbf{s}_{5k-2}) && (9)\\
&= 2nH(\mathsf{s}_1|\mathsf{s}_0) + nH(\mathsf{s}_3|\mathsf{s}_0) && (10)
\end{aligned}
$$

and the rate has to satisfy

$$
R \geq H(\mathsf{s}_1|\mathsf{s}_0) + \frac{1}{4}\left(H(\mathsf{s}_3|\mathsf{s}_0) - 2H(\mathsf{s}_1|\mathsf{s}_0)\right) \qquad (11)
$$

Note that the lower bound on rate in (11) is weaker than (4) since

$$
\begin{aligned}
R^-(B,W,T) &= H(\mathsf{s}_1|\mathsf{s}_0) + \frac{1}{4}I(\mathsf{s}_1;\mathsf{s}_3|\mathsf{s}_0)\\
&= H(\mathsf{s}_1|\mathsf{s}_0) + \frac{1}{4}\left(H(\mathsf{s}_3|\mathsf{s}_0) - H(\mathsf{s}_2|\mathsf{s}_0)\right)\\
&\geq H(\mathsf{s}_1|\mathsf{s}_0) + \frac{1}{4}\left(H(\mathsf{s}_3|\mathsf{s}_0) - H(\mathsf{s}_1,\mathsf{s}_2|\mathsf{s}_0)\right)\\
&= H(\mathsf{s}_1|\mathsf{s}_0) + \frac{1}{4}\left(H(\mathsf{s}_3|\mathsf{s}_0) - 2H(\mathsf{s}_1|\mathsf{s}_0)\right). && (12)
\end{aligned}
$$

We are able to improve on the simple lower bound in (11) to derive the lower bound on rate-recovery function in (4). To this end, we consider N periods of the periodic erasure channel explained before. Rate R should satisfy the following constraint.

$$
\begin{aligned}
4NnR &\geq H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N})\\
&\geq H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0) && (13)
\end{aligned}
$$

where (13) follows from the fact that conditioning reduces the entropy. We provide the proof of the lower bound in four steps.

Step 1: First consider the first period in Fig. 1. According to Fano's inequality and based on the fact that $\mathbf{s}_3$ can be recovered from $\{\mathsf{f}_0, [\mathsf{f}]_2^5, \mathbf{s}_{-1}\}$, we can write

$$
H(\mathbf{s}_3|\mathsf{f}_0, [\mathsf{f}]_2^5, \mathbf{s}_{-1}) \leq n\epsilon_n \qquad (14)
$$

Using this, the entropy term in (13) can be lower bounded as follows.

$$
\begin{aligned}
&H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0) && (15)\\
&\geq H(\mathbf{s}_3, [\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0) - n\epsilon_n && (16)\\
&= nH(\mathsf{s}_3|\mathsf{s}_0) + H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0, \mathbf{s}_3)\\
&\quad - n\epsilon_n && (17)
\end{aligned}
$$

where (16) follows from (14) and the first term in (17) follows from the properties of the source sequences.

Step 2: In this step, based on the fact that conditioning reduces the entropy, we further lower bound the second term in (17) by revealing the erased codewords as follows.

$$
\begin{aligned}
&H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0, \mathbf{s}_3) \geq\\
&H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^1, \mathbf{s}_3, [\mathsf{f}]_0^1) && (18)
\end{aligned}
$$

After revealing the erased codewords of the first period, the source sequences $\mathbf{s}_2$ can be recovered. Thus the following inequality holds.

$$
H(\mathbf{s}_2 | [\mathsf{f}]_0^4, \mathbf{s}_{-1}) \leq n\epsilon_n \qquad (19)
$$

Now the entropy term in (18) can be written as.

$$
\begin{aligned}
&H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^1, \mathbf{s}_3, [\mathsf{f}]_0^1) && (20)\\
&\geq H(\mathbf{s}_2, [\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^1, \mathbf{s}_3, [\mathsf{f}]_0^1) - n\epsilon_n\\
& && (21)\\
&\geq H(\mathbf{s}_2|\mathbf{s}_1, \mathbf{s}_3) +\\
&\quad H([\mathsf{f}]_4^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^3, [\mathsf{f}]_0^3) - n\epsilon_n && (22)\\
&\geq 2nH(\mathsf{s}_1|\mathsf{s}_0) - nH(\mathsf{s}_3|\mathsf{s}_1)\\
&\quad + H([\mathsf{f}]_4^5, [\mathsf{f}]_7^{10}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^3, [\mathsf{f}]_0^3) - n\epsilon_n && (23)
\end{aligned}
$$

Note that (21) follows from (19).

