# International Zurich Seminar on Communications - Proceedings

# International Zurich Seminar on Communications

March 2 – 4, 2016

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.


# Conference Organization

### General Co-Chairs

Amos Lapidoth and Stefan M. Moser

### Technical Program Committee

Ezio Biglieri
Helmut Bölcskei
Terence H. Chan
Alexandros G. Dimakis
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Deniz Gunduz
Martin Hänggi
Franz Hlawatsch
Johannes Huber
Ashish Khisti
Tobias Koch
Gerhard Kramer
Frank Kschischang
Hans-Andrea Loeliger
Thomas Mittelholzer
Bernhard Plattner
Ron Roth
Igal Sason
Jossy Sayir
Robert Schober
Giorgio Taricco
Emre Telatar
Emanuele Viterbo
Michèle Wigger
Armin Wittneben
Ram Zamir

### Organizers of Invited Sessions

Joseph Boutros
Giuseppe Durisi
Albert Guillén i Fàbregas
Ashish Khisti
Anelia Somekh-Baruch

### Local Organization

**Conference Secretary**
Silvia Tempel

**Web and Publications**
Michael Lerjen

# Table of Contents

## Keynote Talks



## Session 1 — Wed 10:00 – 12:00
## Topics in Shannon Theory

Chaired by Yuval Kochman (Hebrew University of Jerusalem)



---

*Invited papers are marked by an asterisk.

## Session 2 — Wed 13:20 – 15:00

## Codes on Graphs and Lattices

Invited session organizer: Joseph J. Boutros (Texas A & M University at Qatar)



## Session 3 — Wed 15:30 – 16:50

## Wireless Communication (with emphasis on Finite Blocklength Information Theory)

Invited session organizer: Giuseppe Durisi (Chalmers University of Technology)

## Session 4 — Wed 17:00 – 18:00

## Privacy and Secrecy

Chaired by Roy Timo (Technische Universität München)



## Session 5 — Thu 10:00 – 12:00

## Communication Techniques

Chaired by Richard Wesel (University of California, Los Angeles)

## Session 6 Thu 13:20 – 15:00

## Information Theoretic Security

Invited session organizer: Ashish Khisti (University of Toronto)



## Session 7 Thu 15:30 – 17:30

## Network Information Theory

Chaired by Elza Erkip (New York University)

## Session 8 — Fri 10:00 – 12:00

## Coding Techniques

Chaired by Erik Agrell (Chalmers University of Technology)



## Session 9 — Fri 13:20 – 15:00

## Shannon Theory

Invited session organizer: Anelia Somekh-Baruch (Bar-Ilan University)

## Session 10 Fri 15:30 – 16:50

## Hypothesis Testing and Converse Bounds

Invited session organizer: Albert Guillén i Fàbregas (Universitat Pompeu Fabra)

# Channel Detection in Coded Communication

Nir Weinberger and Neri Merhav
Dept. of Electrical Engineering, Technion - Israel Institute of Technology
{nirwein@campus, merhav@ee}.technion.ac.il

***Abstract*—We consider the problem of block-coded communication, where in each block, the channel law belongs to one of two disjoint sets. The decoder is aimed to decode only messages that have undergone a channel from one of the sets, and thus has to detect the set which contains the underlying channel. We begin with the simplified case where each of the sets is a singleton. For any given code, we present the optimum detection/decoding rule in the sense of the best trade-off among the probabilities of decoding error, false alarm, and misdetection. Then, we derive the exact single-letter characterization of the random coding exponents for the optimal detector/decoder, as well as an expurgated bound for low rates. We then extend the random coding analysis to general sets of channels, and derive the optimal detector/decoder under a worst case formulation of the error probabilities, and derive its random coding exponents.**

## I. Introduction

Consider communicating over a channel with input $X$ and output $Y$, for which the underlying channel law $P_{Y|X}$ is supposed to belong to a family of channels $\mathcal{W}$, but the receiver would also like to examine an alternative hypothesis, in which the channel $P_{Y|X}$ actually belongs to a different set $\mathcal{V}$, disjoint from $\mathcal{W}$. Examples for which such a detection procedure is useful are: (i) Detection of an abrupt change in the channel statistics, e.g., a deep fading event in wireless communication [1], the presence of excessively large number of interferers, or loss of tracking in adaptive equalization or timing recovery algorithms [2]. (ii) Detection of an imposter, which transmits the same codewords as the authorized transmitter, but via a significantly different channel. The distinctive channel statistics of the authorized transmitter could be used to identify such intrusions, which is crucial, if, e.g., the messages are used to control a sensitive equipment at the receiver side, (iii) detection of the active user in a sparse multiple access channel (with no collisions between the two users). Even if both users use the exact same codebook (as might be dictated by practical considerations), and even if no header is used to identify each of the users, the receiver could still identify the sender with high reliability, utilizing the different channel of each user. Thus, beyond the ordinary task of decoding the message, the receiver would also like to perform *hypothesis testing* between the null hypothesis $P_{Y|X} \in \mathcal{W}$ and the alternative hypothesis $P_{Y|X} \in \mathcal{V}$. For example, if the channel quality is gauged by a single parameter, say, the crossover probability of a binary symmetric channel (BSC), then $\mathcal{W}$ and $\mathcal{V}$ could be two disjoint intervals of this parameter.

This problem of joint detection/decoding belongs to a larger class of hypothesis testing problems, in which after performing the test, another task should be performed, depending on the chosen hypothesis, e.g. Bayesian estimation [3], and lossless source coding [4]. More recently [5], we have studied the related problem of joint detection and decoding for sparse communication [6], which is motivated by strongly asynchronous channels [7]. In these channels the transmitter is either completely silent or transmits a codeword from a given codebook. The task of the detector/decoder is to decide whether transmission has taken place, and if so, to decode the message. The performance is judged by: (i) the *false alarm* (FA) probability - deciding on a message when the transmitter was silent, (ii) the *misdetection* (MD) probability - deciding that the transmitter was silent when it transmitted some message, and (iii) the probability of *inclusive error* (IE) - namely, not deciding on the correct message sent (either misdetection or erroneous decoding). We have then found the optimum detector/decoder that minimizes the IE probability subject to given constraints on the FA and the MD probabilities for a given codebook, and also provided single-letter expressions for the exact random coding exponents. While this is a *joint* detector/decoder, we have also observed that an *asymptotic separation principle* holds, in the following sense: A detector/decoder which achieves the optimal exponents may be comprised of an optimal detector in the Neyman-Pearson sense for the FA and MD probabilities, followed by ordinary maximum likelihood (ML) decoding.

In case of two simple hypotheses, $\mathcal{W} = \{W\}$ and $\mathcal{V} = \{V\}$, the problem of [5] is a special case of the problem studied here, for which the the output of the channel $V$ is completely independent of its input, and plays the role of noise. It turns out that the optimal detector/decoder and its properties for the problem studied here are straightforward generalizations of [5]. However, the analysis of the random coding detection exponents is much more intricate here than in [5]. The detector in [5] compares a likelihood which depends on the codebook with a likelihood that depends on the noise. So, when analyzing the performance of random coding, the random choice of codebook only affects the distribution of the likelihood of the 'codebook hypothesis'. By contrast, here, since we would like to detect the channel, the random choice of codebook affects the likelihood of *both* hypotheses, and consequently, they may be highly dependent.

In this paper, we study the problem of joint channel detection between two disjoint sets of memoryless channels $\mathcal{W}, \mathcal{V}$, and decoding. We begin by considering the case of simple hypotheses, namely $\mathcal{W} = \{W\}$ and $\mathcal{V} = \{V\}$ (Section II). As in [5], we derive the detector/decoder which achieves the optimal trade-off between the FA, MD and IE probabilities,

where here too, an asymptotic separation principle holds (Section III). Then, we derive single-letter expressions for the *exact* random coding exponents, as well as improved (expurgated) bound for low rates (Section IV). Afterwards, we discuss a generalization to composite hypotheses, i.e., $\mathcal{W}$, $\mathcal{V}$ that are not singletons (Section V), and finally, we discuss the archetype example for which $(W, V)$ are a pair BSCs (Section VI). Due to the space limitation, some details and full proofs are omitted, but can be found in [8].

## II. Problem Formulation

We begin with the following notation conventions. Alphabets and other sets will be denoted by calligraphic letters, e.g. $\mathcal{X}$. Random variables and vectors will be denoted by capital letters, e.g. $\mathbf{X} = (X_1, \ldots, X_n)$, ($n$ - positive integer), and specific values for them by lower case letters, $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the $n$-th order Cartesian power of $\mathcal{X}$. A joint distribution of a pair of random variables $(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$, the Cartesian product alphabet of $\mathcal{X}$ and $\mathcal{Y}$, will be denoted by $Q_{XY}$ and similar forms, e.g. $\tilde{Q}_{XY}$. We will usually abbreviate this notation by omitting the subscript $XY$, and denote, e.g, $Q_{XY}$ by $Q$. The $X$-marginal ($Y$-marginal), induced by $Q$ will be denoted by $Q_X$ (respectively, $Q_Y$), and the conditional distributions will be denoted by $Q_{Y|X}$ and $Q_{X|Y}$. The joint distribution induced by $Q_X$ and $Q_{Y|X}$ will be denoted by $Q = Q_X \times Q_{Y|X}$. The mutual information of a joint distribution $Q$ will be denoted by $I(Q)$. The conditional information divergence between the conditional distributions $Q_{Y|X}$ and $P_{Y|X}$, averaged over $Q_X$, will be denoted by $D(Q_{Y|X} \| P_{Y|X} | Q_X)$. The probability of an event $\mathcal{A}$ will be denoted by $\mathbb{P}\{\mathcal{A}\}$, and the expectation operator will be denoted by $\mathbb{E}\{\cdot\}$. The complement of a set $\mathcal{A}$ will be denoted by $\mathcal{A}^c$. Logarithms and exponents will be understood to be taken to the natural base, and we will denote $[t]_+ \triangleq \max\{t, 0\}$. We adopt the standard convention that when a minimization (respectively, maximization) problem is performed on an empty set the result is $\infty$ (respectively, $-\infty$).

Consider a discrete memoryless channel, characterized by a finite input alphabet $\mathcal{X}$, a finite output alphabet $\mathcal{Y}$, and a given matrix of single-letter transition probabilities $\{P_{Y|X}(y|x)\}_{x\in\mathcal{X}, y\in\mathcal{Y}}$. Let $\mathcal{C}_n = \{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_M\} \subset \mathcal{X}^n$, denote a codebook for blocklength $n$ and rate $R$, for which the transmitted codeword is chosen with a uniform probability distribution over the $M = \lceil e^{nR} \rceil$ codewords. The conditional distribution $P_{Y|X}$ may either satisfy $P_{Y|X} = W$ (the *null hypothesis*), or $P_{Y|X} = V$ (the *alternative hypothesis*). It is required to design a detector/decoder which is oriented to decode messages only arriving via the channel $W$. Formally, such a detector/decoder $\phi$ is a partition of $\mathcal{Y}^n$ into $M + 1$ regions, denoted by $\{\mathcal{R}_m\}_{m=0}^M$. If $\mathbf{y} \in \mathcal{R}_m$ for some $1 \le m \le M$ then the $m$-th message is decoded. If $\mathbf{y} \in \mathcal{R}_0$ (the *rejection region*) then the channel $V$ is identified, and no decoding takes place.

For a codebook $\mathcal{C}_n$ and a detector/decoder $\phi$, we define[1]

$$P_{\text{FA}}(\mathcal{C}_n, \phi) \triangleq \frac{1}{M} \sum_{m=1}^{M} W(\mathcal{R}_0 | \mathbf{x}_m), \tag{1}$$

$$P_{\text{MD}}(\mathcal{C}_n, \phi) \triangleq \frac{1}{M} \sum_{m=1}^{M} V(\mathcal{R}_0^c | \mathbf{x}_m), \tag{2}$$

$$P_{\text{IE}}(\mathcal{C}_n, \phi) \triangleq \frac{1}{M} \sum_{m=1}^{M} W\left(\mathcal{R}_m^c | \mathbf{x}_m\right) \tag{3}$$

as the *false alarm* (FA) probability, *misdetection* (MD) probability, and *inclusive error* (IE) probability, respectively. Thus, the IE event is the total error event, i.e., when the correct codeword is not decoded either because of a FA or an ordinary erroneous decoding. When possible, we will omit the notation of the dependence of these probabilities on $\mathcal{C}_n$ and $\phi$.

For a given code $\mathcal{C}_n$, we are interested in achievable tradeoffs between $P_{\text{FA}}$, $P_{\text{MD}}$ and $P_{\text{IE}}$. Consider the following problem:

$$\begin{aligned} &\text{minimize} \quad P_{\text{IE}} \\ &\text{subject to} \quad P_{\text{FA}} \le \epsilon_{\text{FA}},\ P_{\text{MD}} \le \epsilon_{\text{MD}} \end{aligned} \tag{4}$$

where $\epsilon_{\text{FA}}$ and $\epsilon_{\text{MD}}$ are given prescribed quantities, and it is assumed that these two constraints are not contradictory. Indeed, there is some tension between $P_{\text{MD}}$ and $P_{\text{FA}}$ as they are related via the Neyman-Pearson lemma [9, Theorem 11.7.1]. For a given $\epsilon_{\text{FA}}$, the minimum achievable $P_{\text{MD}}$ is positive, in general. It is assumed then that the prescribed value of $\epsilon_{\text{MD}}$ is not smaller than this minimum. In the problem under consideration, it makes sense to relax the tension between the two constraints to a certain extent, in order to allow some freedom to minimize $P_{\text{IE}}$ under these constraints. While this is true for any *finite* blocklength, as we shall see (Proposition 2), an *asymptotic separation principle* holds, and the optimal detector in terms of exponents has full tension between the FA and MD exponents. The optimal detector/decoder for the problem (4) will be denoted by $\phi^*$. Our goal is to find the optimum detector/decoder for the problem (4), and then analyze the achievable exponents associated with the resulting error probabilities.

## III. The Optimal Joint Detector/Decoder

Let $a, b \in \mathbb{R}$, and define the detector/decoder $\phi^* = \{\mathcal{R}_m^*\}_{m=0}^M$, where $\mathcal{R}_0^*$ is defined as

$$\left\{ \mathbf{y} :\ a \sum_{m=1}^{M} W(\mathbf{y}|\mathbf{x}_m) + \max_m W(\mathbf{y}|\mathbf{x}_m) \le b \sum_{m=1}^{M} V(\mathbf{y}|\mathbf{x}_m) \right\} \tag{5}$$

and

$$\mathcal{R}_m^* \triangleq [\mathcal{R}_0^*]^c \bigcap \left\{ \mathbf{y} :\ W(\mathbf{y}|\mathbf{x}_m) > \max_{k \ne m} W(\mathbf{y}|\mathbf{x}_k) \right\}. \tag{6}$$

**Lemma 1.** *Let a codebook $\mathcal{C}_n$ be given, let $\phi^*$ be as above, and let $\phi$ be any other partition of $\mathcal{Y}^n$ into $M+1$ regions. If $P_{\text{FA}}(\mathcal{C}_n, \phi) \le P_{\text{FA}}(\mathcal{C}_n, \phi^*)$ and $P_{\text{MD}}(\mathcal{C}_n, \phi) \le P_{\text{MD}}(\mathcal{C}_n, \phi^*)$ then $P_{\text{IE}}(\mathcal{C}_n, \phi) \ge P_{\text{IE}}(\mathcal{C}_n, \phi^*)$.*

Note that this detector/decoder is optimal (in the Neyman-Pearson sense) for any *given* blocklength $n$ and codebook

[1] The meaning of FA and MD here is opposite to their respective meaning in [5], as sanctioned by the motivating applications.