Step 3: In this step we exploit the fact that the source sequences in the interval $[4, 5]$ can also be recovered according to the following inequality.

$$
H([\mathbf{s}]_4^5 | [\mathsf{f}]_0^5, \mathsf{f}_7, \mathbf{s}_{-1}) \leq 2n\epsilon_n \qquad (24)
$$

(24) can be used to lower bound the last entropy term in (23) as follows.

$$
\begin{aligned}
&H([\mathsf{f}]_4^5, [\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{NP} | [\mathbf{s}]_{-1}^3, [\mathsf{f}]_0^3) && (25)\\
&\geq H([\mathbf{s}]_4^5, [\mathsf{f}]_4^5, [\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^3, [\mathsf{f}]_0^3)\\
&\quad - 2n\epsilon_n && (26)\\
&= H([\mathbf{s}]_4^5 | \mathbf{s}_3) +\\
&\quad H([\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^5, [\mathsf{f}]_0^5) - 2n\epsilon_n && (27)\\
&= 2nH(\mathsf{s}_1|\mathsf{s}_0) +\\
&\quad H([\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^5, [\mathsf{f}]_0^5) - 2n\epsilon_n && (28)
\end{aligned}
$$

where (26) follows from (24).

Step 4: The last step is considering all the N periods simultaneously and repeatedly exploiting the same methods in steps 1 to 3. In particular by combining (17), (23) and (28) we have

$$
\begin{aligned}
4NnR &\geq H([\mathsf{f}]_2^5, [\mathsf{f}]_7^{10}, [\mathsf{f}]_{12}^{15}, \ldots, [\mathsf{f}]_{5(N-1)+2}^{5N} | [\mathbf{s}]_{-1}^0) && (29)\\
&\geq 4nH(\mathsf{s}_1|\mathsf{s}_0) + nI(\mathsf{s}_1;\mathsf{s}_3|\mathsf{s}_0)
\end{aligned}
$$

[1] Bold fonts indicate the n-length source sequences, i.e. $\mathbf{s}_i \triangleq \mathsf{s}_i^n$.

$$+H([f]_7^{10},[f]_{12}^{15},\dots,[f]_{5(N-1)+2}^{5N}|[\mathsf{s}]_{-1}^{5},[f]_0^5)-4n\epsilon_n \tag{30}$$

By $(N-1)$ times repeating the same methods used in steps 1–3 for $(N-1)$ periods and leaving the entropy terms of the N-th period, the entropy term in (30) can be lower bounded as follows.

$$4NnR \geq H([f]_2^5,[f]_7^{10},[f]_{12}^{15},\dots,[f]_{5(N-1)+2}^{5N}|[\mathsf{s}]_{-1}^{0}) \tag{31}$$

$$\begin{aligned}&\geq 4n(N-1)H(\mathsf{s}_1|\mathsf{s}_0)+n(N-1)I(\mathsf{s}_1;\mathsf{s}_3|\mathsf{s}_0)\\ &\quad + H([f]_{5(N-1)+2}^{5N}|[\mathsf{s}]_{-1}^{5(N-1)})-4(N-1)\epsilon_n\end{aligned} \tag{32}$$

Finally by dividing (32) by $4Nn$ and taking $n \to \infty$ and thereafter $N \to \infty$ we recover (4).

This completes the proof of the lower bound for this example. The proof of general case is similar and is omitted due to the lack of space.

V. Additional Results

In this section we provide two additional results. The proofs are omitted here.

A. Rate-Recovery for Sliding-Window Burst Erasure Channel

In order to investigate the effect of channels with multiple erasures, we consider the sliding-window burst erasure channel model. In this model the channel can introduce multiple erasure bursts each of length up to B during the transmission period, however there is a guaranteed guard interval of length at least L between each consecutive erasure bursts. The rest of the set up is similar to single erasure case. Note that in our setting $L > W$, i.e. the guard between the erasures has to be larger than the waiting non-recovery period. The following theorem characterizes the upper and lower bounds on rate-recovery function for sliding-window burst erasure channel model denoted as $R_{\mathrm{ME}}(B,W,T,L)$.

Theorem 2. *The rate-recovery function for sliding-window burst erasure channel satisfies*

$$R^-_{ME}(B,W,L,T) \leq R_{ME}(B,W,T,L) \leq R^+_{ME}(B,W,L,T)$$

where

$$\begin{aligned}&R^-_{ME}(B,W,L,T)\\ &\triangleq H(s_1|s_0)+\frac{1}{\min\{L,T+W+1\}}I(s_B;s_{B+W+1}|s_0)\end{aligned} \tag{33}$$

$$\begin{aligned}&R^+_{ME}(B,W,L,T)\\ &\triangleq H(s_1|s_0)+\frac{1}{\min\{L,T+W+1\}}I(s_B;s_{B+1}|s_0)\end{aligned} \tag{34}$$

□

It can be observed from Theorem 2 that for $T \leq L-W-1$, the results of Theorem 1 for rate-recovery function of single burst erasure channel model also hold for the sliding-window burst erasure model. The main intuition behind this fact is that as soon as the decoder recovers the source sequences at a specific time, because of the Markov property of the source model, it becomes oblivious to the erasure bursts happened in the past. Thus it treats the new burst erasure as a single burst erasure as if there has been no previous erasures. On the other hand when $T \geq L-W-1$ our lower and upper bounds in Theorem 2 does not depend on the delay parameter T. The upper bound is based on random binning scheme and reveals that if $T > L-W-1$ there is no benefit of delay more than $L-W-1$. These bounds indicate that restricting the decoder to perform within the delay of $L-W-1$ may not affect capacity.