$\mathcal{C}_n$. Thus, upon a suitable choice of $a$ and $b$, it solves the problem (4) *exactly*. As common, to assess the achievable performance, we resort to large blocklength analysis of error exponents. For a given sequence of codes $\mathcal{C} \triangleq \{\mathcal{C}_n\}_{n=1}^{\infty}$ and a detector/decoder $\phi$, the FA exponent is defined as

$$E_{\text{FA}}(\mathcal{C}, \phi) \triangleq \liminf_{n\to\infty} -\frac{1}{n} \log P_{\text{FA}}(\mathcal{C}_n, \phi), \tag{7}$$

and the MD exponent $E_{\text{MD}}(\mathcal{C}, \phi)$ and the IE exponent $E_{\text{IE}}(\mathcal{C}, \phi)$ are defined similarly. The asymptotic version of (4) is then stated as finding the detector/decoder which achieves the largest $E_{\text{IE}}$ under constraints on $E_{\text{FA}}$ and $E_{\text{MD}}$. To affect these error exponents, the coefficients $a, b$ in (5) need to exponentially increase/decrease as a functions of $n$, and so we set $a \triangleq e^{n\alpha}$ and $b \triangleq e^{n\beta}$. Evidently, for $\alpha > 0$, the ML term on the right-hand side (r.h.s.) of (5) is negligible w.r.t. the left-hand side (l.h.s.), and the obtained rejection region is asymptotically equivalent to

$$\mathcal{R}_0' \triangleq \left\{ \mathbf{y} : \; e^{n\alpha} \sum_{m=1}^{M} W(\mathbf{y}|\mathbf{x}_m) \le e^{n\beta} \sum_{m=1}^{M} V(\mathbf{y}|\mathbf{x}_m) \right\} \tag{8}$$

which corresponds to an ordinary Neyman-Pearson test between the hypotheses that the channel is $W$ or $V$. Thus, unlike the fixed blocklength case, asymptotically, we obtain a complete tension between the FA and MD probabilities. The resulting detector (8) is indeed quite intuitive: It compares the likelihoods of each of the channels, assuming nothing on the transmitted message. The 'max' term in the optimal detector (5), is, however, more difficult to intuitively explain, but it naturally arises in the proof of Lemma 1.

Consequently, as the next proposition shows, there is no loss in error exponents when using the $\phi'$, whose rejection region is as in (8), and if $\mathbf{y} \notin \mathcal{R}_0'$ then ordinary ML decoding for $W$ is used, as in (6). This implies an *asymptotic separation principle* between detection and decoding: the optimal detector can be used without considering the subsequent decoding, and the optimal decoder can be used without considering the preceding detection. As a result, asymptotically, there is only a single degree of freedom to control the exponents. Thus, when analyzing error exponents in Section IV, we will assume that $\phi'$ is used, and since (8) depends on the difference $\alpha - \beta$ only, we will set henceforth $\beta = 0$ for $\phi'$. The parameter $\alpha$ will be used to control the trade-off between the FA and MD exponents, just as in ordinary hypothesis testing.

**Proposition 2.** *For any given sequence of codes $\mathcal{C} = \{\mathcal{C}_n\}_{n=1}^{\infty}$, and given constraints on the FA and MD exponents, the detector/decoder $\phi'$ achieves the same IE exponent as $\phi^*$.*

Proposition 2 is proved by noticing that conditioned on the $m$th codeword, the IE probability (3) is the union of the FA event and the event $\{W(\mathbf{Y}|\mathbf{x}_m) < \max_{k\neq m} W(\mathbf{Y}|\mathbf{x}_k)\}$, i.e., an ordinary ML decoding error. So, as the union bound is asymptotically exponentially *tight* for a union of two events

$$E_{\text{IE}}(\mathcal{C}, \phi^*) = \min\left\{E_{\text{o}}(\mathcal{C}, \phi^*), E_{\text{FA}}(\mathcal{C}, \phi^*)\right\}, \tag{9}$$

which implies that the best IE exponent is obtained for the best FA exponent.

The achievable exponent bounds will be proved by random coding over some ensemble of codes. Letting over-bar denote an average w.r.t. some ensemble, we will define the random coding exponents, as

$$E_{\text{FA}}(\phi) \triangleq \lim_{l\to\infty} -\frac{1}{n_l} \log \overline{P_{\text{FA}}}(\mathcal{C}_{n_l}, \phi), \tag{10}$$

where $\{n_l\}_{l=1}^{\infty}$ is a sub-sequence of blocklengths. When we assume a fixed composition ensemble with distribution $P_X$, this sub-sequence will simply be the blocklengths such that the type class associated with $P_X$ is not empty. To comply with definition (7), one can obtain codes which are good for *all* sufficiently large blocklength by slightly modifying the input distribution. The MD exponent $E_{\text{MD}}(\phi)$ and the IE exponent $E_{\text{IE}}(\phi)$ are defined similarly, where the three exponents share the *same* sequence of blocklengths.

Now, if we provide random coding exponents for the FA, MD and ordinary decoding exponents, then the existence of a good sequence of codes can be easily shown. Indeed, the Markov inequality implies that

$$\mathbb{P}\left(P_{\text{FA}}(\mathcal{C}_{n_l}, \phi) \ge \exp\left[-n_l\left(E_{\text{FA}}(\phi) - \delta\right)\right]\right) \le e^{-n_l \frac{\delta}{2}}, \tag{11}$$

for all $l$ sufficiently large. Thus, with probability tending to 1, the chosen codebook will have FA probability not larger than $\exp\left[-n\left(E_{\text{FA}}(\phi) - \delta\right)\right]$. As the same can be said on the MD probability and the ordinary error probability, then one can find a sequence of codebooks with simultaneously good FA, MD and ordinary decoding error probabilities, and from (9), also good IE probability. For this reason, henceforth we will only focus on the FA and MD exponents. The IE exponent can be simply obtained by (9) and the known bounds of ordinary decoding, namely, the random coding bounds [10, Theorem 10.2] (and its tightness [10, Problem 10.34]) and the expurgated bound [10, Problem 10.18].

## IV. Achievable Error Exponents

Let $\tilde{Q}$ represent the joint type of the true transmitted codeword and the output, and $\overline{Q}$ is some type of a competing codeword. We denote the *normalized log-likelihood ratio* of a channel $W$ by

$$f_W(Q) \triangleq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} Q(x,y) \log W(y|x), \tag{12}$$

with the convention $f_W(Q) = -\infty$ if $Q(x,y) > 0 \Rightarrow W(y|x) = 0$. We define the set

$$\mathcal{Q}_W \triangleq \{Q : \; f_W(Q) > -\infty\} \tag{13}$$

and for $\gamma \in \mathbb{R}$,

$$\mathbf{s}(\tilde{Q}_Y, \gamma) \triangleq \min_{Q\in\mathcal{Q}_W : \, Q_Y = \tilde{Q}_Y} I(Q) + \left[-\alpha - f_W(Q) + \gamma\right]_+. \tag{14}$$

Further, define the sets

$$\mathcal{J}_1 \triangleq \left\{\tilde{Q} : \; f_W(\tilde{Q}) \le -\alpha + f_V(\tilde{Q})\right\}, \tag{15}$$

$$\mathcal{J}_2 \triangleq \left\{\tilde{Q}:\ \mathbf{s}\left(\tilde{Q}_Y, f_V(\tilde{Q})\right) \geq R\right\}, \tag{16}$$

the exponent

$$E_A \triangleq \min_{\tilde{Q}\in\cap_{i=1}^2 \mathcal{J}_i} D(\tilde{Q}_{Y|X}\|W|P_X), \tag{17}$$

the sets

$$\mathcal{K}_1 \triangleq \left\{(\tilde{Q},\overline{Q}):\ \overline{Q}_Y = \tilde{Q}_Y\right\}, \tag{18}$$

$$\mathcal{K}_2 \triangleq \left\{(\tilde{Q},\overline{Q}):\ f_W(\overline{Q}) \leq -\alpha + f_V(\overline{Q})\right\}, \tag{19}$$

$$\mathcal{K}_3 \triangleq \left\{(\tilde{Q},\overline{Q}):\ f_V(\overline{Q}) \geq \alpha + f_W(\tilde{Q}) - \left[R - I(\overline{Q})\right]_+\right\}, \tag{20}$$

$$\mathcal{K}_4 \triangleq \left\{(\tilde{Q},\overline{Q}):\ \mathbf{s}\left(\tilde{Q}_Y, f_V(\overline{Q}) + \left[R - I(\overline{Q})\right]_+\right) \geq R\right\}, \tag{21}$$

and the exponent

$$E_B \triangleq \min_{(\tilde{Q},\overline{Q})\in\cap_{i=1}^4 \mathcal{K}_i} \left\{D(\tilde{Q}_{Y|X}\|W|P_X) + \left[I(\overline{Q}) - R\right]_+\right\}. \tag{22}$$

In addition, let us define the *type-enumeration detection random coding exponent* as

$$E_{\text{TE}}^{\text{RC}}(R,\alpha,P_X,W,V) \triangleq \min\{E_A, E_B\}. \tag{23}$$

**Theorem 3.** *Let $P_X$ and$\alpha \in \mathbb{R}$ be given. Then, there exists a sequence of codes $\mathcal{C} = \{\mathcal{C}_n\}_{n=1}^{\infty}$ of rate $R$ such that for any $\delta > 0$*

$$E_{\text{FA}}(\mathcal{C},\phi^*) \geq E_{\text{TE}}^{\text{RC}}(R,\alpha,P_X,W,V) - \delta, \tag{24}$$

$$E_{\text{MD}}(\mathcal{C},\phi^*) \geq E_{\text{TE}}^{\text{RC}}(R,\alpha,P_X,W,V) - \alpha - \delta. \tag{25}$$

The proof of Theorem 3 relies on the *type-enumeration method*, and omitted due to space limitations. For the FA exponent, the exponent $E_A$ ($E_B$, respectively) pertains to the event that the scaled likelihood of the true codeword (sum of scaled likelihoods of all competing codewords, respectively) under $V$ is larger than $\sum_{m=1}^M W(\mathbf{Y}|\mathbf{X}_m)$. The MD exponent is obtained by proving that the difference between the FA and MD exponents of $\phi'$ is always $\alpha$. The main challenge in analyzing the random coding FA exponent, is that the *likelihoods* of both hypotheses, namely $\sum_{m=1}^M W(\mathbf{Y}|\mathbf{X}_m)$ and $\sum_{m=1}^M V(\mathbf{Y}|\mathbf{X}_m)$ are very correlated due to the fact the once the codewords are drawn, they are common for both likelihoods. This is significantly different from the situation in [5], in which the likelihood $\sum_{m=1}^M W(\mathbf{Y}|\mathbf{X}_m)$ was compared to the noise likelihood $Q_0(\mathbf{Y})$, of a completely different distribution.

We next turn to derive expurgated exponents. Throughout, $P_{X\tilde{X}}$ will represent a joint type of a pair of codewords. Let us define

$$d_s(x,\tilde{x}) \triangleq -\log[\sum_{y\in\mathcal{Y}} W^{1-s}(y|x)V^s(y|\tilde{x})] \tag{26}$$

and the set

$$\mathcal{L} \triangleq \{P_{X\tilde{X}}:\ P_{\tilde{X}} = P_X,\ I(P_{X\tilde{X}}) \leq R\}. \tag{27}$$

In addition, let us define the *type-enumeration detection expurgated exponent* $E_{\text{TE}}^{\text{EX}}(R,\alpha,P_X,W,V)$ as

$$\max_{0\leq s\leq 1}\min_{P_{X\tilde{X}}\in\mathcal{L}} \left\{\alpha s + \mathbb{E}\left[d_s(X,\tilde{X})\right] + I(P_{X\tilde{X}}) - R\right\}. \tag{28}$$

**Theorem 4.** *Let a distribution $P_X$ and a parameter $\alpha \in \mathbb{R}$ be given. Then, there exists a sequence of codes $\mathcal{C} = \{\mathcal{C}_n\}_{n=1}^{\infty}$ of rate $R$ such that for any $\delta > 0$*

$$E_{\text{FA}}(\mathcal{C},\phi^*) \geq E_{\text{TE}}^{\text{EX}}(R,\alpha,P_X,W,V) - \delta, \tag{29}$$

$$E_{\text{MD}}(\mathcal{C},\phi^*) \geq E_{\text{TE}}^{\text{EX}}(R,\alpha,P_X,W,V) - \alpha - \delta. \tag{30}$$

We summarize this section with the following discussion.

1) *Monotonicity in the rate*: Unlike the ordinary random coding exponent, the detection exponents do not necessarily decrease with the rate $R$. Thus, the required detection does not necessarily cause a rate loss, i.e. nulls the IE exponent at rates below $I(P_X \times W)$.

2) *Choice of input distribution*: Thus far, the input distribution $P_X$ was assumed fixed, but it can obviously be optimized. Nonetheless, there might be a tension between the optimal choice for channel coding versus the optimal choice for detection, and so a compromise should be made.

3) *Simplified Detectors*: The optimal detector (5) is rather difficult to implement, as it requires computations of the form $\sum_{m=1}^M W(\mathbf{y}|\mathbf{x}_m)$, which involve the sum of exponentially many likelihood terms, and each likelihood term is exponentially small. For low rates, a plausible approximation in (5) is $\sum_{m=1}^M W(\mathbf{y}|\mathbf{x}_m) \approx \max_{1\leq m\leq M} W(\mathbf{y}|\mathbf{x}_m)$ (and the same for $V$), and in [8] we have shown that the random coding exponents of th resulting approximated detector, denoted by $\phi_{\text{L}}$, can be derived just as for the optimal detector. For high rates, the output distribution of a capacity achieving code tends to be close to a memoryless distribution $\tilde{W} \triangleq (P_X \times W)_Y$, and an appropriate approximation is $\frac{1}{M}\sum_{m=1}^M W(\mathbf{y}|\mathbf{x}_m) \approx \tilde{W}(\mathbf{y})$. The random coding FA and MD exponents of such a detector follow directly from standard results [9, Section 11.7].

4) *Gallager/Forney style exponents:* In [8], we also derived Gallager/Forney-style lower bounds on the exponents. While these bounds can be strictly loose, it is indeed useful to derive them since: (i) They are simpler to compute, as they require solving at most two-dimensional (four-dimensional) optimization problems when there are no input constraints (with input constraints, respectively), irrespective of the input/output alphabet sizes. (ii) The bounds are translated almost verbatim to memoryless channels with continuous input/output alphabets, like the AWGN channel.

## V. Composite Detection

Up until now, we have assumed that detection is performed between two simple hypotheses, $W$ and $V$. Next, we discuss the generalization of the random coding analysis to composite hypotheses, i.e., a detection between a channel $W \in \mathcal{W}$ and a channel $V \in \mathcal{V}$, where $\mathcal{W}$ and $\mathcal{V}$ are some disjoint compact sets. Due to the nature of the problems outlined in the introduction (Section I), we adopt a *worst case* approach. For a codebook $\mathcal{C}_n$ and a given detector/decoder $\phi$, we generalize the FA probability to

$$P_{\text{FA}}(\mathcal{C}_n,\phi) \triangleq \max_{W\in\mathcal{W}} \frac{1}{M}\sum_{m=1}^M W(\mathcal{R}_0|\mathbf{x}_m), \tag{31}$$

and analogously, the MD and IE probabilities are obtained by maximizing over $V \in \mathcal{V}$ and $W \in \mathcal{W}$, respectively. Then, the trade-off between the IE probability and the FA and MD probabilities in (4) is defined exactly the same way.

Just as we have seen in (9), for any sequence of codebooks $\mathcal{C}$ and decoder $\phi$

$$E_{\text{IE}}(\mathcal{C}_n, \phi) = \min\left\{E_{\text{O}}(\mathcal{C}_n, \phi), E_{\text{FA}}(\mathcal{C}_n, \phi)\right\} \tag{32}$$

where here, $E_{\text{O}}(\mathcal{C}_n, \phi)$ is the exponent achieved by an ordinary decoder, which is not aware of $W$. Thus, the asymptotic separation principle holds here too, in the sense that the optimal detector/decoder may first use a detector which achieves the optimal trade-off between the FA and MD exponents, and then a decoder which achieves the optimal ordinary exponent.

We next discuss the achievable random coding exponents. As is well known, the *maximum mutual information* [10, Chapter 10, p. 147] universally achieves the random coding exponents for ordinary decoding. So, as in the simple hypotheses case, it remains to focus on the optimal trade-off between the FA and MD exponents, namely, solve

$$\begin{aligned} &\text{minimize} \quad P_{\text{FA}} \\ &\text{subject to} \quad P_{\text{MD}} \le e^{-n\overline{E}_{\text{MD}}} \end{aligned} \tag{33}$$

for some given exponent $\overline{E}_{\text{MD}} > 0$. The next lemma shows that the following *universal* detector/decoder $\phi^{\text{U}}$, whose rejection region $\mathcal{R}_0^{\text{U}}$ is defined as

$$\left\{\mathbf{y}:\ e^{n\alpha} \cdot \sum_{m=1}^{M} \max_{W\in\mathcal{W}} W(\mathbf{y}|\mathbf{x}_m) \le \sum_{m=1}^{M} \max_{V\in\mathcal{V}} V(\mathbf{y}|\mathbf{x}_m)\right\}, \tag{34}$$

solves (33). The universality here is in the sense of (33), i.e., achieving the best worst-case (over $W$) FA exponent, under a worst case constraint (over $V$) on the MD exponent. There might be, however, a loss in exponents compared to a detector which is aware of the actual pair $(W, V)$ (cf. Corollary 6).