B. Diagonally Correlated Deterministic Sources

As stated so far, the upper and lower bounds of Theorem 1 do not coincide in general. A natural question arises as to whether the binning based scheme is always optimal in the streaming setup or whether it can be improved. Clearly one feature of the binning scheme is that it forces the *simultaneous recovery* of all the sources in the recovery window whose bin indices are used by the decoder. We show that such a simultaneous recovery within a specific delay in general is suboptimal at least for a class of sources. In fact for our special class of sources, the lower bound is tight and the binning based upper bound is loose. This counterexample shows that the rate recovery function does not coincide with the binning based upper bound in general.

Definition 1. *(Diagonally Correlated Deterministic Sources) The alphabet of a* diagonally correlated deterministic source *consists of $K+1$ sub-symbols i.e.,*

$$\mathbf{s}_i = (\mathbf{s}_{i,0},\dots,\mathbf{s}_{i,K}) \in \mathcal{S}_0 \times \mathcal{S}_1 \times \dots \times \mathcal{S}_K, \tag{35}$$

where each $\mathcal{S}_i = \{0,1\}^{N_i}$ is a binary sequence. Suppose that the sub-sequence $\{\mathbf{s}_{i,0}\}_{i\geq 0}$ is an i.i.d. sequence sampled uniformly over $\mathcal{S}_0$ and for $1 \leq j \leq K$, the sub-symbol $\mathbf{s}_{i,j}$ is a linear deterministic function[2] *of $\mathbf{s}_{i-1,j-1}$ i.e.,*

$$\mathbf{s}_{i,j} = \mathbf{R}_{j,j-1}\cdot\mathbf{s}_{i-1,j-1}, \qquad 1 \leq j \leq K. \tag{36}$$

for fixed matrices $\mathbf{R}_{1,0},\mathbf{R}_{2,1}\dots,\mathbf{R}_{K,K-1}$ each of full row-rank i.e., $\mathrm{rank}(\mathbf{R}_{j,j-1}) = N_j$.

For such a class of sources we establish that the lower bound in Theorem 1 is tight and the binning based scheme is sub-optimal.

Proposition 1. *For the class of Diagonally Correlated Deterministic Sources in Def. 1 the rate-recovery function is also given by:*

$$\begin{aligned}&R(B,W,T) = R^-(B,W,T)\\ &= H(\mathbf{s}_1|\mathbf{s}_0)+\frac{1}{W+T+1}I(\mathbf{s}_B;\mathbf{s}_{B+W+1}|\mathbf{s}_0)\end{aligned} \tag{37}$$

$$= N_0 + \frac{1}{W+T+1}\sum_{k=1}^{\min\{[K-W]^+,B\}} N_{W+k}. \tag{38}$$

Our coding scheme exploits the special structure of such sources and achieves a rate that is strictly lower than the binning based scheme.

References

[1] F. Etezadi, A. Khisti, and M. Trott, “Zero-delay sequential transmission of markov sources over burst erasure channels,” To Appear, IEEE Trans. on Info. Theory. available at http://www.comm.utoronto.ca/akhisti/eks.pdf.

[2] A. Khisti, F. Etezadi, and M. Trott, “Real-time coding of markov sources over erasure channels: When is binning optimal?” International Zurich Seminar on Communications, 2012.

[3] H. Viswanathan and T. Berger, “Sequential coding of correlated sources,” *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 236 –246, jan 2000.

[2] All multiplication is over the binary field.

Information Loss and Anti-Aliasing Filters in Multirate Systems

Bernhard C. Geiger and Gernot Kubin
Graz University of Technology, Austria
{geiger,g.kubin}@ieee.org

***Abstract*—This work investigates the information loss in a decimation system, i.e., in a downsampler preceded by an anti-aliasing filter. It is shown that, without a specific signal model in mind, the anti-aliasing filter cannot reduce information loss, while for a simple signal-plus-noise model it can. For the Gaussian case, the optimal anti-aliasing filter is shown to coincide with the one obtained from energetic considerations. For a non-Gaussian signal corrupted by Gaussian noise, the Gaussian assumption yields an upper bound on the information loss, suggesting filter design principles based on second-order statistics.**

I. Introduction

Multi-rate systems are ubiquitously used in digital systems to increase (upsample) or decrease (downsample) the rate at which a signal is processed. Especially downsampling is a critical operation since it can introduce aliasing, like sampling, and thus can cause information loss. Standard textbooks on signal processing deal with this issue by recommending an anti-aliasing filter prior to downsampling – resulting in a cascade which is commonly known as a decimator [1, Ch. 4.6]. In these books, this anti-aliasing filter is usually an ideal low-pass filter with a cut-off frequency of π/M, for an M-fold decimation system (cf. Fig. 1). Unser showed that this choice is optimal in terms of the mean-squared reconstruction error (MSE) only if the input process is such that the passband portion of its power spectral density (PSD) exceeds all aliased components [2]. Similarly, it was shown by Tsatsanis and Giannakis [3], that the filter minimizing the MSE is piecewise constant, M-aliasing-free (i.e., the aliased components of the M-fold downsampled frequency response do not overlap), and has a passband depending on the PSD of the input process. Specifically, the filter which permits most of the energy to pass aliasing-free is optimal in the MSE sense.