**Lemma 5.** *Let $\mathcal{C}$ be a sequence of codebooks, let $\phi^{\text{U}}$ be as above, and let $\phi$ be any other detector/decoder. Then, if $E_{\text{FA}}(\mathcal{C}, \phi) \ge E_{\text{FA}}(\mathcal{C}, \phi^{\text{U}})$ then $E_{\text{MD}}(\mathcal{C}, \phi) \le E_{\text{MD}}(\mathcal{C}_n, \phi^{\text{U}})$.*

It remains to evaluate the random coding exponents for some $(W, V) \in \mathcal{W} \times \mathcal{V}$ when $\phi^{\text{U}}$ is used. Fortunately, this is a simple corollary to Theorem 3. Let us define the *generalized normalized log-likelihood ratio* of the set of channels $\mathcal{W}$ as

$$f_{\mathcal{W}}(Q) \triangleq \max_{W\in\mathcal{W}} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} Q(x,y) \log W(y|x). \tag{35}$$

**Corollary 6.** *Let $P_X$ and $\alpha \in \mathbb{R}$ be given. Then, there exists a sequence of codes $\mathcal{C} = \{\mathcal{C}_n\}_{n=1}^{\infty}$ of rate $R$, such that for any $\delta > 0$*

$$E_{\text{FA}}\left(\mathcal{C}, \phi^{\text{U}}\right) \ge E_{\text{TE,U}}^{\text{RC}}\left(R, \alpha, P_X, W, V\right) - \delta, \tag{36}$$

$$E_{\text{MD}}\left(\mathcal{C}, \phi^{\text{U}}\right) \ge E_{\text{TE,U}}^{\text{RC}}\left(R, \alpha, P_X, W, V\right) - \alpha - \delta \tag{37}$$

*where $E_{\text{TE,U}}^{\text{RC}}\left(R, \alpha, P_X, W, V\right)$ is defined as $E_{\text{TE}}^{\text{RC}}\left(R, \alpha, P_X, W, V\right)$ of (23), but replacing $f_W(Q)$ ($f_V(Q)$) with with $f_{\mathcal{W}}(Q)$ ($f_{\mathcal{V}}(Q)$, respectively) in all the definitions preceding Theorem 3.*

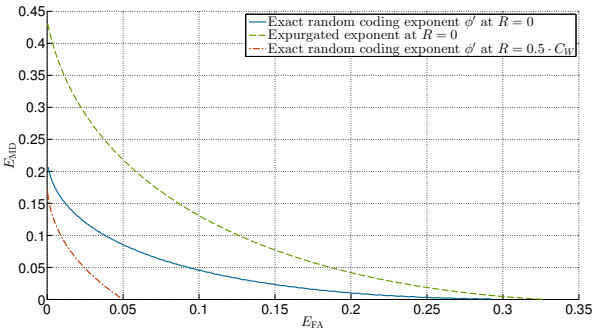



Figure 1. Achievable MD and FA exponents for $\phi'$ .

## VI. An Example

Let $W$ and $V$ be a pair of BSCs with crossover probabilities $w = 0.1$ and $v = 0.4$, respectively, and assume that $P_X = (\frac{1}{2}, \frac{1}{2})$, which results a capacity of $C_W \triangleq I(P_X \times W) \approx 0.37$ (nats). Interestingly, the output distributions $(P_X \times W)_Y$ and $(P_X \times V)_Y$ are both uniform, and so a detector which assumes a memoryless output distribution is useless, whereas the optimal decoder $\phi'$ produces strictly positive exponents.

We have plotted the the trade-off between the FA exponent and the MD exponent. Figure 1 shows that at zero rate, the expurgated bound significantly improves the random coding bound, and that in this case the exponents decrease as the rate increases to $R = 0.5 \cdot C_W$. Following the discussion at the end of Section IV, in this example, the simplified low-rate detector/decoder $\phi_{\text{L}}$ performs as well as the optimal detector/decoder $\phi'$ for all rates less than $R \approx 0.8 \cdot C_W$. In addition, the Gallager/Forney-style random coding exponents turn out to be strictly loose when $R = 0.5 \cdot C_W$, which exemplifies the importance of the ensemble-tight bounding technique of the type enumeration method used here.

## References


[1] D. N. C. T. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

[2] J. R. Barry, D. G. Messerschmitt, and E. A. Lee, *Digital communication: Third edition*. Norwell, MA, USA: Kluwer Academic Publishers, 2003.

[3] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *Information Theory, IEEE Transactions on*, vol. 58, no. 7, pp. 4215–4229, July 2012.

[4] N. Merhav, "Asymptotically optimal decision rules for joint detection and source coding," *Information Theory, IEEE Transactions on*, vol. 60, no. 11, pp. 6787–6795, Nov 2014.

[5] N. Weinberger and N. Merhav, "Codeword or noise? exact random coding exponents for joint detection and decoding," *Information Theory, IEEE Transactions on*, vol. 60, no. 9, pp. 5077–5094, Sept 2014.

[6] D. Wang, "Distinguishing codes from noise : fundamental limits and applications to sparse communication," MS.c thesis, Massachusetts Institute of Technology, June 2010.

[7] A. Tchamkerten, V. Chandar, and G. W. Wornell, "Communication under strong asynchronism," *Information Theory, IEEE Transactions on*, vol. 55, no. 10, pp. 4508–4528, Oct 2009.

[8] N. Weinberger and N. Merhav, "Channel detection in coded communication," *Submitted to Information Theory, IEEE Transactions on*, September 2015, available online: http://arxiv.org/pdf/1509.01806.pdf.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

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

# A Converse for Lossy Source Coding in the Finite Blocklength Regime

Lars Palzer and Roy Timo
Technische Universität München
{lars.palzer, roy.timo}@tum.de

***Abstract*—We present a converse bound for lossy source coding in the finite blocklength regime. The bound is based on $\mathsf{d}$-tilted information, and it combines ideas from two different converse techniques by Kostina and Verdú. When particularised to the binary and Gaussian memoryless sources, the new bound gives slightly tighter results in certain blocklength regimes.**

## I. Introduction and Main Results

Kostina and Verdú recently presented two general converse bounds in [1] for the problem of lossy source coding at finite blocklengths. These bounds were respectively based on $\mathsf{d}$-*tilted information* and *binary hypothesis testing* arguments. Matsuta and Uyematsu [2] presented a converse at ISIT'15 that is tighter than Kostina and Verdú's meta-converse, but, unfortunately, this bound is not (yet) numerically computable.

The main purpose of this paper is to report a new converse bound for general sources, and to particularise the bound to the *binary memoryless source* (BMS) with Hamming distortions and the *Gaussian memoryless source* (GMS) with squared-error distortions. The new bound is simply stated and (we believe) quite intuitive. Its proof combines ideas from Kostina and Verdú's $\mathsf{d}$-tilted and meta-converse bounds, and it gives slightly better numerical results for the BMS and GMS at certain blocklengths.

### A. Problem Statement & Basic Functions

Our presentation will follow the one-shot paradigm in [1]: We first consider an abstract rate-distortion (RD) problem that consists of compressing and reconstructing a single random variable. We then specialise this one-shot problem setup to the block encoding and decoding of memoryless sources.

Let $X$ be the output of a general source with distribution $p_X$ on an alphabet $\mathcal{X}$. A (possibly stochastic) encoder

$$f : \mathcal{X} \to \mathcal{M} := \{1, 2, \ldots, M\}$$

maps the source output $X$ to an index $T := f(X)$ from which a (possibly stochastic) decoder

$$g : \mathcal{M} \to \hat{\mathcal{X}}$$

outputs $\hat{X} := g(T)$ as its estimate of $X$. Let

$$\mathsf{d} : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$$

denote the distortion function.

**Definition.** An $(M, d, \varepsilon)$-code consists of an encoder $f$ and decoder $g$, as described above, with $\mathbb{P}\big[\,\mathsf{d}(X, \hat{X}) > d\big] \leq \varepsilon$.

In this paper, the main problem is to find lower bounds on the smallest $M$ for which there exists an $(M, d, \varepsilon)$-code.

The abstract problem formulation above can be specialised to block encoding/decoding of memoryless sources as follows.

**Definition.** An $(n, M, d, \varepsilon)$ code for a memoryless source with distribution $p_{\mathbb{X}} = p_X^n := p_X \times \ldots \times p_X$ putting out strings $\mathbb{X}$ of length $n$ from $\mathcal{X}^n = \mathcal{X} \times \ldots \times \mathcal{X}$ and reconstruction alphabet $\hat{\mathcal{X}}^n$ consists of an encoder $f : \mathcal{X}^n \to \mathcal{M}$ and a decoder $g : \mathcal{M} \to \hat{\mathcal{X}}^n$ satisfying $\mathbb{P}\big[\,\mathsf{d}(\mathbb{X}, \hat{\mathbb{X}}) > d\big] \leq \varepsilon$.

Let

$$R(d) := \inf_{p_{\hat{X}|X}:\ \mathsf{E}[\mathsf{d}(X,\hat{X})] \leq d} \mathsf{E}\big[\imath_{X;\hat{X}}(X; \hat{X})\big], \tag{1}$$

denote the usual *RD function*, where

$$\imath_{X;\hat{X}}(x; \hat{x}) := \log \frac{\mathsf{d}p_{\hat{X}|X=x}}{\mathsf{d}p_{\hat{X}}}(\hat{x}),$$

is the *information density* of $p_{X\hat{X}} = p_{\hat{X}|X} p_X$. As in [1], we make the following two basic assumptions:

A1. The distortion constraint $d$ satisfies $R(d) < \infty$.

A2. The infimum in (1) is achieved by a unique[1] $p^*_{\hat{X}|X}$.

Let $p^*_{\hat{X}}$ denote the $\hat{X}$-marginal on $\hat{\mathcal{X}}$ induced by $p^*_{\hat{X}|X}$ and $p_X$, and define $\lambda := -R'(d)$ to be the negative slope of the RD function at distortion $d$. Let

$$\jmath_X(x, d) := \log \frac{1}{\mathsf{E}_{p^*_{\hat{X}}}[\exp(\lambda(d - \mathsf{d}(x, \hat{X}))]},$$

where the expectation is taken with respect to $p^*_{\hat{X}}$. The function $\jmath_X(x, d)$ is called $\mathsf{d}$-*tilted information*, and, intuitively, it corresponds to the number of bits required to represent a particular source realisation $x$ to within distortion $d$. For example, one can show [3], [4] that $R(d) = \mathsf{E}[\jmath_X(X, d)]$. We now summarise the main results of the paper. These results are proved in Sections II, III and IV.

### B. General Sources

We start with a converse for general sources.

**Theorem 1.** *Any $(M, d, \varepsilon)$ code must satisfy*

$$M \geq \sup_{\beta \in \mathbb{R}} \left( \frac{\mathbb{P}\big[\jmath_X(X, d) \geq \beta\big] - \varepsilon}{\sup_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{P}\big[\jmath_X(X, d) \geq \beta,\, \mathsf{d}(X, \hat{x}) \leq d\big]} \right). \tag{2}$$

[1] We make this assumption for clarity of presentation. As mentioned in [1, Remark 9], it can be relaxed.

### C. Binary Memoryless Sources (BMS)

The next corollary specialises Theorem 1 to the special case of a BMS with Hamming distortions. Let $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ be a string of $n$ iid instances of $X \sim \text{Bernoulli}(p)$, and choose the distortion function to be

$$\mathsf{d}(\mathbb{X}, \hat{\mathbb{X}}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \neq \hat{X}_i\}.$$

**Corollary 2** (BMS). *Fix $p \in (0, 1/2)$ and $d \in [0, p)$. Any $(n, M, d, \varepsilon)$ code must satisfy*

$$M \geq \max_{0 \leq b \leq n} \left( \frac{\sum_{k=b}^{n} \binom{n}{k} p^k (1-p)^{n-k} - \varepsilon}{\alpha_{n,d,p}(b)} \right), \tag{3}$$

*where*

$$\alpha_{n,d,p}(b) = \max_{\hat{n}_1} \sum_{k=0}^{\lfloor nd \rfloor} \sum_{l=0}^{k} \binom{\hat{n}_1}{l} \binom{n-\hat{n}_1}{k-l} \cdot p^{\hat{n}_1 + k - 2l} (1-p)^{n - \hat{n}_1 - k + 2l} \mathbb{1}\{\hat{n}_1 + k - 2l \geq b\}$$

*and the maximisation is taken over all $\hat{n}_1 \in \mathbb{N}$ satisfying*

$$\max\{0, b - \lfloor nd \rfloor\} \leq \hat{n}_1 \leq \min\{n, b + \lfloor nd \rfloor\}.$$

It is worth noting that Corollary 2 does not weaken Theorem 1; that is, the right hand sides of (2) and (3) are equal for the BMS with Hamming distortions.

*R*emark: For $p = 1/2$, $\jmath_X(\mathbb{X}, d)$ does not depend on $\mathbb{X}$ [1, Example 1]. In this case, Theorem 1 coincides with [1, Thm. 20] which is derived from the meta-converse bound.

### D. Gaussian Memoryless Sources (GMS)

Now let $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ be a string of $n$ iid instances of $X \sim \mathcal{N}(0, 1)$, and consider the squared-error distortions

$$\mathsf{d}(\mathbb{X}, \hat{\mathbb{X}}) = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{X}_i|^2.$$

A slight weakening of Theorem 1 yields the next corollary. Here $f_{\chi^2_n}(\cdot)$ denotes the $\chi^2_n$ probability density function.

**Corollary 3** (GMS). *Fix $d \in (0, 1)$. Any $(n, M, d, \varepsilon)$ code must satisfy*

$$M \geq \sup_{\gamma \geq nd} \left( \frac{\int_{\gamma}^{\infty} f_{\chi^2_n}(w)\, \mathsf{d}w - \varepsilon}{\frac{1}{2} I_{nd/\gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right) \int_{\gamma}^{\gamma^\star} f_{\chi^2_n}(w)\, \mathsf{d}w} \right), \tag{4}$$

*where $I_{(\cdot)}(\cdot, \cdot)$ is the regularized incomplete beta function and*

$$\gamma^\star := \left[ \frac{2(nd)^{n/2}}{I_{nd/\gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right)} + \gamma^{n/2} \right]^{2/n}. \tag{5}$$

### E. Comparions to Existing Bounds

We now compare Theorem 1 and Corollaries 2 and 3 to Kostina and Verdú's d-*tilted information* and *meta-converse* bounds in [1]. Let us first recall the d-tilted information bound.

**Theorem KV-1.** Any $(M, d, \varepsilon)$ code must satisfy [1, Thm. 7]

$$\varepsilon \geq \sup_{\gamma \geq 0} \Big( \mathbb{P}[\jmath_X(X, d) \geq \log M + \gamma] - e^{-\gamma} \Big). \tag{6}$$

To compare Theorem 1 with Theorem KV-1, it is helpful to first rewrite (2) as a lower bound on $\varepsilon$:

$$\varepsilon \geq \sup_{\beta \in \mathbb{R}} \Bigg( \mathbb{P}\big[\jmath_X(X, d) \geq \beta\big] - M \sup_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{P}\big[\jmath_X(X, d) \geq \beta,\, \mathsf{d}(X, \hat{x}) \leq d\big] \Bigg). \tag{7}$$

Given the similarities between (6) and (7), one might guess that Theorem KV-1 can be recovered as a special case of Theorem 1 by choosing $\beta$ appropriately in (7). We now show that this is indeed the case, and, therefore, Theorem KV-1 cannot be stronger than Theorem 1.

Choose $\beta = \log M + \gamma$ and consider the rightmost term in (7). We have[2]

$$\begin{aligned} & M \sup_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{P}\big[\jmath_X(X, d) \geq \log M + \gamma,\, \mathsf{d}(X, \hat{x}) \leq d\big] \\ & \overset{\text{a}}{=} M \sup_{\hat{x} \in \hat{\mathcal{X}}} \mathsf{E}\left[ \mathbb{1}\left\{ \frac{1}{M}\, e^{\jmath_X(X,d) - \gamma} \geq 1,\; e^{\lambda(d - \mathsf{d}(X,\hat{x}))} \geq 1 \right\} \right] \\ & \overset{\text{b}}{\leq} M \sup_{\hat{x} \in \hat{\mathcal{X}}} \mathsf{E}\left[ \frac{1}{M}\, e^{\jmath_X(X,d) - \gamma}\, \mathbb{1}\big\{ e^{\lambda(d - \mathsf{d}(X,\hat{x}))} \geq 1 \big\} \right] \\ & \overset{\text{c}}{\leq} e^{-\gamma} \sup_{\hat{x} \in \hat{\mathcal{X}}} \mathsf{E}\Big[ e^{\jmath_X(X,d) + \lambda(d - \mathsf{d}(X,\hat{x}))} \Big] \\ & \overset{\text{d}}{\leq} e^{-\gamma}, \end{aligned} \tag{8}$$

where (a) follows because the RD function $R(d)$ is non-increasing in $d$ and therefore $\lambda \geq 0$; (b) and (c) follow from Markov's inequality; and (d) applies the next lemma[3].