In this paper we consider a design objective vastly different from the MSE: information. The fact that information, compared to energy, can yield more successful system designs has long been recognized, e.g., for (non-linear) adaptive filters [4] or for state estimation using linear filters [5]. In information theory, transceiver filter design based on mutual information is covered in, e.g., [6], [7]. That information-theoretic design seems to become a trend recently is understandable: After all, it is information one wants to transmit, not energy. Finally, quantifying information relieves us from having to specify a reconstruction procedure: The information lost in the decimation system is independent from signal reconstruction, therefore a separate design of these two system components should be possible.

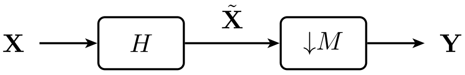


Fig. 1. Simple decimation system consisting of a linear filter H and an M-fold downsampler.

Our first result is surprising: Given mild assumptions on the input process of the decimation system, the information loss can be bounded *independently* of the anti-aliasing filter (see Section III). The reason is that under these assumptions every bit of the input process is treated equivalently, regardless of the amount of energy by which it is represented. In order to remedy this counter-intuitivity, Section IV considers Gaussian processes with a specific signal model in mind: The input to the decimation system is a relevant data signal corrupted by noise. As a corollary to a more general result, we show that for white noise the anti-aliasing filter minimizing the information loss coincides with the optimal filter of [3]. Since in most cases the Gaussian assumption is too restrictive, we let the signal process have arbitrary distribution in Section V, but keep the noise Gaussian. Following the approach of Plumbley in [8], we prove that the Gaussian assumption for the signal process yields an upper bound on the information loss in the general case. In other words, designing a filter based on the PSDs of the signal and noise processes guarantees a minimum information transfer over the decimation system. This justifies a filter design based on second-order statistics, i.e., on energetic considerations, also from an information-theoretic perspective. In Section VI we illustrate our results in a simple toy example.

Due to the lack of space, we only give an outline of our proofs. An extended version of this manuscript is currently in preparation.

II. Preliminaries and Notation

Throughout this work we adopt the following notation: $\mathbf{Z}$ is a real-valued random process, whose n-th sample is the random variable (RV) Z_n. We abbreviate $Z_i^j := \{Z_i, Z_{i+1}, \dots, Z_j\}$. The differential entropy [9, Ch. 8] and the Rényi information dimension [10] of Z_i^j are $h(Z_i^j)$ and $d(Z_i^j)$, respectively, provided these quantities exist and are finite. Finally, we define the M-fold blocking $\mathbf{Z}^{(M)}$ of $\mathbf{Z}$

as the sequence of M-dimensional RVs $Z_1^{(M)} := Z_1^M$, $Z_2^{(M)} := Z_{M+1}^{2M}$, and so on.

In this work, we often consider a process $\mathbf{Z}$ satisfying

Assumption 1. $\mathbf{Z}$ is stationary, has finite marginal differential entropy $h(Z_n)$, finite Shannon entropy of the quantized RV $\lfloor Z_n \rfloor$, and finite differential entropy rate

$$\bar{h}\,(\mathbf{Z}) := \lim_{n\to\infty} \frac{1}{n} h(Z_1^n) = \lim_{n\to\infty} h(Z_n|Z_1^{n-1}). \qquad (1)$$

As a direct consequence of Assumption 1, the information dimension satisfies $d(Z_1^n) = n$ for all n, and the mutual information rate with a process $\mathbf{W}$ jointly stationary with $\mathbf{Z}$ exists and equals [11, Thm. 8.3]

$$\bar{I}(\mathbf{Z};\mathbf{W}) := \lim_{n\to\infty} \frac{1}{n} I(Z_1^n; W_1^n). \qquad (2)$$

We introduce two measures of information loss for stationary stochastic processes: The first is an extension of the *relative information loss*[1] $l(Z \to g(Z))$ to stochastic processes (cf. [12], [13]):

$$l(\mathbf{Z} \to g(\mathbf{Z})) := \lim_{n\to\infty} l((Z_1^n \to g(Z_1^n)) \qquad (3)$$

where we abused notation by applying g coordinate-wise. The second notion is an extension of [14], where we introduced the *relevant information loss*. Let $\mathbf{W}$ be a process statistically related to and jointly stationary with $\mathbf{Z}$, representing the relevant information content of $\mathbf{Z}$; for example, $\mathbf{W}$ might be the sign of $\mathbf{Z}$, or $\mathbf{Z}$ might be a noisy observation of $\mathbf{W}$. Then, the information loss rate relevant w.r.t. $\mathbf{W}$ is

$$\bar{L}_{\mathbf{W}}(\mathbf{Z} \to g(\mathbf{Z})) := \bar{I}(\mathbf{W};\mathbf{Z}) - \bar{I}(\mathbf{W}; g(\mathbf{Z})) \qquad (4)$$

provided the quantities exists.