**Lemma 4** (Csiszár). *For all $\hat{x} \in \hat{\mathcal{X}}$ [3, Eq. (1.22)],*

$$\mathsf{E}\big[ e^{\jmath_X(X,d) + \lambda(d - \mathsf{d}(X,\hat{x}))} \big] \leq 1,$$

*with equality for $p_{\hat{X}^\star}$-almost all $\hat{x}$.*

The second converse result from Kostina and Verdú that we will consider is based on binary hypothesis testing. Let

$$\beta_\alpha(p, q) = \min_{\substack{p_{W|X} \\ \mathbb{P}[W=1] \geq \alpha}} \mathbb{Q}\big[W = 1\big], \tag{9}$$

denote the optimal performance achievable among all randomised tests $p_{W|X} : \mathcal{X} \to \{0, 1\}$ between probability distributions $p$ and $q$ on $\mathcal{X}$ where 1 indicates that the test chooses $p$ and $\mathbb{Q}[\cdot]$ is the probability of an event if $X$ has distribution $q$.

**Theorem KV-2.** Any $(M, d, \varepsilon)$ code must satisfy [1, Thm. 8]

$$M \geq \sup_{q_X} \inf_{\hat{x} \in \hat{\mathcal{X}}} \frac{\beta_{1-\varepsilon}(p_X, q_X)}{\mathbb{Q}\big[\mathsf{d}(X, \hat{x}) \leq d\big]}, \tag{10}$$

where the supremum is taken over all distributions on $\mathcal{X}$.

*Remark:* After submitting this paper, we found that one can derive Theorem 1 by making a (suboptimal) choice for $q_X$ in Theorem KV-2; see the Appendix. This also shows that Theorem KV-2 is never weaker than Theorem KV-1.

[2] The following arguments are based on the proof of Theorem KV-1 in [1].
[3] See also Property 2 in [1, p. 3311].

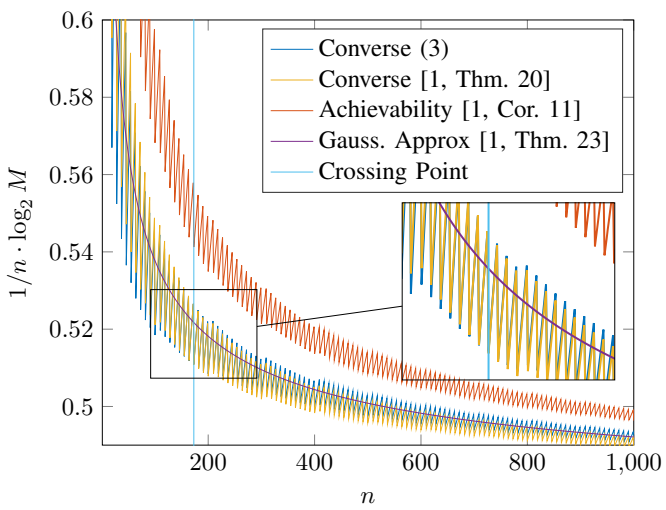



Fig. 1. BMS, $d = 0.11$, $p = 2/5$, $\varepsilon = 10^{-2}$.

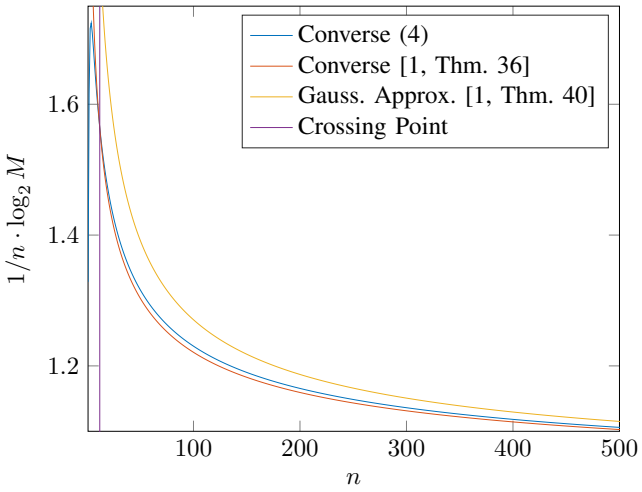



Fig. 2. GMS, $d = 0.25$, $\sigma^2 = 1$, $\varepsilon = 10^{-2}$.

### F. Numerical Results

Consider the BMS under Hamming distortions with the following parameters: $p = 2/5$, $d = 0.11$ and $\varepsilon = 10^{-2}$. Figure 1 plots the lower bound (3) from Corollary 2. For comparison, the best converse bound from in [1, Thm. 20] is also plotted. For this particular setup, [1, Thm. 20] is tighter for $n < 173$, but weaker for $n \geq 173$. For completeness, we have also plotted the Gaussian approximation [1, Thm. 23] and an achievability result based on random coding [1, Theorem 10]. There, we chose $p_{\hat{X}} = p_{\hat{X}}^n$ and set $p_{\hat{X}}(1) = \frac{p-d'}{1-2d'}$ with $d' := \lfloor nd \rfloor / n$, which is slightly better than choosing $p_{\hat{X}}(1) = \frac{p-d}{1-2d}$. Computations with other parameters indicate that the crossing point moves to smaller $n$ when increasing $d$ or $\varepsilon$ and to larger $n$ otherwise.

Now consider the GMS under squared error distortions with the following parameters: $d = 0.25$, $\sigma^2 = 1$ and $\varepsilon = 10^{-2}$. Figure 2 plots the bound in (4) and, for comparison, the converse bound [1, Theorem 36], which can be derived from (10). Our result is tighter for $n \geq 12$. We also included the Gaussian approximation [1, Theorem 40]. Here, choosing small values for $d$ shifts the crossing point to larger $n$ whereas varying $\varepsilon$ does not seem to have a significant influence.

## II. General Sources: Proof of Theorem 1

For ease of notation, we assume that $\mathcal{X}$ and $\hat{\mathcal{X}}$ are finite sets but note that the result applies to general abstract sources.

Let $\beta \in \mathbb{R}$ be arbitrary. In the same manner as the proof of Theorem KV-1 in [1], we start by bounding

$$\begin{aligned}\mathbb{P}\big[\jmath_X(X,d) \geq \beta\big] &= \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{X}) > d\big] \\ &\quad + \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{X}) \leq d\big] \\ &\leq \varepsilon + \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{X}) \leq d\big]. \quad (11)\end{aligned}$$

Now consider the second probability of the RHS. Using similar arguments as the proof of Theorem KV-2 in [1],

$$\begin{aligned}&\mathbb{P}\big[\jmath_X(X,d) \geq \beta,\ \mathsf{d}(X,\hat{X}) \leq d\big] \\ &\quad= \sum_{x\in\mathcal{X}} p_X(x) \sum_{t\in\mathcal{M}} \underbrace{p_{T|X}(t|x)}_{\leq 1} \\ &\qquad\qquad \sum_{\hat{x}\in\hat{\mathcal{X}}} p_{\hat{X}|T}(\hat{x}|t)\ \mathbb{1}\big\{\jmath_X(x,d) \geq \beta,\, \mathsf{d}(x,\hat{x}) \leq d\big\} \\ &\quad\leq \sum_{t\in\mathcal{M}} \sum_{\hat{x}\in\hat{\mathcal{X}}} p_{\hat{X}|T}(\hat{x}|t) \sum_{x\in\mathcal{X}} p_X(x) \\ &\qquad\qquad \mathbb{1}\big\{\jmath_X(x,d) \geq \beta,\, \mathsf{d}(x,\hat{x}) \leq d\big\} \\ &\quad= \sum_{t\in\mathcal{M}} \sum_{\hat{x}\in\hat{\mathcal{X}}} p_{\hat{X}|T}(\hat{x}|t) \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{x}) \leq d\big] \\ &\quad\leq \sum_{t\in\mathcal{M}} \sup_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{x}) \leq d\big] \\ &\quad= M \sup_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{P}\big[\jmath_X(X,d) \geq \beta,\, \mathsf{d}(X,\hat{x}) \leq d\big]. \quad (12)\end{aligned}$$

To complete the proof, combine (11) and (12) and take the supremum over $\beta$ to get (2) or (7). □

## III. BMS: Proof of Corollary 2

Fix $p \in (0, 1/2)$, $d \in [0, p)$ and $\beta \in \mathbb{R}$. Let $h_2(\cdot)$ denote the binary entropy function. We have [1, Eqn. (21)]

$$\jmath_{\mathbb{x}}(\mathbb{x}, d) = \mathsf{N}(1|\mathbb{x}) \log \frac{1}{p} + (n - \mathsf{N}(1|\mathbb{x})) \log \frac{1}{1-p} - n h_2(d),$$

where

$$\mathsf{N}(1|\mathbb{x}) := \sum_{k=1}^{n} \mathbb{1}\{x_i = 1\}.$$

Since $p \in (0, 1/2)$, it follows that $p < 1 - p$ and $\jmath_{\mathbb{x}}(\mathbb{x}, d)$ grows linearly in $\mathsf{N}(1|\mathbb{x})$ for fixed $n$. Let

$$b := \min\Big\{ n' \in \{0, \ldots, n\} : \\ n' \log \frac{1}{p} + (n - n') \log \frac{1}{1-p} - n h_2(d) \geq \beta \Big\},$$

and note that

$$\big\{\mathbb{x} \in \mathcal{X}^n : \jmath_{\mathbb{X}}(\mathbb{x}, d) \geq \beta\big\} = \big\{\mathbb{x} \in \mathcal{X}^n : \mathsf{N}(1|\mathbb{x}) \geq b\big\}.$$

Hence,

$$\begin{aligned}\mathbb{P}[\jmath_{\mathbb{X}}(\mathbb{X}, d) \geq \beta] &= \mathbb{P}\big[\mathsf{N}(1|\mathbb{X}) \geq b\big] \\ &= \sum_{k=b}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{13}$$

Now consider the denominator of (2). Let $\hat{n}_1 := \mathsf{N}(1|\hat{\mathbb{x}})$. Using Vandermonde's identity, the number of binary sequences in a Hamming ball of size $\lfloor nd \rfloor$ centered at a sequence of Hamming weight $\hat{n}_1$ is given by

$$\sum_{k=0}^{\lfloor nd \rfloor} \binom{n}{k} = \sum_{k=0}^{\lfloor nd \rfloor} \sum_{l=0}^{k} \binom{\hat{n}_1}{l} \binom{n-\hat{n}_1}{k-l},$$

where, in each summand, $\binom{\hat{n}_1}{l}\binom{n-\hat{n}_1}{k-l}$ is the number of sequences of Hamming weight $\hat{n}_1 + k - 2l$. We can thus write

$$\begin{aligned}&\sup_{\hat{\mathbb{x}} \in \hat{\mathcal{X}}^n} \mathbb{P}\big[\jmath_{\mathbb{X}}(\mathbb{X}, d) \geq \beta,\, \mathsf{d}(\mathbb{X}, \hat{\mathbb{x}}) \leq nd\big] \\ &\quad \overset{*}{=} \max_{\hat{n}_1} \mathbb{P}\big[\mathsf{N}(1|\mathbb{X}) \geq b,\, \mathsf{d}(\mathbb{X}, \hat{\mathbb{x}}) \leq nd\big] \\ &\quad = \max_{\hat{n}_1} \sum_{\mathbb{x}} p_{\mathbb{X}}(\mathbb{x}) \mathbb{1}\big\{\mathsf{N}(1|\mathbb{x}) \geq b,\, \mathsf{d}(\mathbb{x}, \hat{\mathbb{x}}) \leq nd\big\} \\ &\quad = \max_{\hat{n}_1} \sum_{k=0}^{\lfloor nd \rfloor} \sum_{l=0}^{k} \binom{\hat{n}_1}{l} \binom{n-\hat{n}_1}{k-l} p^{\hat{n}_1 + k - 2l} \\ &\qquad \cdot (1-p)^{n - \hat{n}_1 - k + 2l} \mathbb{1}\big\{\underbrace{\hat{n}_1 + k - 2l}_{=\mathsf{N}(1|\mathbb{x})} \geq b\big\}, \end{aligned} \tag{14}$$

where (*) follows since, by symmetry, the probability depends on $\hat{\mathbb{x}}$ only through $\hat{n}_1$.

In fact, we only need to consider $b - \lfloor nd \rfloor \leq \hat{n}_1 \leq b + \lfloor nd \rfloor$ for the maximisation. This is because for $\hat{n}_1 < b - \lfloor nd \rfloor$, we have $\mathbb{1}\big\{\hat{n}_1 + k - 2l \geq b\big\} = 0$ for all summands and for $\hat{n}_1 > b + \lfloor nd \rfloor$, $\mathbb{1}\big\{\hat{n}_1 + k - 2l \geq b\big\} = 1$ for all summands in which case the sum is monotonically decreasing in $\hat{n}_1$ (we omit the proof of this fact). □

## IV. GMS: Proof of Corollary 3

The $\mathsf{d}$-tilted information for the GMS with $d < \sigma^2 = 1$ is given by [1, Example 2]

$$\jmath_{\mathbb{X}}(\mathbb{x}, d) = \frac{n}{2} \log \frac{1}{d} + \frac{\|\mathbb{x}\|^2 - n}{2} \log e,$$

which grows linearly in $\|\mathbb{x}\|^2$. Hence, we can rewrite (2) as

$$M \geq \sup_{\gamma \geq 0} \left( \frac{\mathbb{P}\big[\|\mathbb{X}\|^2 \geq \gamma\big] - \varepsilon}{\sup\limits_{\hat{\mathbb{x}} \in \mathbb{R}^n} \mathbb{P}\big[\|\mathbb{X}\|^2 \geq \gamma,\, \mathsf{d}(\mathbb{X}, \hat{\mathbb{x}}) \leq d\big]} \right). \tag{15}$$

We will lower bound (15) using a geometric argument for the denominator. By the circular symmetry of the GMS, we only need to consider those $\hat{\mathbb{x}} \in \mathbb{R}^n$ for the supremum that lie on an arbitrary straight line through the origin. Denote

$$\mathcal{A} := \big\{\mathbb{x} \in \mathbb{R}^n : \|\mathbb{x}\|^2 \geq \gamma\big\},$$

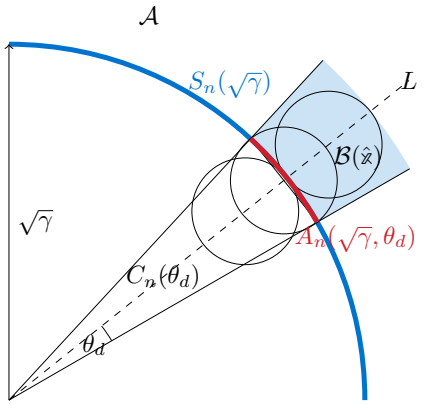



Fig. 3. Intersection of $\mathcal{A}$ with possible distortion balls.

$$\mathcal{B}(\hat{\mathbb{x}}) := \big\{\mathbb{x} \in \mathbb{R}^n : \|\mathbb{x} - \hat{\mathbb{x}}\|^2 \leq nd\big\},$$

and observe that

$$\begin{aligned}&\sup_{\hat{\mathbb{x}} \in \mathbb{R}^n} \mathbb{P}\big[\|\mathbb{X}\|^2 \geq \gamma,\, \mathsf{d}(\mathbb{X}, \hat{\mathbb{x}}) \leq d\big] \\ &\qquad = \sup_{\hat{\mathbb{x}} \in \mathbb{R}^n} \mathbb{P}\big[\mathbb{X} \in \mathcal{A} \cap \mathcal{B}(\hat{\mathbb{x}})\big] \\ &\qquad = \sup_{\hat{\mathbb{x}} \in L} \mathbb{P}\big[\mathbb{X} \in \mathcal{A} \cap \mathcal{B}(\hat{\mathbb{x}})\big], \end{aligned} \tag{16}$$

where $L$ denotes the set of points lying on a straight line through the origin, see Figure 3.

Denote the surface area of an $n$-dimensional sphere of radius $r$ by $S_n(r)$ and the surface area of a $n$-dimensional spherical cap of radius $r$ and half angle $\theta$ by $A_n(r, \theta)$. The following relation holds [5]:

$$A_n(r, \theta) := \frac{1}{2} S_n(r) I_{\sin^2(\theta)}\left(\frac{n-1}{2}, \frac{1}{2}\right),$$

where $I_{(\cdot)}(\cdot, \cdot)$ is the regularized incomplete beta function. Using the law of sines and taking $\gamma \geq nd$, we can determine the half angle $\theta_d$ such that $A_n(\sqrt{\gamma}, \theta_d)$ is the largest spherical cap at radius $\sqrt{\gamma}$ contained in some $\mathcal{B}(\hat{\mathbb{x}})$:

$$\theta_d = \sin^{-1} \sqrt{nd/\gamma}.$$

Let $C_n(\theta_d)$ be the $n$-dimensional infinite cone of half angle $\theta_d$ that passes through $A_n(\sqrt{\gamma}, \theta_d)$. Clearly, $\mathcal{A} \cap \mathcal{B}(\hat{\mathbb{x}}) \subset C_n(\theta_d)$, for any $\hat{\mathbb{x}} \in L$. This setup is visualized in Figure 3. Next, denote the volume of $\mathcal{B}(\hat{\mathbb{x}})$ for any $\hat{\mathbb{x}} \in \mathbb{R}^n$ by

$$V_n\big(\sqrt{nd}\big) := \frac{\pi^{n/2}}{\Gamma\left(\frac{n+2}{2}\right)} (nd)^{n/2},$$

with $\Gamma(\cdot)$ being the gamma function. To upper bound (16), we consider the largest probability of any set in $\mathcal{A} \cap C_n(\theta_d)$ (the shaded area in Figure 3) that has the same volume as a distortion ball. We denote this set by

$$\mathcal{K}^\star := \operatorname*{arg\,max}_{\substack{\mathcal{K} \subset \mathcal{A} \cap C_n(\theta_d): \\ \mathrm{Vol}(\mathcal{K}) = V_n(\sqrt{nd})}} \mathbb{P}\big[\mathbb{X} \in \mathcal{K}\big] \tag{17}$$

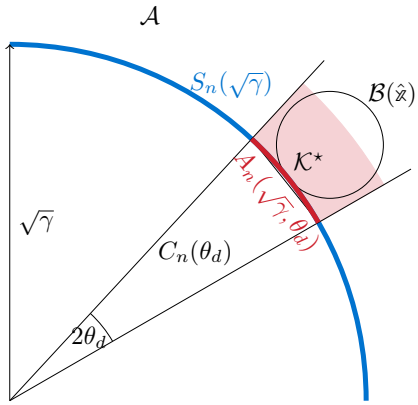



Fig. 4. Geometry of $\mathcal{K}^\star$.

The geometry of the $\arg\max$ problem is depicted in Figure 4. By the circular symmetry, $\mathcal{K}^\star$ is the slice of the cone $C_n(\theta_d)$ that lies on the surface of $S_n(\sqrt{\gamma},\theta_d)$ and has volume $V_n\big(\sqrt{nd}\big)$. More precisely, we can describe $\mathcal{K}^\star$ as the difference between spherical sectors of half angle $\theta_d$ whose volumes differ by exactly $V_n(\sqrt{nd})$, see Figure 4. The volume of a hypershperical sector of half angle $\theta$ and radius $r$ is given by [5]

$$V_n^{\text{sec}}(r,\theta) := \frac{1}{2}V_n(r)I_{\sin^2(\theta)}\left(\frac{n-1}{2},\frac{1}{2}\right).$$

Now let $\gamma^\star$ be the solution to

$$V_n^{\text{sec}}(\sqrt{\gamma^\star},\theta_d) - V_n^{\text{sec}}(\sqrt{\gamma},\theta_d) = V_n(\sqrt{nd}), \tag{18}$$

which, using $\sin^2(\theta_d) = nd/\gamma$, can be rewritten as (5).

Now, we can use the tools developed in (16)–(18) to bound

$$\begin{aligned}
&\sup_{\hat{x}\in L}\mathbb{P}\left[\mathbb{X}\in\mathcal{A}\cap\mathcal{B}(\hat{x})\right]\\
&\quad\le \mathbb{P}\left[\mathbb{X}\in\mathcal{K}^\star\right] &(19)\\
&\quad= \mathbb{P}\left[\gamma\le\|\mathbb{X}\|^2\le\gamma^\star,\mathbb{X}\in C_n(\theta_d)\right] &(20)\\
&\quad= \mathbb{P}\left[\gamma\le\|\mathbb{X}\|^2\le\gamma^\star\right]\mathbb{P}\left[\mathbb{X}\in C_n(\theta_d)\right], &(21)\\
&\quad= \mathbb{P}\left[\gamma\le\|\mathbb{X}\|^2\le\gamma^\star\right]\frac{A_n(\sqrt{\gamma},\theta_d)}{S_n(\sqrt{\gamma})} &(22)\\
&\quad= \frac{1}{2}I_{nd/\gamma}\left(\frac{n-1}{2},\frac{1}{2}\right)\int_\gamma^{\gamma^\star} f_{\chi_n^2}(w)\,\mathrm{d}w &(23)
\end{aligned}$$

where (19) follows from the definition of $\mathcal{K}^\star$ (17), (20) follows from the definition of $\gamma^\star$ (18), and the geometry of $\mathcal{K}^\star$, (21)–(22) are a result of the circular symmetry of the multivariate Gaussian and $f_{\chi_n^2}(\cdot)$ is the $\chi_n^2$ probability density function. Combining (23) and (15) then yields (4).


## Acknowledgements

This work was supported by the Alexander von Humboldt Foundation and the German Research Foundation. We would like to thank Rana Ali Amjad for helpful discussions and Victoria Kostina for helpful comments on different drafts.


## Appendix

As noted in [1, Rem. 5], Theorem KV-2 can be strengthened by relaxing the requirement that $q_X$ be a probability measure in (9): Instead, we may allow $q_X$ to be any $\sigma$-finite measure to obtain the following bound.

**Theorem KV-3.** Any $(M,d,\varepsilon)$-code must satisfy

$$M \ge \sup_{q_X}\inf_{\hat{x}\in\hat{\mathcal{X}}}\frac{\beta_{1-\varepsilon}(p_X,q_X)}{\mathbb{Q}\big[\mathsf{d}(X,\hat{x})\le d\big]}, \tag{24}$$

where the supremum is taken over all $\sigma$-finite measures $q_X$.

We will now recover Theorem 1 from Theorem KV-3. Choose $q_X$ such that

$$\mathbb{Q}[X\in\mathcal{A}] = \mathbb{P}[X\in\mathcal{A}, \jmath_X(X,d)\ge\beta], \quad \forall\mathcal{A}\subseteq\mathcal{X}.$$

An optimal randomised test between $p_X$ and $q_X$ is

$$p_{W|X}(1|x) := \begin{cases} 1, & \text{if } \jmath_X(x,d)<\beta\\ \dfrac{\mathbb{P}[\jmath_X(X,d)\ge\beta]-\varepsilon}{\mathbb{P}[\jmath_X(X,d)\ge\beta]}, & \text{if } \jmath_X(x,d)\ge\beta, \end{cases}$$

The probability that this test succeeds under $p_X$ is

$$\begin{aligned}
\mathbb{P}[W=1] &= \mathbb{P}[\jmath_X(X,d)<\beta]\ \mathbb{P}[W=1|\jmath_X(X,d)<\beta]\\
&\quad+\mathbb{P}[\jmath_X(X,d)\ge\beta]\ \mathbb{P}[W=1|\jmath_X(X,d)\ge\beta]\\
&= 1-\varepsilon.
\end{aligned}$$

Moreover, the probability that the test fails under $q_X$ is

$$\begin{aligned}
\mathbb{Q}[W=1] &= \mathbb{P}[\jmath_X(X,d)\ge\beta, W=1]\\
&= \mathbb{P}[\jmath_X(X,d)\ge\beta]\cdot\frac{\mathbb{P}[\jmath_X(X,d)\ge\beta]-\varepsilon}{\mathbb{P}[\jmath_X(X,d)\ge\beta]}\\
&= \mathbb{P}[\jmath_X(X,d)\ge\beta]-\varepsilon.
\end{aligned}$$

Substituting $q_X$ into Theorem KV-3 gives

$$M \ge \inf_{\hat{x}\in\hat{\mathcal{X}}}\frac{\mathbb{P}[\jmath_X(X,d)\ge\beta]-\varepsilon}{\mathbb{P}[\jmath_X(X,d)\ge\beta, d(X,\hat{x})\le d]}.$$

Taking the supremum over $\beta$ gives Theorem 1.

*Remark:* The above discussion together with that in Section I-E demonstrates that the $\mathsf{d}$-tilted converse in Theorem KV-1 cannot be tighter than the (generalised $\sigma$-finite measure) meta converse in Theorem KV-3. To the best of our knowledge, this fact has not been observed in the literature before.

## References


[1] V. Kostina, and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012

[2] T. Matsuta and T. Uyematsu, "Non-asymptotic bounds for fixed-length lossy compression," *Proc IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015

[3] I. Csiszár, "On an extremum problem of information theory", *Studia Sci. Math. Hungarica*, vol. 9, p. 57–71, 1974

[4] R. Gray, "Entropy and Information Theory," *Springer*, 2nd Edition, 2011

[5] S. Li, "Concise formulas for the area and volume of a hyperspherical cap", *Asian Journal of Math. and Stat.* vol. 4, no. 1, pp. 66–70, 2011

# Frequency hopping does not increase anti-jamming resilience of wireless channels

Moritz Wiese and Panos Papadimitratos

Networked Systems Security Group
KTH Royal Institute of Technology, Stockholm, Sweden
{moritzw, papadim}@kth.se

***Abstract*—The effectiveness of frequency hopping for anti-jamming protection of wireless channels is analyzed from an information-theoretic perspective. The sender can input its symbols into one of several frequency subbands at a time. Each subband channel is modeled as an additive noise channel. No common randomness between sender and receiver is assumed. It is shown that capacity is positive, and then equals the common randomness assisted (CR) capacity, if and only if the sender power strictly exceeds the jammer power. Thus compared to transmission over any fixed frequency subband, frequency hopping is not more resilient towards jamming, but it does increase the capacity. Upper and lower bounds on the CR capacity are provided.**

## I. Introduction

A wireless channel is open to inputs from anybody operating on the same frequency. Therefore communication has to be protected against deliberate jamming. This means that communication protocols have to be devised whose application enables reliable data transmission even if attacked by a jammer.

If a sufficiently broad frequency band is available, and if the jammer does not have simultaneous access to the complete band, a method which suggests itself is frequency hopping (FH). The frequency spectrum is divided into subbands. In each time slot, the sender chooses a subband in a random way and uses only that frequency to transmit data in that time slot. In some models [4], [6], the receiver hops over frequencies, too, and only listens to one subband at a time. The idea is that in this way, the channel will not be jammed all the time with positive probability, and some information will go through.

To succeed, the basic FH idea requires common randomness known to sender and receiver, but unknown to the jammer. A careful analysis of that situation has been performed in [4]. It is clearly necessary that the common randomness realization be known before transmission starts. As the channel cannot be used to distribute this knowledge, this leads to a circle called anti-jamming/key-establishment dependency in [6].

In [6] it has been investigated for the first time whether FH can be used for data transmission without the availability of common randomness. Moreover, the jammer is allowed to distribute its power arbitrarily over all frequency subbands and use these simultaneously. It is assumed that whether the jammer inserts, modifies or jams messages only depends on the relation of its own and the sender's power. A protocol is found which achieves a positive throughput whose value depends on the jammer's strategies, e.g. whether or not it can listen to the sender's signals.

We take a different perspective in this work. The central figure of merit for our communication system is the message transmission error incurred under a jamming attack. A good FH protocol should make this error small. We assume that the jammer cannot listen to symbols sent through the channel (this in particular differs from [6]), that it knows the channel and the code, but not the specific message sent, and that it knows when the transmission of a new codeword begins. It can input symbols into any frequency subset of a given size. We also assume that the receiver listens to all frequencies simultaneously.

Within these boundaries, any jammer strategy is allowed. The jammer is successful if no coding strategy can be found making the transmission error vanish with increasing coding blocklength for any jamming strategy. This is an operational approach to measure the success of jamming, in contrast to the approach of [6] described above.

Using the information-theoretic model of an additive Arbitrarily Varying Channel (AVC) and the analysis in [2], we find that the success of a jammer indeed depends on the relation between its own and the sender's power. In fact, if the sender power is strictly larger than the jammer power, the same, positive capacity is achieved as in the case where sender and receiver have access to common randomness which is unknown to the jammer. If the converse relation between sender and jammer power holds, then no data transmission at all is possible. This is independent of the number $J$ of subchannels the jammer can influence at the same time.

On the other hand, it is known that for each frequency subband the same holds: If the jammer has more power than the sender, no communication is possible over this band, whereas the common randomness assisted capacity is achieved in case the sender power exceeds the jammer power. Thus in the case that no single frequency subband has a positive capacity without common randomness, then no FH scheme achieves a positive capacity either. Seen from this perspective, FH does not provide any additional protection against jamming compared to schemes which stick to one single frequency. However, FH does in general increase the common randomness assisted capacity compared to the use of one single

subchannel, and hence also the capacity without common randomness if positive – the FH sequence may depend on the message and thus reveal additional information. (In [9], [8] this is called message-driven frequency hopping.)

The common randomness assisted capacity will in general depend on the number $J$ of subchannels the jammer can simultaneously influence. Thus the capacity achievable without common randomness, if positive, also depends on $J$. We give a lower bound for the common randomness assisted capacity. If the noise is Gaussian and $J$ is sufficiently large, we also provide an upper bound which differs from the lower bound by the logarithm of the number of frequency subbands. The bounds involve a waterfilling strategy for the distribution of the jammer's power over the frequencies.

Due to space limitations, some parts of the proofs will be omitted or only sketched. The full version of this paper is available online [7].

*Organization of the paper:* Section II presents the channel model and the main results. Sections III-IV contain the proofs of these results. A discussion concludes the paper in Section V.

## II. System model and main results

The total frequency band available for communication is divided into $K$ frequency subbands. These are modeled as parallel channels with additive noise. The receiver listens to all frequencies simultaneously. Frequency hopping (FH) means that the sender at each time instant chooses one of the $K$ subchannels into which it inputs a signal. For a fixed number $J$ with $1 \leq J \leq K$, the jammer can at each time instant choose a subset $\mathcal{I}$ of the $K$ subchannels with $|\mathcal{I}| = J$ and input its own signals in subchannels belonging to this subset.

The overall channel, called FH channel in the following, can be described as an additive Arbitrarily Varying Channel (AVC) with additive noise. For any $k \in \mathcal{K} = \{1, \ldots, K\}$, we set $(e_{k1}, \ldots, e_{kK})^\top = \mathbf{e}_k$ to be the vector with $e_{kk} = 1$ and $e_{kl} = 0$ for $l \neq k$. Further for any $\mathcal{I}$ with $|\mathcal{I}| = J$, we set $(e_{\mathcal{I},1}, \ldots, e_{\mathcal{I},K})^\top = \mathbf{e}_\mathcal{I}$ to be the vector satisfying $e_{\mathcal{I},l} = 1$ if $l \in \mathcal{I}$ and $e_{\mathcal{I},l} = 0$ else.

If the sender chooses symbol $x \in \mathbb{R}$ to transmit over subchannel $k$, it inputs $x\mathbf{e}_k$ into the channel. We denote the set $\mathbb{R} \times \mathcal{K}$ by $\mathcal{X}$. The jammer choooses a subset $\mathcal{I} \subset \mathcal{K}$ of subchannels for possible jamming ($|\mathcal{I}| = J$) and a vector $(s_1, \ldots, s_K)^\top = \mathbf{s} \in \mathbb{R}^K$ of real numbers satisfying $s_l = 0$ if $l \notin \mathcal{I}$. Then it inputs $\mathbf{s} \circ \mathbf{e}_\mathcal{I}$ into the channel, where the symbol $\circ$ denotes component-wise multiplication. We denote the set of possible jammer choices by $\mathcal{S}$.

The noise on different frequencies is assumed to be independent. For subchannel $k$, let $N_k$ be the noise random variable. Its mean is assumed to be zero and its variance is denoted by $\sigma_k^2$. The random vector $(N_1, \ldots, N_K)^\top$ is denoted by $\mathbf{N}$.

Given sender input $x\mathbf{e}_k$ and jammer input $\mathbf{s} \circ \mathbf{e}_\mathcal{I}$, the receiver obtains a real $K$-dimensional output vector $(y_1, \ldots, y_K)^\top = \mathbf{y}$ through the FH channel which satisfies

$$\mathbf{y} = x\mathbf{e}_k + \mathbf{s} \circ \mathbf{e}_\mathcal{I} + \mathbf{N}.$$

In particular, on frequencies without sender or jammer inputs, the output is pure noise. The channel is memoryless over time, i.e. outputs at different time instants are independent conditional on the sender and jammer inputs. Note that this is an additive AVC, but as its input alphabet is a strict subset of $\mathbb{R}^K$, the special results of [2] on additive-noise AVCs do not apply here. The general theory developed in [2] is applicable, though: All alphabets involved are complete, separable metric spaces[1], the channel output distribution continuously depends on the sender and jammer inputs, and the constraints on sender and jammer inputs to be defined below are continuous. Hence the central hypotheses (H.1)-(H.4) of [2] are satisfied.