III. Relative Information Loss in a Downsampler

Consider the scenario depicted in Fig. 1, where $\mathbf{X}$ satisfies Assumption 1. It can be shown that if the linear filter H is stable and causal, also the output process $\tilde{\mathbf{X}}$ satisfies Assumption 1. Moreover, such a filter has no effect on the information content of the stochastic process in the sense that, for $\mathbf{S}$ jointly stationary with $\mathbf{X}$, $\bar{I}(\mathbf{X};\mathbf{S}) = \bar{I}(\tilde{\mathbf{X}};\mathbf{S})$.

To analyze the information loss rate in the downsampling device, we employ the relative information loss rate,

$$l(\tilde{\mathbf{X}}^{(M)} \to \mathbf{Y}) := \lim_{n\to\infty} l((\tilde{X}^{(M)})_1^n \to Y_1^n) \qquad (5)$$

where we applied M-fold blocking to ensure that the mapping between $(\tilde{X}^{(M)})_1^n$ and Y_1^n is static. Downsampling, $Y_n := \tilde{X}_{nM}$, is now a projection to a single coordinate, hence [12]

$$l((\tilde{X}^{(M)})_1^n \to Y_1^n) = \frac{d((\tilde{X}^{(M)})_1^n | Y_1^n)}{d((\tilde{X}^{(M)})_1^n)} = \frac{n(M-1)}{nM}. \qquad (6)$$

Since the filter H is stable and causal and, thus, has no influence on the information content of the stochastic process, we abuse notation in (a) below and combine (3) with (6) to

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) \overset{(a)}{=} l(\tilde{\mathbf{X}}^{(M)} \to \mathbf{Y}) = \frac{M-1}{M}. \qquad (7)$$

The amount of information lost in the decimation system in Fig. 1 is the same for all stable, causal filters H.

The question remains whether an *ideal* anti-aliasing filter can prevent information loss, since it guarantees that the downsampling operation is invertible. To show that the answer to this question is negative, take, for example,

$$H(\mathrm{e}^{\jmath\theta}) = \begin{cases} 1, & \text{if } |\theta| < \frac{\pi}{M} \\ 0, & \text{else} \end{cases}. \qquad (8)$$

We decompose $\mathbf{X}$ in an M-channel filterbank: The k-th channel is characterized by analysis and synthesis filters being constant in the frequency band $(k-1)/M \le |\theta| < k/M$ and zero elsewhere. Let $\mathbf{Y}_{(k)}$ be the (M-fold downsampled) process in the k-th channel – clearly, $\mathbf{Y} \equiv \mathbf{Y}_{(1)}$. It can be shown that every $\mathbf{Y}_{(k)}$ satisfies Assumption 1 if $\mathbf{X}$ is Gaussian. Thus we obtain

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) \overset{(a)}{=} l(\mathbf{Y}_{(1)}, \dots, \mathbf{Y}_{(M)} \to \mathbf{Y}_{(1)}) = \frac{M-1}{M} \qquad (9)$$

where the information is again lost in a projection and where we abused notation in (a) since the filterbank decomposition is invertible. The ideal anti-aliasing low-pass filter prevents information from being lost in the downsampler by *destroying information itself.*

If the filter H is a cascade of a causal, stable filter and of one with a piecewise-constant transfer function (with less trivial intervals as pass-bands), the analysis still holds; Information is either lost in the filter or in the downsampler:

Theorem 1. *For a Gaussian process* $\mathbf{X}$ *satisfying Assumption 1, the relative information loss rate in the decimation system depicted in Fig. 1 satisfies*

$$l(\mathbf{X}^{(M)} \to \mathbf{Y}) \ge \frac{M-1}{M} \qquad (10)$$

for every anti-aliasing filter H *with finitely many pass-band intervals.*

Sketch of the proof: The inequality is trivial, since H can destroy an arbitrarily large amount of information. We sketch the proof for H being piecewise constant with passband intervals having rational endpoints, i.e., being integer multiples of $1/L$ with L sufficiently large. As before, we apply a filterbank decomposition of $\mathbf{X}$, this time with LM channels. The cascade of H and the filters of the filterbank is either identical to zero or to one. Using polyphase decomposition, it can be shown that M-fold downsampling amounts to adding M bands, some of which are set to zero by the filter. Summation is a memoryless operation, hence the information loss depends on the information dimension of the random sums. Since the information dimension of a scalar cannot exceed one, the bound is obtained. ■

[1] Roughly speaking, $l(Z \to g(Z))$ captures the *percentage* of information lost by applying the function g to the RV Z.

The reason for this seemingly counter-intuitive result is that, without a specific signal model, the amount of information is not necessarily proportional to the amount of energy by which it is represented: There is no reason to prefer a specific frequency band over another. This in some sense parallels our result on the relative information loss in principal components analysis (PCA), where we showed that PCA cannot reduce the amount of information being lost in reducing the dimensionality of the data [12].