The protocols used for data transmission are block codes. A blocklength-$n$ code is defined as follows. We assume without loss of generality that the set of messages $\mathcal{M}_n$ is the set $\{1, \ldots, |\mathcal{M}_n|\}$. An encoder is a mapping $f_n$ from $\mathcal{M}_n$ into the set of sequences of sender channel inputs of length $n$,

$$\{(x_1\mathbf{e}_{k_1}, \ldots, x_n\mathbf{e}_{k_n}) : (x_i, k_i) \in \mathcal{X}\ (1 \leq i \leq n)\}.$$

Note that this means that the sequence of frequency bands used by the sender may depend on the message to be sent. Every codeword can be considered as a $K \times n$-matrix whose $i$-th column is the $i$-th channel input vector. The decoder at blocklength $n$ is a mapping $\varphi_n : \mathbb{R}^{K \times n} \longrightarrow \mathcal{M}_n$.

Additionally, for some $\Gamma > 0$, the sender has the power constraint $\sum_{i=1}^n \|f_n(m)_i\|^2 \leq n\Gamma$ for all $m \in \mathcal{M}_n$, where $f_n(m)_i$ denotes the $i$-th column of the $K \times n$-matrix $f_n(m)$ and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^K$. A code $(f_n, \varphi_n)$ with blocklength $n$ which satisfies the power constraint for $\Gamma$ is called an $(n, \Gamma)$-code.

We are interested in the transmission error incurred by a code $(f_n, \varphi_n)$. This error should be small for all possible jammer input sequences. Thus we first define the transmission error for a given length-$n$ jamming sequence $((\mathcal{I}_1, \mathbf{s}_1), \ldots, (\mathcal{I}_n, \mathbf{s}_n))$. This sequence can be given matrix form as well. We denote by $\tilde{S}$ the $K \times n$-matrix whose $i$-th column equals $\mathbf{s}_i$. By $\tilde{E} \in \mathbb{R}^{K \times n}$, we denote the matrix with columns $\mathbf{e}_{\mathcal{I}_1}, \ldots, \mathbf{e}_{\mathcal{I}_n}$. Of course, $\tilde{S} \circ \tilde{E} = \tilde{S}$. We keep $\tilde{E}$ explicit because $\tilde{S}$ itself does not in general uniquely determine the sequence $(\mathcal{I}_1, \ldots, \mathcal{I}_n)$, as some components of $\mathbf{s}_i$ could be zero $(1 \leq i \leq n)$.

Just like the sender, the jammer has a power constraint. We require that $\sum_{i=1}^n \|\mathbf{s}_i\|^2 \leq n\Lambda$ for some $\Lambda > 0$ and denote the set of $\tilde{S} \circ \tilde{E}$ satisfying this power constraint by $\mathcal{J}_\Lambda$. It is clear that a realistic jammer cannot transmit at arbitrarily large powers, so this is a reasonable assumption. Note that the jammer is free to distribute its power over the subchannel subset it has chosen for jamming. In particular, the power can be concentrated on one single frequency no matter what $J$ is.

Now let $(f_n, \varphi_n)$ be a blocklength-$n$ code and $\tilde{S} \circ \tilde{E} \in \mathbb{R}^{K \times n}$ a jammer input. Then the average error incurred by

[1] Giving a discrete set $\mathcal{K}$ the metric $\rho(k, l) = 1$ if $k \neq l$ and $\rho(k, k) = 0$ for all $k, l \in \mathcal{K}$ makes $\mathcal{K}$ a complete metric space whose Borel algebra is its complete power set.

$(f_n, \varphi_n)$ under this jamming sequence is defined to equal

$$\bar{e}(f_n, \varphi_n, \tilde{S} \circ \tilde{E}) = \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \mathbb{P}[\varphi_n(f_n(m) + \tilde{S} \circ \tilde{E} + \tilde{N}) \neq m],$$

where $\tilde{N}$ is a matrix whose columns are $n$ independent copies of the noise random vector $\mathbf{N}$. The overall transmission error for $(f_n, \varphi_n)$ under jammer power constraint $\Lambda$ is given by

$$\bar{e}(f_n, \varphi_n, \Lambda) = \sup_{\tilde{S} \circ \tilde{E} \in \mathcal{J}_\Lambda} \bar{e}(f_n, \varphi_n, \tilde{S} \circ \tilde{E}).$$

This error criterion makes the FH channel an AVC.

A nonnegative real number is said to be an *achievable rate* under sender power constraint $\Gamma$ and jammer power constraint $\Lambda$ if there exists a sequence of codes $((f_n, \varphi_n))_{n=1}^{\infty}$, where $(f_n, \varphi_n)$ is an $(n, \Gamma)$-code, satisfying

$$\liminf_{n \to \infty} \frac{1}{n} \log |\mathcal{M}_n| \geq R, \quad \lim_{n \to \infty} \bar{e}(f_n, \varphi_n, \Lambda) = 0.$$

The supremum $C(\Gamma, \Lambda)$ of the set of achievable rates under power constraints $\Gamma$ and $\Lambda$ is called the $(\Gamma, \Lambda)$-*capacity* of the channel.

Now we ask under which conditions the $(\Gamma, \Lambda)$-capacity of the FH channel is positive, and in case it is positive, how large it is. A precise statement can be made upon introduction of the common randomness assisted capacity $C_r(\Gamma, \Lambda)$. This is the maximal rate achievable if sender and receiver have a common secret key unknown to the jammer. The key size is not restricted. As noted in the introduction, the presence of a certain amount of common randomness is a frequent assumption in the literature on frequency hopping.

For given power constraint $\Gamma > 0$, we describe a common randomness assisted $(n, \Gamma)$-code as a random variable $(F_n, \Phi_n)$ on the set of $(n, \Gamma)$-codes with common message size and $(F_n, \Phi_n)$ independent of channel noise. The error it incurs under jamming sequence $\tilde{S} \circ \tilde{E}$ is defined to equal the mean $\mathbb{E}[\bar{e}(F_n, \Phi_n, \tilde{S} \circ \tilde{E})]$ over all possible realizations of $(F_n, \Phi_n)$, and the overall transmission error under jammer power constraint $\Lambda > 0$ is set to equal $\sup_{\tilde{S} \circ \tilde{E} \in \mathcal{J}_\Lambda} \mathbb{E}[\bar{e}(F_n, \Phi_n, \tilde{S} \circ \tilde{E})]$. The definition of common randomness assisted achievable rate under power constraints $\Gamma$ and $\Lambda$ is now a straightforward extension of the corresponding notion for the deterministic case. The supremum of all common randomness assisted rates under power constraints $\Gamma$ and $\Lambda$ is called the common randomness assisted $(\Gamma, \Lambda)$-capacity and denoted by $C_r(\Gamma, \Lambda)$.

**Theorem 1.** *$C(\Gamma, \Lambda)$ is positive if and only if $\Gamma > \Lambda$. If it is positive, it equals $C_r(\Gamma, \Lambda)$.*

**Corollary.** *1) If $C(\Gamma, \Lambda) > 0$, then every fixed-frequency subchannel also has a positive capacity. In this sense FH is not necessary to achieve a positive rate.*

*2) If $C(\Gamma, \Lambda) > 0$, then common randomness does not increase the maximal transmission rate.*

For $\Gamma > \Lambda$, it is thus desirable to have bounds on $C_r(\Gamma, \Lambda)$. These can be provided for all pairs $(\Gamma, \Lambda)$. Note that the choice of $\Lambda_1, \ldots, \Lambda_K$ below is a waterfilling strategy.

**Theorem 2.** *1) Let $\Lambda_1, \ldots, \Lambda_K$ be nonnegative numbers satisfying*

$$\begin{cases} \sigma_k^2 + \Lambda_k = c & \text{if } \sigma_k^2 < c, \\ \Lambda_k = 0 & \text{if } \sigma_k^2 \geq c \end{cases}$$

*with $c$ such that $\Lambda_1 + \cdots + \Lambda_K = \Lambda$. Then*

$$C_r(\Gamma, \Lambda) \geq \frac{1}{2} \log \left(1 + \frac{\Gamma}{c}\right). \tag{1}$$

*In particular, $C_r(\Gamma, \Lambda) > 0$.*

*2) If the noise is Gaussian and $J \geq |\{k \in \mathcal{K} : \sigma_k^2 < c\}|$, then*

$$C_r(\Gamma, \Lambda) \leq \frac{1}{2} \log \left(1 + \frac{\Gamma}{c}\right) + \log K. \tag{2}$$

*Remark.* 1) Set $\mathcal{K}' := \{k \in \mathcal{K} : \sigma_k^2 < c\}$. As comparison with (2) shows, (1) is a good bound if $J \geq |\mathcal{K}'|$ and the noise is Gaussian. The lack of a similar bound for the case $J < |\mathcal{K}'|$ can be explained by the fact that the jammer in this case has to leave some of the highest-throughput subchannels unjammed. $C_r(\Gamma, \Lambda)$ in general depends on $J$, and should increase for decreasing $J$.

2) The proof of Theorem 2 shows that the $\frac{1}{2} \log(1 + \frac{\Gamma}{c})$ terms in (1), (2) are achievable without frequency hopping, whereas frequency hopping contributes at most $\log K$ bits to capacity. According to the lower bound, the common randomness assisted capacity grows to infinity as $\Lambda$ is kept fixed and $\Gamma$ tends to infinity. Thus asymptotically for large $\Gamma$, the relative contribution to $C_r(\Gamma, \Lambda)$ of information transmitted through the FH sequence vanishes.

3) Non-trivial frequency hopping will in general be necessary both to achieve $C_r(\Gamma, \Lambda)$ and $C(\Gamma, \Lambda)$. Although we will not prove this, this is implied by the mutual information characterization of $C_r(\Gamma, \Lambda)$ (see the proof of Theorem 2).

## III. Proof of Theorem 2

Although Theorem 1 and its corollary are our main results, we first prove Theorem 2, which is needed for the proof of Theorem 1. From [2, Theorem 4] it follows that

$$C_r(\Gamma, \Lambda) = \min_{\substack{(\iota, \mathbf{S}): \\ \mathbb{E}[\|\mathbf{S}\|^2] \leq \Lambda}} \sup_{\substack{(X, \kappa): \\ \mathbb{E}[X^2] \leq \Gamma}} I(X\mathbf{e}_\kappa; X\mathbf{e}_\kappa + \mathbf{S} \circ \mathbf{e}_\iota + \mathbf{N}).$$

Here $X\mathbf{e}\kappa$ is a random variable on the possible sender inputs determined by an $\mathcal{X}$-valued random pair $(X, \kappa)$. Similarly, $\mathbf{S} \circ \mathbf{e}_\iota$ is the jammer's random channel input determined by a random $\mathcal{S}$-valued pair $(\iota, \mathbf{S})$ independent of $(X, \kappa)$.

Define $\mathbf{Y} = X\mathbf{e}_\kappa + \mathbf{S} \circ \mathbf{e}_\iota + \mathbf{N}$. The expression $I(X\mathbf{e}_\kappa; \mathbf{Y})$ is concave in the distribution $P_\kappa$ of $\kappa$ and convex in the distribution $P_\iota$ of $\iota$. Therefore the sender will in general have to use frequency hopping to approach capacity and likewise, the jammer will not stick to one constant frequency subset $\mathcal{I}$ for jamming.

The mutual information term appearing in the above formula for $C_r(\Gamma, \Lambda)$ can be written as

$$\begin{aligned} I(X\mathbf{e}_\kappa; \mathbf{Y}) &= I(X\mathbf{e}_\kappa, \kappa; \mathbf{Y}) - I(\kappa; \mathbf{Y}|X\mathbf{e}_\kappa) \\ &= I(X; \mathbf{Y}|\kappa) + I(\kappa; \mathbf{Y}), \end{aligned} \tag{3}$$

upon application of the chain rule in each of the equalities and observing that the sequence $\kappa \leftrightarrow X\mathbf{e}_\kappa \leftrightarrow \mathbf{Y}$ is Markov.

The second term in (3) is between 0 and $\log K$. Thus to bound $C_r(\Gamma,\Lambda)$, it remains to bound

$$\min_{(\iota,\mathbf{S}):\mathbb{E}[\|\mathbf{S}\|^2]\le\Lambda} \sup_{(X,\kappa):\mathbb{E}[X^2]\le\Gamma} I(X;\mathbf{Y}|\kappa) \tag{4}$$

$$= \min_{\substack{(\iota,\mathbf{S}):\\ \mathbb{E}[\|\mathbf{S}\|^2]\le\Lambda}} \sup_{(\kappa,\mathbf{\Gamma})} \sum_{k=1}^{K} P_\kappa(k) \sup_{X:\mathbb{E}[X^2|\kappa=k]\le\Gamma_k} I(X;\mathbf{Y}|\kappa=k),$$

where the supremum over $(\kappa,\mathbf{\Gamma})$ is over $\kappa$ and nonnegative vectors $\mathbf{\Gamma}=(\Gamma_1,\dots,\Gamma_K)$ satisfying $\sum P_\kappa(k)\Gamma_k \le \Gamma$.

We now sketch the proof of the lower bound. For any $k\in\mathcal{K}$, one has $I(X;\mathbf{Y}|\kappa=k) \ge I(X;Y_k|\kappa=k)$. Fix any $(\iota,\mathbf{S})$ with $\mathbb{E}[\|\mathbf{S}\|^2]\le\Lambda$. Then the $k$-th coordinate output conditional on the event $\kappa=k$ has the form

$$y_k = x + Z_k, \tag{5}$$

where $Z_k$ is a real-valued random variable of variance $\sigma_k^2+\Lambda_k$ for some nonnegative $\Lambda_1,\dots,\Lambda_K$ summing up to at most $\Lambda$. As (5) is an additive channel with the real numbers as input and output alphabet, it is a well-known fact [5, Theorem 7.4.3] that

$$\sup_{X:\mathbb{E}[X^2|\kappa=k]\le\Gamma_k} I(X;Y_k|\kappa=k) \ge \frac{1}{2}\log\left(1+\frac{\Gamma_k}{\sigma_k^2+\Lambda_k}\right).$$

Hence the right-hand side of (4) can be lower-bounded by

$$\min_{\mathbf{\Lambda}} \max_{(\kappa,\mathbf{\Gamma})} \frac{1}{2}\sum_{k=1}^{K} P_\kappa(k)\log\left(1+\frac{\Gamma_k}{\sigma_k^2+\Lambda_k}\right), \tag{6}$$

where the minimum is over vectors $\mathbf{\Lambda}=(\Lambda_1,\dots,\Lambda_K)$ with nonnegative components satisfying $\Lambda_1+\cdots+\Lambda_K\le\Lambda$. It is now straightforward to show that waterfilling for the jammer is the optimal choice of $\mathbf{\Lambda}$. This lower bound on (4) together with (3) proves (1). The proof of the upper bound is omitted due to space limitations. The full proof of Theorem 2 can be found in [7].

## IV. Proof of Theorem 1 and its corollary

The proof of Theorem 1 bases on the sufficient criterion for $C(\Gamma,\Lambda)=C_r(\Gamma,\Lambda)$ provided by the corollary to [2, Theorem 4]. To formulate this criterion, we first have to say what it means for the FH channel to be *symmetrized* by a stochastic kernel.

A stochastic kernel $U$ with inputs from $\mathcal{X}$ and outputs in $\mathcal{S}$ gives, for every $(x,k)\in\mathcal{X}$, a probability measure $U(\cdot|x,k)$ on the Borel algebra of $\mathcal{S}$ such that for every Borel-measurable $\mathcal{A}\subset\mathcal{S}$, the mapping $(x,k)\mapsto U(\mathcal{A}|x,k)$ is measurable. $U(\cdot|x,k)$ is specified by its values on all pairs $(\mathcal{I},\mathcal{B})$, where $|\mathcal{I}|=J$ and $\mathcal{B}$ is a Borel set on $\mathbb{R}^K$ such that for all $\mathbf{b}\in\mathcal{B}$, it holds that $l\notin\mathcal{I}$ implies $b_l=0$. One can thus write

$$U(\mathcal{I},\mathcal{B}|x,k) = U_1(\mathcal{I}|x,k)U_2(\mathcal{B}|x,k,\mathcal{I}).$$

$U_1(\cdot|x,k)$ determines a random variable $\iota^U(x,k)$ on the set of subsets of $\mathcal{K}$ with cardinality $J$. $U(\cdot|x,k)$ then determines a random variable $\mathbf{S}^U(x,k)$ which, conditional on the event $\iota^U(x,k)=\mathcal{I}$, has the distribution $U_2(\cdot|x,k,\mathcal{I})$. These random variables give rise to a random jammer input, $Z^U_{x,k} := \mathbf{S}^U(x,k)\circ\mathbf{e}_{\iota^U(x,k)}$. Thus any pair $(x',k')\in\mathcal{X}$ together with $U$ defines the following channel:

$$\mathbf{y} = x\mathbf{e}_k + \mathbf{Z}^U_{x',k'} + \mathbf{N},$$

where $(x,k)\in\mathcal{X}$ is the sender input, the output set is $\mathbb{R}^K$, and the noise is $\mathbf{Z}^U_{x',k'}+\mathbf{N}$.

By definition, the FH channel is *symmetrized* by $U$ if all sender input pairs $(x,k)$ and $(x',k')$ satisfy

$$x\mathbf{e}_k + \mathbf{Z}^U_{x',k'} + \mathbf{N} \overset{\mathcal{D}}{=} x'\mathbf{e}_{k'} + \mathbf{Z}^U_{x,k} + \mathbf{N},$$

where $\overset{\mathcal{D}}{=}$ means that the left-hand and the right-hand side have the same distribution. In particular, as the noise is mean-zero, this implies

$$x\mathbf{e}_k + \mathbb{E}\big[\mathbf{Z}^U_{x',k'}\big] = x'\mathbf{e}_{k'} + \mathbb{E}\big[\mathbf{Z}^U_{x,k}\big]. \tag{7}$$

To state the criterion for the equality of the $(\Gamma,\Lambda)$-capacities with and without common randomness, some more definitions are necessary. Let $\mathcal{U}_0$ be the class of stochastic kernels $U$ that symmetrize the FH channel and for which $\mathbf{Z}^U_{x,k}$ has finite variance for all $(x,k)$. Let $\tilde{\mathcal{X}}\subset\mathcal{X}$ be finite and $(X,\kappa)$ be concentrated on $\tilde{\mathcal{X}}$. Assume that for every $(x,k)\in\mathcal{X}$, the conditional distribution of the random variable $Z^U_{X,\kappa}$ given $\{X=x,\kappa=k\}$ equals that of $Z^U_{x,k}$. Then define

$$\tau_{\tilde{\mathcal{X}}}(X,\kappa,\Lambda) = \frac{1}{\Lambda}\inf_{U\in\mathcal{U}_0}\mathbb{E}\big[\|\mathbf{Z}^U_{X,\kappa}\|^2\big].$$

We also write $C_{r,\tilde{\mathcal{X}}}(\Gamma,\Lambda)$ for the common randomness assisted capacity of the FH channel with the same power constraints, but whose inputs are restricted to the finite subset $\tilde{\mathcal{X}}$ of $\mathcal{X}$.

By the corollary of [2, Theorem 4], $C(\Gamma,\Lambda)=C_r(\Gamma,\Lambda)$ if there exists a family $\mathcal{F}$ of finite subsets of $\mathcal{X}$ satisfying that every finite subset of $\mathcal{X}$ is contained in some member of $\mathcal{F}$ and that for every $\tilde{\mathcal{X}}\in\mathcal{F}$, there is an $(X,\kappa)$ concentrated on $\tilde{\mathcal{X}}$ and satisfying $\mathbb{E}[X^2]\le\Gamma$ with $I(X\mathbf{e}_\kappa;\mathbf{Y})=C_{r,\tilde{\mathcal{X}}}(\Gamma,\Lambda)$ and $\tau_{\tilde{\mathcal{X}}}(X,\kappa,\Lambda)>1$.

We will now closely follow the proof of [2, Theorem 5] to prove that the above criterion is satisfied for the FH channel if $\Gamma>\Lambda$. Fix $\Gamma,\Lambda>0$. Let $\tilde{\mathcal{X}}_0$ be a finite set satisfying $C_{r,\tilde{\mathcal{X}}_0}(\Gamma',\Lambda)>C_r(\Gamma,\Lambda)$ for some $\Gamma'>\Gamma$. Such a set exists by the fact ([2, Theorem 4]) that for all $\Gamma,\Lambda$,

$$C_r(\Gamma,\Lambda) = \sup_{\tilde{\mathcal{X}}\subset\mathcal{X}\text{ finite}} C_{r,\tilde{\mathcal{X}}}(\Gamma,\Lambda)$$

and by the lower bound on $C_r(\Gamma,\Lambda)$ of Theorem 2 showing that $C_r(\Gamma,\Lambda)$ tends to infinity as $\Lambda$ is fixed and $\Gamma$ tends to infinity. We choose $\mathcal{F}$ as the family of finite subsets $\tilde{\mathcal{X}}$ of $\mathcal{X}$ satisfying $\tilde{\mathcal{X}}_0\subset\tilde{\mathcal{X}}$ and

$$\tilde{\mathcal{X}} = \bigcup_{k=1}^{K}\tilde{\mathcal{X}}_k\times\{k\},$$

where $\tilde{\mathcal{X}}_k$ is symmetric about the origin. Obviously, every finite subset of $\mathcal{X}$ is contained in some $\tilde{\mathcal{X}}\in\mathcal{F}$. We first need

that for every finite input set $\tilde{\mathcal{X}} \in \mathcal{F}$ there exist $C_{r,\tilde{\mathcal{X}}}(\Gamma,\Lambda)$-achieving channel input distributions which exhaust all the power and are symmetric on every frequency subband.

**Lemma.** *Let $\tilde{\mathcal{X}} \in \mathcal{F}$. Then there exists a pair $(X,\kappa)$ of random variables with values in $\tilde{\mathcal{X}}$ satisfying*

$$\min_{\substack{(\iota,\mathbf{S}):\\ \mathbb{E}[\|\mathbf{S}\|^2]\le\Lambda}} I(X\mathbf{e}_\kappa; X\mathbf{e}_\kappa + \mathbf{S}\circ\mathbf{e}_\iota + \mathbf{N}) = C_{r,\tilde{\mathcal{X}}}(\Gamma,\Lambda) \quad (8)$$

*and*

$$\mathbb{E}\big[X^2\big] = \Gamma, \quad (9)$$

$$P_{X|\kappa}(\cdot|k) = P_{-X|\kappa}(\cdot|k) \qquad (1 \le k \le K) \quad (10)$$

*Here $P_{X|\kappa}$ denotes the conditional probability of $X$ given $\kappa$, and $P_{-X|\kappa}$ is defined analogously.*

The proof of the lemma is very similar to that of (4.42) in [2] and omitted here, but can be found in [7]. Let $\tilde{\mathcal{X}} \in \mathcal{F}$ and $(X,\kappa)$ as in the Lemma. We now show that $\tau_{\tilde{\mathcal{X}}}(X,\kappa,\Lambda) > 1$ if $\Gamma > \Lambda$. To do so, choose any $U \in \mathcal{U}_0$. Then for any $(x',k') \in \mathcal{X}$, using Jensen's inequality,

$$\mathbb{E}\big[\|\mathbf{Z}^U_{X,\kappa}\|^2\big] \ge \sum_{(x,k)\in\tilde{\mathcal{X}}} P_{(X,\kappa)}(x,k)\,\big\|\mathbb{E}\big[\mathbf{Z}^U_{x,k}\big]\big\|^2. \quad (11)$$

As $U$ symmetrizes the FH channel, we can apply (7) and lower-bound (11) by

$$\sum_{(x,k)\in\tilde{\mathcal{X}}} P_{(X,\kappa)}(x,k)\,\big\|x\mathbf{e}_k - x'\mathbf{e}_{k'} + \mathbb{E}[\mathbf{Z}^U_{x',k'}]\big\|^2 \quad (12)$$

$$\ge \sum_k P_\kappa(k) \sum_{x\in\tilde{\mathcal{X}}_k} P_{X|\kappa}(x|k)|x - x' e_{k'k} + \mathbb{E}[Z^U_{x',k'}(k)]|^2,$$

where we denote by $Z^U_{x,k}(k)$ the $k$-th component of $\mathbf{Z}^U_{x,k}$. By (10), $P_{X|\kappa}(\cdot|k)$ is symmetric for every $k$, so its mean equals 0. Hence with (9) the right-hand side side of (12) can be lower-bounded by

$$\sum_{(x,k)\in\tilde{\mathcal{X}}} P(x,k)|x|^2 = \mathbb{E}[X^2] = \Gamma.$$

We conclude that $\tau_{\tilde{\mathcal{X}}}(X,\kappa,\Lambda) > 1$ for all $\tilde{\mathcal{X}} \in \mathcal{F}$ and the corresponding $(X,\kappa)$ if $\Gamma > \Lambda$, implying that $C(\Gamma,\Lambda) = C_r(\Gamma,\Lambda)$. As the common randomness assisted $(\Gamma,\Lambda)$-capacity is positive for positive $\Gamma$, this further implies that $C(\Gamma,\Lambda) > 0$ if $\Gamma > \Lambda$, and the proof of the direct part of Theorem 1 is complete. The proof of the converse is analogous to the proof of the converse of [3, Theorem 1], so we omit it here, but it can be found in [7].

The second claim of the corollary of Theorem 1 is obvious. The first statement follows from [2, Theorem 5], which says that an additive-noise channel with $\mathbb{R}$ as sender, jammer and output alphabet has positive capacity (then equal to the common randomness assisted capacity) if and only if the sender power exceeds the jammer power. So if both the sender and the jammer in the FH channel concentrate their power on any frequency band $k \in \mathcal{K}$ and $\Gamma > \Lambda$, already a positive capacity lower-bounded by $\log(1 + \frac{\Gamma}{\sigma_k^2+\Lambda})/2 > 0$ will be achievable. In particular, this rate can be obtained without frequency hopping. On the other hand, if no transmission is possible over the subchannels, then $\Gamma \le \Lambda$, and the FH channel also has zero capacity.

## V. Discussion

For non-discrete AVCs, there is no general statement that capacity without common randomness always equals 0 or the common randomness assisted capacity like the Ahlswede dichotomy in [1] for discrete AVCs. Thus it is not possible to justify Theorem 1 just by observing that the capacity of every subchannel is positive if $\Gamma > \Lambda$.

Like [9], [8] we assume here that the receiver simultaneously listens on all frequencies. A different approach is taken in [6], [4], where the receiver listens randomly on only one frequency band at a time. The above analysis can be performed in a similar way for this situation and leads to analogous results: The capacity without common randomness shared between sender and receiver is positive if and only if the sender power exceeds the jammer power. Of course, the capacity will in general be smaller than if the receiver listens on all frequencies.

The converse in [3, Theorem 1] shows that in order to find a good jamming sequence, the jammer needs knowledge of the channel and the transmission protocol. Further, it should know when the transmission of a codeword starts, so it has to be synchronized with the sender. If this is given, then the successful jamming strategy in the case $\Gamma \le \Lambda$ is to confuse the receiver: There exists a legitimate codeword such that if the jammer inputs this into the FH channel, the receiver cannot distinguish the sender's messages.

The case of a jammer listening to the sender's input into the channel like in [6], [4] was not treated here because there exist few results on AVCs in this direction.

## References


[1] R. Ahlswede. Elimination of correlation in random codes for arbitrarily varying channels. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 44:159–175, 1978.

[2] I. Csiszar. Arbitrarily varying channels with general alphabets and states. *Information Theory, IEEE Transactions on*, 38(6):1725–1742, Nov 1992.

[3] I. Csiszar and P. Narayan. Capacity of the gaussian arbitrarily varying channel. *Information Theory, IEEE Transactions on*, 37(1):18–26, Jan 1991.

[4] Y. Emek and R. Wattenhofer. Frequency hopping against a powerful adversary. In Y. Afek, editor, *Distributed Computing*, volume 8205 of *LNCS*, pages 329–343. Springer, 2013.

[5] R. G. Gallager. *Information theory and reliable communication*. Wiley, New York, 1968.

[6] M. Strasser, C. Popper, S. Capkun, and M. Cagalj. Jamming-resistant key establishment using uncoordinated frequency hopping. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 64–78, May 2008.

[7] M. Wiese and P. Papadimitratos. Frequency hopping does not increase anti-jamming resilience of wireless channels. Available online at http://arxiv.org/abs/1512.06645, December 2015.

[8] L. Zhang and T. Li. Anti-jamming message-driven frequency hopping—part ii: Capacity analysis under disguised jamming. *Wireless Communications, IEEE Transactions on*, 12(1):80–88, January 2013.

[9] L. Zhang, H. Wang, and T. Li. Anti-jamming message-driven frequency hopping—part i: System design. *Wireless Communications, IEEE Transactions on*, 12(1):70–79, January 2013.

# Tracking Unstable Autoregressive Sources over Discrete Memoryless Channels

Roy Timo
Technische Universität München
roy.timo@tum.de

Badri N. Vellambi
New Jersey Institute of Technology
badri.n.vellambi@ieee.org

Alex Grant
Cohda Wireless
alex.grant@cohdawireless.com

Khoa D. Nguyen
University of South Australia
khoa.nguyen@unisa.edu.au

***Abstract*—We consider the problem of tracking, in realtime, an unstable autoregressive (AR) source over a discrete memoryless channel (DMC). We present computable achievable bounds on the optimal tracking error for general DMCs, and we particularise these bounds to the binary erasure and packet erasure channels.**

## I. Problem Setup

Consider the scalar *unstable AR* source

$$W_n := \lambda W_{n-1} + V_n, \quad n = 1, 2, \dots, \tag{1}$$

where $W_0 = 0$, $\lambda > 1$ and $V_1, V_2, \dots$ is iid and uniform on the real interval $[-v_{\max}, v_{\max}]$ for some positive $v_{\max}$. Suppose that a transmitter causally encodes and communicates (1) over a DMC to a receiver, and suppose that the receiver attempts to track (1) while operating with a finite decoding delay $\Delta$.

More formally, let us fix $\Delta$ to be any non-negative integer. Let $\mathcal{X}$ denote the DMC's input alphabet, $\mathcal{Y}$ its output alphabet and $T_{Y|X} : \mathcal{X} \to \mathcal{Y}$ its transition probabilities. A $\Delta$-*code* for tracking (1) consists of a sequence of mappings

$$(F_1, G_1), (F_2, G_2), \dots,$$

where $F_n : \mathbb{R}^n \to \mathcal{X}$ and $G_n : \mathcal{Y}^{n+\Delta} \to \mathbb{R}$. The $n$-th channel symbol sent by the transmitter over the channel is

$$X_n := F_n(W_1, W_2, \dots, W_n),$$

The receiver estimates the $n$-th source symbol $W_n$ by

$$\hat{W}_n := g_n(Y_1, Y_2, \dots, Y_{n+\Delta}).$$

Fix $\rho \geq 1$, and let

$$\mathsf{ME}_\rho(\Delta, N) := \mathbb{E}\, \frac{1}{N} \sum_{n=1}^{N} \left|\hat{W}_n - W_n\right|^\rho \tag{2}$$

denote the *mean error* in tracking. We call $N$ the *blocklength*, and we are interested in determining the following quantities.

*Definition 1:*

(i) The optimal $\mathsf{ME}_\rho$ for a given $\Delta$ and $N$,

$$\mathsf{ME}^*_\rho(\Delta, N) := \inf_{\Delta\text{-codes}} \mathsf{ME}_\rho(\Delta, N).$$

(ii) The optimal $\mathsf{ME}_\rho$ for a given $\Delta$ and *all* $N$,

$$\mathsf{ME}^*_\rho(\Delta) := \sup_{N \in \{1,2,\dots\}} \mathsf{ME}^*_\rho(\Delta, N).$$

The purpose of this paper is to report some useful achievable (upper) bounds on $\mathsf{ME}^*_\rho(\Delta, N)$ and $\mathsf{ME}^*_\rho(\Delta)$.

*Remark 1:* It can be shown that

$$\sup_{N \in \{1,2,\dots\}} \mathsf{ME}^*_\rho(\Delta, N) = \limsup_{N \to \infty} \mathsf{ME}^*_\rho(\Delta, N).$$

Consequently, $\mathsf{ME}^*_\rho(\Delta)$ measures the *worst case* tracking error as $N \to \infty$, and it is an appropriate engineering benchmark for problems where $N$ is large, varies or is otherwise unknown.

*Remark 2:* If $\Delta = 0$, then for each $n = 1, 2, \dots$ the receiver is required to output its estimate $\hat{W}_n$ of $W_n$ immediately upon observing the first $n$ channel outputs $Y_1, Y_2, \dots, Y_n$. Here we have *instantaneous communications* in the sense that any "new information" in $W_n$ can only be communicated over the channel using $X_n$. If $\Delta > 0$, then the receiver delays making an estimate of $W_n$ by $\Delta$ channel symbols. The transmitter now has more channel symbols from which to communicate each source symbol, and the receiver's estimates can therefore be improved. In essence, one can trade *tracking reliability* against *decoding timeliness* by varying $\Delta$.