IV. Relevant Information Loss: Gaussian Case

To remove the counter-intuitivity of the previous section, we adapt the signal model: Let $\mathbf{X}$ be a noisy observation of a signal process $\mathbf{S}$, i.e., $X_n = S_n + N_n$, where $\mathbf{S}$ and $\mathbf{N}$ are independent, jointly stationary Gaussian processes with smooth PSDs $S_S(\mathrm{e}^{\jmath\theta})$ and $S_N(\mathrm{e}^{\jmath\theta})$, respectively, and which satisfy Assumption 1. The information loss rate relevant w.r.t. $\mathbf{S}$ is given by

$$\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) := \bar{I}(\mathbf{S}^{(M)}; \mathbf{X}^{(M)}) - \bar{I}(\mathbf{S}^{(M)}; \mathbf{Y}) \quad (11)$$

and measures how much of the information $\mathbf{X}$ conveys about $\mathbf{S}$ is lost in each time step due to downsampling.

While in the general case the filter which minimizes $\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y})$ is hard to find, for this particular signal model the solution is surprisingly intuitive:

Theorem 2. *Let* $\mathbf{S}$ *and* $\mathbf{N}$ *be jointly stationary Gaussian processes with smooth PSDs* $S_S(\mathrm{e}^{\jmath\theta})$ *and* $S_N(\mathrm{e}^{\jmath\theta})$ *and which satisfy Assumption 1. Let* $X_n = S_n + N_n$*. Then, the* M*-aliasing-free energy compaction filter for* $S_S(\mathrm{e}^{\jmath\theta})/S_N(\mathrm{e}^{\jmath\theta})$ *minimizes the relevant information loss rate in the decimation system depicted in Fig. 1.*

The energy compaction filter for a given PSD can be constructed easily: The M-fold downsampled PSD consists of M aliased components; for each frequency point $\theta \in [-\pi/M, \pi/M]$, at least one of them is maximal. The passbands of the energy compaction filter correspond to exactly these maximal components [2], [3]. See also (12) below.

In particular, since for white Gaussian noise $\mathbf{N}$ the energy compaction filter for $S_S(\mathrm{e}^{\jmath\theta})/S_N(\mathrm{e}^{\jmath\theta})$ coincides with the energy compaction filter for $S_S(\mathrm{e}^{\jmath\theta})$, the filter that lets most of the signal's energy pass aliasing-free is also optimal in terms of information.

Sketch of the proof: Instead of minimizing $\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y})$ we maximize

$$\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n} \left(h(Y_1^n) - h(Y_1^n | \tilde{S}_1^{nM}) \right).$$

with $\tilde{\mathbf{S}}$ being $\mathbf{S}$ filtered by H. But $h(Y_1^n) = h(\tilde{X}_M, \dots, \tilde{X}_{nM})$ and $h(Y_1^n|\tilde{S}_1^{nM}) = h(\tilde{N}_M, \dots, \tilde{N}_{nM})$, where $\tilde{\mathbf{N}}$ is the noise process filtered by H. By Gaussianity, the mutual information rate reads

$$\bar{I}(\tilde{\mathbf{S}}^{(M)}; \mathbf{Y}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln \left(1 + \frac{\sum_{k=0}^{M-1} \frac{S_S(\mathrm{e}^{\jmath\theta_k})}{S_N(\mathrm{e}^{\jmath\theta_k})} H_k(\theta)}{\sum_{k=0}^{M-1} H_k(\theta)} \right) d\theta$$

where $\theta_k := \frac{\theta - 2k\pi}{M}$ and $H_k(\theta) := S_N(\mathrm{e}^{\jmath\theta_k})|H(\mathrm{e}^{\jmath\theta_k})|^2$. Maximizing the integral is done by maximizing the fraction inside the logarithm, which is, essentially, a weighted average of the ratios $S_S(\mathrm{e}^{\jmath\theta_k})/S_N(\mathrm{e}^{\jmath\theta_k})$. The maximum is obtained if

$$H_l(\theta) = \begin{cases} 1, & \text{for smallest } l \text{ s.t. } \forall k: \frac{S_S(\mathrm{e}^{\jmath\theta_l})}{S_N(\mathrm{e}^{\jmath\theta_l})} \geq \frac{S_S(\mathrm{e}^{\jmath\theta_k})}{S_N(\mathrm{e}^{\jmath\theta_k})} \\ 0, & \text{else} \end{cases} \quad (12)$$

i.e., if H is related to the energy compaction filter via $H_k(\theta) = S_N(\mathrm{e}^{\jmath\theta_k})|H(\mathrm{e}^{\jmath\theta_k})|^2$. Since $|\bar{h}(\mathbf{N})| < \infty$, a filter H' with $|H'(\mathrm{e}^{\jmath\theta})|^2 = 1/S_N(\mathrm{e}^{\jmath\theta})$ does not change the information content; thus, H can be chosen as the energy compaction filter. ■

V. Relevant Information Loss: General Case

Although the result for Gaussian processes is interesting due to its closed form, it is of little practical relevance. In many cases, at least the relevant part of $\mathbf{X}$, the signal process $\mathbf{S}$, will be non-Gaussian. We thus drop the restriction that $\mathbf{S}$ is Gaussian, but we still assume Gaussianity of $\mathbf{N}$.

One can expect that in this more general case a closed-form solution for H will not be available. However, *assuming* that $\mathbf{S}$ is Gaussian, yields an upper bound on the information rate $\bar{I}(\mathbf{S}^{(M)}; \mathbf{Y})$. It can also be shown that the Gaussian assumption provides an upper bound on the relevant information loss rate. To this end, we employ the approach of Plumbley [8], who showed that, with a specific signal model, PCA can be justified from an information-theoretic perspective (cf. also [14]).

Theorem 3. *Let* H *be stable and causal, let* $\mathbf{S}$ *and* $\mathbf{N}$ *be jointly stationary and satisfy Assumption 1, and let* $X_n = S_n + N_n$*.* $\mathbf{N}$ *is Gaussian, and* $\mathbf{S}_G$ *is Gaussian with the same PSD as* $\mathbf{S}$*. Let* $X_{G,n} = S_{G,n} + N_n$*, and let* $\mathbf{Y}_G$ *be the corresponding output processes of the decimation system, respectively. Then,*

$$\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) \leq \bar{L}_{\mathbf{S}_G^{(M)}}(\mathbf{X}_G^{(M)} \to \mathbf{Y}_G). \quad (13)$$

Sketch of the proof: We start from

$$\begin{aligned} &\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y}) \\ &= \lim_{n\to\infty} \frac{1}{n} \left(h(\tilde{X}_1^{nM}) - h(\tilde{N}_1^{nM}) - h(Y_1^n) + h(Y_1^n|\tilde{S}_1^{nM}) \right) \end{aligned}$$

where we exploited the fact that the filter does not change the information content of the process. We then apply $h(Y_1^n|\tilde{S}_1^{nM}) = h(\tilde{N}_M, \dots, \tilde{N}_{nM})$ and $h(Y_1^n) = h(\tilde{X}_M, \dots, \tilde{X}_{nM})$. Then, the only term depending on non-Gaussian RVs is the conditional differential entropy

$$h(\tilde{X}_1^{M-1}, \dots, \tilde{X}_{(n-1)M+1}^{nM-1} | \tilde{X}_M, \dots, \tilde{X}_{nM})$$

which is positive in above equation for the relevant information loss rate. This differential entropy is upper bounded by the one of Gaussian RVs $(X_G)_1^{nM}$ with the same first and second moments. The bound is achieved by replacing $\mathbf{S}$ by $\mathbf{S}_G$. ■

A consequence of this theorem is that filter design by energetic considerations, i.e., by considering the PSDs of the signal only, has performance guarantees also in information-theoretic terms. One has to consider, though, that the filter H

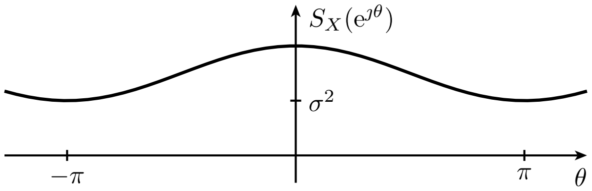


Fig. 2. Power spectral density of $\mathbf{X}$

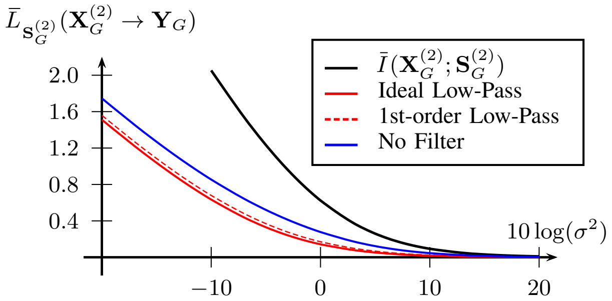


Fig. 3. Upper bounds on the relevant information loss rate in nats as a function of the noise variance σ^2 for various filter options.

optimal in the sense of the upper bound might not coincide with the filter optimal w.r.t. $\bar{L}_{\mathbf{S}^{(M)}}(\mathbf{X}^{(M)} \to \mathbf{Y})$.

VI. Example

We now illustrate our results with an example: Let the PSD of $\mathbf{S}$ be given by $S_S(e^{\jmath\theta}) = 1 + \cos\theta$ and let $\mathbf{N}$ be independent white Gaussian noise with variance σ^2, i.e., $S_N(e^{\jmath\theta}) = \sigma^2$. The PSD of $\mathbf{X}$ is depicted in Fig. 2. We consider downsampling by a factor of $M = 2$. Were $\mathbf{S}$ Gaussian too, the optimal filter would be an ideal low-pass filter with cut-off frequency $\pi/2$ (cf. Theorem 2).

If we assume that $\mathbf{S}$ is non-Gaussian, Theorem 3 allows us to design a finite-order filter which minimizes an upper bound on the relevant information loss rate. In particular, it can be shown that among all first-order FIR filters with impulse response $h[n] = \delta[n] + c\delta[n-1]$, the filter with $c = 1$ minimizes the Gaussian bound.

Fig. 3 shows the upper bound on the relevant information loss rate as a function of the noise variance σ^2 for the ideal low-pass filter and the optimal first-order FIR filter compared to the case where no filter is used. In addition, the available information $\bar{I}(\mathbf{X}_G^{(2)}; \mathbf{S}_G^{(2)}) = 2\bar{I}(\mathbf{X}_G; \mathbf{S}_G)$ is plotted, which decreases with increasing noise variance. Indeed, filtering can reduce the relevant information loss rate compared to omitting the filter. This is in stark contrast with the results of Section III, in which we showed that the relative information loss rate equals $1/2$ regardless of the filter. The reason is that in Section III we did not have a signal model in mind, treating every bit of information equally. As soon as one knows which aspect of a stochastic process is relevant, one can successfully apply signal processing methods to retrieve as much information as possible (or to remove as much of the irrelevant information as possible, cf. [14]).

Interestingly, as Fig. 3 shows, the improvement of a first-order FIR filter over direct downsampling is significant. Using low-order filters is beneficial also from a computational perspective: To the best of our knowledge, the optimization problem does not permit a closed-form solution for the filter coefficients in general. Thus, numerical procedures will benefit from the fact that the number of coefficients can be kept small. Moreover, while the optimal first-order FIR filter is independent of the noise variance σ^2, numerical calculations suggest that the optimal second-order FIR with impulse response $h[n] = \delta[n] + c\delta[n-1] + \delta[n-2]$ has a coefficient c depending on σ^2.

VII. Conclusion

In this work we analyzed the information loss in a decimation system as a function of its constituting anti-aliasing filter. In particular, we showed that without a signal model in mind, anti-aliasing filtering is futile since it cannot reduce the information loss even if ideal filters are permitted. The situation changes for a simple signal-plus-Gaussian-noise model, where the information loss w.r.t. the signal process can be reduced by properly choosing the filter. As a direct consequence, we concluded that filter design based on second-order statistics of the process can be justified from an information-theoretic perspective.

References

[1] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Pearson Higher Ed., 2010.

[2] M. Unser, "On the optimality of ideal filters for pyramid and wavelet signal approximation," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3591–3596, Dec. 1993.

[3] M. Tsatsanis and G. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1766–1777, Aug. 1995.

[4] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, ser. Information Science and Statistics. New York, NY: Springer, 2010.

[5] J. I. Galdos and D. E. Gustafson, "Information and distortion in reduced-order filter design," *IEEE Trans. Inf. Theory*, vol. 23, no. 2, pp. 183–194, Mar. 1977.

[6] N. Al-Dhahir and J. M. Cioffi, "Block transmission over dispersive channels: transmit filter optimization and realization, and MMSE-DFE receiver performance," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 137–160, Jan. 1996.

[7] A. Scaglione, S. Barbarossa, and G. B. Giannakis, "Filterbank transceivers optimizing information rate in block transmissions over dispersive channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 1019–1032, Apr. 1999.

[8] M. Plumbley, "Information theory and unsupervised neural networks," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR. 78, 1991.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley Interscience, 2006.

[10] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 193–215, Mar. 1959.

[11] R. M. Gray, *Entropy and Information Theory*. New York, NY: Springer, 1990.

[12] B. C. Geiger and G. Kubin, "Relative information loss in the PCA," in *Proc. IEEE Information Theory Workshop (ITW)*, Lausanne, Sep. 2012, pp. 562–566, extended version available: `arXiv:1204.0429 [cs.IT]`.

[13] ——, "On the rate of information loss in memoryless systems," Apr. 2013, `arXiv:1304.5057 [cs.IT]`.

[14] ——, "Signal enhancement as minimization of relevant information loss," in *Proc. ITG Conf. on Systems, Communication and Coding*, Munich, Jan. 2013, pp. 1–6, extended version available: `arXiv:1205.6935 [cs.IT]`.

Constrained Entropy Maximisation

Terence H. Chan and Alex Grant,
Institute for Telecommunications Research,
University of South Australia

***Abstract*—A fundamental problem in designing distributed storage networks is to determine the optimal tradeoffs among various design parameters, including storage cost, repair cost, and reliability. Such a problem can be formulated as an entropy maximisation problem subject to functional a set of dependency constraints. In fact, many problems in network coding and error correcting codes can also be formulated as the same entropy maximisation problem.**

Unfortunately, solving such an optimisation problem can be extremely difficult in general. To reduce the complexity, various relaxations have been considered, which are based on techniques in association schemes, information inequalities, and functional dependency bounds (a generalisation of cut-set bounds). This paper compares these relaxations and showed that both linear programming bounds (derived from association schemes and information inequalities) are at least better than the functional dependency bounds.

Author Index