*Remark 3:* The reader will have noticed that the unstable AR source in (1) does not fit within classical rate-distortion (RD) theory [1] (e.g., the AR source is non-stationary and non-ergodic), and $\Delta$-codes and Definition 1 do not fit within the classical joint source-channel coding framework [2, Sec. 9.6]. Indeed, channels with the same capacity often behave quite differently under Definition 1, and the standard RD function, channel capacity and separation theorem offer little guidance on how to best approximate $\mathsf{ME}^*_\rho(\Delta, N)$ and $\mathsf{ME}^*_\rho(\Delta)$. Such behaviour has been observed throughout the realtime communications literature and is by no means unique to this paper, see, for example, the early work of [3]. A more complete literature review can be found in the longer version of this paper [4], and excellent literature reviews of realtime communications can be found in [5]–[7].

## II. Examples: Tracking over Erasure Channels

Let us first demonstrate the usefulness of this work by specialising the following achievable bound to the *binary erasure channel* (BEC) and *packet erasure channel* (PEC). (The extended paper [4] also considers the *binary symmetric channel*.) In this section, we restrict attention to *mean absolute errors* (MAEs); that is, we fix $\rho = 1$ in (2).

*Theorem 1:*

$$\mathsf{ME}^*_\rho(\Delta, N)\Big|_{\rho=1} \leq \min_M \big[\alpha(M) + \beta(M)\big],$$

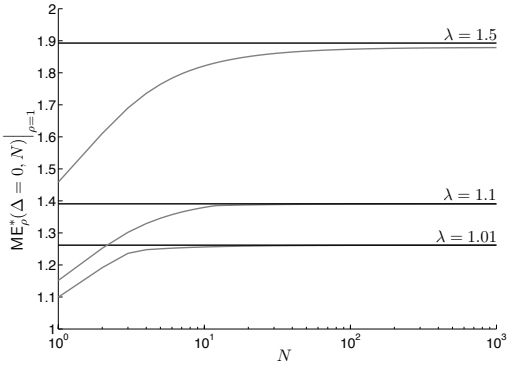



Fig. 1. The achievable bound on $\mathsf{ME}^*_\rho(\Delta, N)|_{\rho=1}$ given in Proposition 1 (grey lines) and the achievable bound on $\mathsf{ME}^*_\rho(\Delta)|_{\rho=1}$ given in Proposition 2 (black lines). The bounds are plotted for $\lambda = 1.01, 1.1$ and $1.5$ with bandwidth expansion $\kappa = 5$, zero decoding delay $\Delta = 0$ and $v_{\max} = 1$.

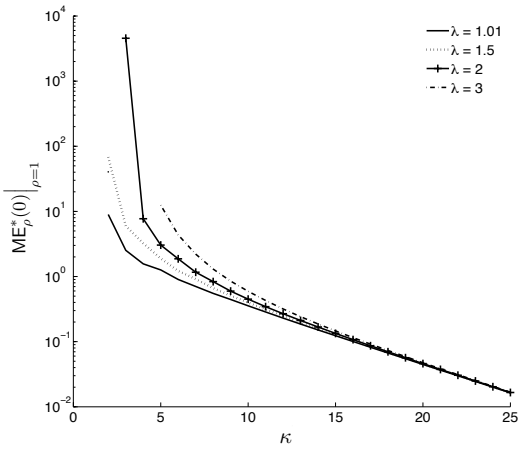



Fig. 2. The achievable bound in Proposition 2 for tracking (1) over a BEC with erasure probability $\epsilon = 0.1$ and bandwidth expansion $\kappa$. The bound is plotted as a function of $\kappa$ for $\lambda = 1.1, 1.5, 2$ and 3 with zero decoding delay $\Delta = 0$ and $v_{\max} = 1$.

where the minimisation is over all integers $M \geq \max\{2, \lambda\}$,

$$\alpha(M) := \frac{v_{\max}}{M - \lambda}$$

and

$$\beta(M) := \frac{2v_{\max}(M-1)}{N(M-\lambda)} \sum_{n=1}^{N} \sum_{k=1}^{n} \lambda^{n-k} \min\Big\{1, \sum_{i=0}^{k-1} \tau_M(n + \Delta - i)\Big\}.$$

Theorem 1 follows from Lemma 1 and Theorem 3, which will be presented later in the paper. The upper bound depends on the channel law $T_{Y|X}$ via the function $\tau_M(\cdot)$, and we particularise $\tau_M(\cdot)$ to the BEC and PEC in the next two subsections[1]. Intuitively, $\alpha(M)$ can be understood as the quantisation error associated with discretising the source (it is the average error induced by an adaptive $M$ level scalar quantiser); and $\beta(M)$ can be understood as the *channel distortion* associated with streaming this $M$-level discrete approximation over the DMC $T_{Y|X}$. In general, a small $M$ will induce a large quantiser error and a small channel distortion, while a large $M$ will induce a small quantisation error and a large channel distortion. We now particularise $\tau_M(\cdot)$ to the BEC and PEC.

### A. Binary Erasure Channel (BEC)

Suppose that the DMC consists of $\kappa$ independent BECs, each with the same erasure probability $0 < \varepsilon < 1$. Let

$$\mathcal{X} = \{0,1\}^\kappa \quad \text{and} \quad \mathcal{Y} = \{0, 1, \mathsf{e}\}^\kappa,$$

where $\mathsf{e}$ represents the erasure event; and

$$T_{Y|X}(y|x) = \begin{cases} \varepsilon^{\mathsf{N}(y)}(1-\varepsilon)^{\kappa - \mathsf{N}(y)} & \text{if } x \text{ and } y \text{ agree on unerased positions,} \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathsf{N}(y)$ denotes the number of erased symbols in $y$. We call $\kappa$ the *bandwidth expansion factor*.

*Proposition 1:* For the BEC with erasure probability $\epsilon$ and bandwidth expansion $\kappa$, Theorem 1 holds with

$$\tau_M(k) = \sum_{t=0}^{\kappa k} \binom{\kappa k}{t} \varepsilon^t (1-\varepsilon)^{\kappa k - t} 2^{-[(\kappa k - t) - k \log M - \log(M-1)]^+}.$$

*Proof:* A proof can be found in [4, Appendix F-B]. ■

A slight weakening of Proposition 1 yields the next proposition.

*Proposition 2:* For the BEC with erasure probability $\varepsilon$ and bandwidth expansion $\kappa$, we have

$$\mathsf{ME}^*_\rho(\Delta)\Big|_{\rho=1} \leq \min_M \Big[\alpha(M) + \gamma(M)\ 2^{-\Delta(\kappa \mathsf{R}_0(\varepsilon) - \log M)}\Big],$$

where

$$\mathsf{R}_0(\varepsilon) := 1 - \log(1+\varepsilon)$$

is the *cutoff rate*[2] of the BEC,

$$\gamma(M) := \left(\frac{2v_{\max}\ (M-1)^2\ M}{2^{\kappa \mathsf{R}_0(\varepsilon)}\ (M-\lambda)}\right)\left(\frac{1}{1 - M2^{-\kappa \mathsf{R}_0(\varepsilon)}}\right)\left(\frac{1}{1 - \lambda M 2^{-\kappa \mathsf{R}_0(\varepsilon)}}\right),$$

and the minimisation is taken over all integers $M$ satisfying

$$\max\{\lambda, 2\} \leq M < (1/\lambda)\ 2^{\kappa \mathsf{R}_0(\varepsilon)}.$$

*Proof:* A proof can be found in [4, Appendix F-B]. ■

The bounds in Propositions 1 and 2 are plotted in Figure 1 as a function of the blocklength $N$ for three different values

[1] We defer giving a formal definition of $\tau_M(\cdot)$ for general DMCs until Section IV, because this definition requires additional notation and ideas that are not needed for the BEC and PEC.

[2] With a slight abuse of terminology, we use *cutoff rate* to refer to the standard $\mathsf{R}_0$ parameter (see [12, Eqn. 14] or [11, p. 628]). The *operational cutoff rate*, which concerns the computational complexity of sequential decoding [12], does not appear to be related to this work.

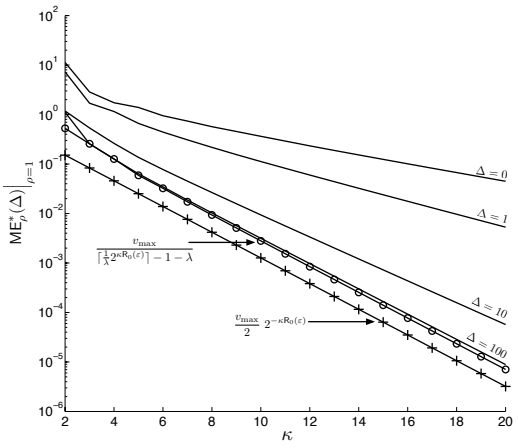



Fig. 3. The achievable bound in Proposition 2 for tracking (1) over a BEC with erasure probability $\epsilon = 0.1$ and bandwidth expansion $\kappa$. The bound is plotted as a function of $\kappa$ for $\lambda = 1.1, 1.5, 2$ and 3 with $\Delta = 0$ and $v_{\max} = 1$ fixed. The lower bound in Proposition 3 is also shown.

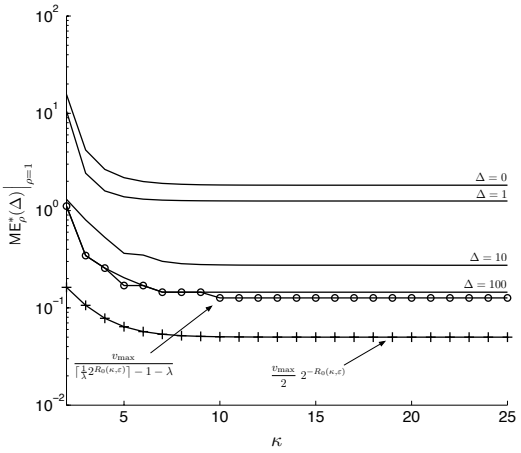



Fig. 4. The achievable bound on $\mathsf{ME}^*_\rho(\Delta)|_{\rho=1}$ in Proposition 5 for tracking (1) with $\lambda = 1.1$ and $v_{\max} = 1$ over a $\kappa$ bit packet erasure channel with erasure probability $\varepsilon = 0.1$. The bound is plotted as a function of $\kappa$ for decoding delays $\Delta = 0, 1, 10$ and 100. The lower bound in Proposition 6 is also shown.

of $\lambda$. Figure 2 plots the bound in Proposition 2 as a function of the bandwidth expansion factor $\kappa$ for four different values of $\lambda$. Figure 3 illustrates the role of decoding delay $\Delta$ in Proposition 2. As $\Delta \to \infty$ the bound tends to

$$\frac{v_{\max}}{\lceil \frac{1}{\lambda} 2^{\kappa \mathsf{R}_0(\varepsilon)} \rceil - 1 - \lambda} \approx \lambda v_{\max}\, 2^{-\kappa \mathsf{R}_0(\varepsilon)},$$

so the cutoff rate $\mathsf{R}_0(\varepsilon)$ governs its asymptotic accuracy. To help get a feel for how useful Proposition 2 is, we now give a simple lower bound. This lower bound is shown in Figure 3.

*Proposition 3:* For the BEC with erasure probability $\varepsilon$, bandwidth expansion $\kappa$ and zero decoding delay $\Delta = 0$,

$$\mathsf{ME}^*_\rho(0)\Big|_{\rho=1} \geq \frac{v_{\max}}{2}\, 2^{-\kappa \mathsf{R}_0(\varepsilon)}.$$

*Proof:* A proof can be found in [4, Appendix F-C]. ■

### B. Packet Erasure Channel (PEC)

Now imagine that the transmitter communicates with the receiver over a network that can be modelled[3] by a $\kappa$ bit PEC.

Fix $0 < \varepsilon < 1$, and let

$$\mathcal{X} := \{0,1\}^\kappa \quad \text{and} \quad \mathcal{Y} := \{0,1\}^\kappa \cup \{\mathsf{e}\}$$

and

$$T_{Y|X}(y|x) := \begin{cases} 1-\varepsilon & \text{if } y = x \\ \varepsilon & \text{if } y = \mathsf{e} \\ 0 & \text{otherwise.} \end{cases}$$

*Proposition 4:* For the $\kappa$ bit PEC with erasure probability $\varepsilon$, Theorem 1 holds with

$$\tau_M(k) = \sum_{t=0}^{k} \binom{k}{t} \varepsilon^t\, (1-\varepsilon)^{k-t}\, 2^{-[\kappa(k-t) - k \log M - \log(M-1)]^+}.$$

[3]Here we assume that bit-level errors within a packet are handled on a link-by-link basis using physical layer error-correction techniques, and packets arrive at the receiver promptly or they are lost to, for example, congestion and buffer overflows.

*Proof:* A proof can be found in [4, Appendix G-A]. ■

*Proposition 5:* For the $\kappa$ bit PEC with erasure probability $\varepsilon$,

$$\mathsf{ME}^*_\rho(\Delta)\Big|_{\rho=1} \leq \min_M \Big[\alpha(M) + \gamma(M)\, 2^{-\Delta(\mathsf{R}_0(\kappa,\varepsilon) - \log M)}\Big],$$

where

$$\gamma(M) := \left(\frac{2 v_{\max}\, (M-1)^2\, M}{2^{\mathsf{R}_0(\kappa,\varepsilon)}(M-\lambda)}\right)\left(\frac{1}{1 - M 2^{-\mathsf{R}_0(\kappa,\varepsilon)}}\right)\left(\frac{1}{1 - \lambda M 2^{-\mathsf{R}_0(\kappa,\varepsilon)}}\right),$$

the minimisation is over all integers $M$ satisfying

$$\max\{2, \lambda\} \leq M < (1/\lambda)\, 2^{\mathsf{R}_0(\kappa,\varepsilon)}$$

and

$$\mathsf{R}_0(\kappa,\varepsilon) = \kappa - \log\big(\varepsilon(2^\kappa - 1) + 1\big)$$

is the cutoff rate of the $\kappa$-bit PEC.

*Proof:* A proof can be found in [4, Appendix G-B]. ■

*Proposition 6:* For the $\kappa$ bit PEC with erasure probability $\varepsilon$ and zero decoding delay ($\Delta = 0$),

$$\mathsf{ME}^*_\rho(0)\Big|_{\rho=1} \geq \frac{v_{\max}}{2}\, 2^{-\mathsf{R}_0(\kappa,\varepsilon)}.$$

*Proof:* A proof can be found in [4, Appendix G-C]. ■

Figure 4 illustrates the role of decoding delay in Proposition 5. Each curve exhibits a tracking 'error floor' because

$$\lim_{\kappa \to \infty} \mathsf{R}_0(\kappa,\varepsilon) = -\log \varepsilon,$$

so $M$ cannot grow without bound in $\kappa$. Interestingly, this error floor is intrinsic to the problem, because, for example, it also appears in Proposition 6.

### C. Discussion

Propositions 1 to 6 give computable upper and lower bounds for tracking (1) over two basic erasure channels. To the best of our knowledge, no computable bounds have appeared before in the literature[4]. Although we do not believe that the bounds are tight, we do believe that they provide useful performance benchmarks for practitioners in code design. For example, Figure 3 illustrates that the optimal MAE for the BSC improves exponentially fast with the bandwidth expansion factor $\kappa$; therefore, one might consider optimising $\kappa$ against the symbol error probability $\varepsilon$ on a system level. While, on the other hand, Figure 4 suggests that it does not make sense to use large packets over a PEC (unless one can decrease the erasure probability with increasing packet length). Finally, Theorem 1 is proved using a method that partially separates source quantisation from channel coding. Separation is often necessary in practice, and these bounds illustrate what one might achieve with such an approach.

## III. Bounds on MAE via Channel Coding with an AR-Distortion Function

The upper bounds on $\mathsf{ME}^*_\rho(\Delta,N)|_{\rho=1}$ and $\mathsf{ME}^*_\rho(\Delta)|_{\rho=1}$ presented in Section II all follow from the method outlined in this section, which can be applied to any DMC and $\rho \geq 1$. The key idea here will be to partially separate channel coding from quantisation, and to optimise the channel code with respect to an *AR Hamming (ARH) distortion function* that is determined by the source statistics. This section extends our earlier work [13] on streaming discrete AR sources with a MAE criterion to continuous AR sources with $\rho \geq 1$ in (2).

Arbitrarily fix an integer $M \geq 2$, and suppose that a discrete memoryless source (DMS) emits a sequence $U_1, U_2, \ldots,$ of independent and uniformly distributed random variables on

$$\mathcal{U} := \{0, \ldots, M-1\}.$$

An $(M,\Delta)$-*channel code* for streaming $U_1, U_2, \ldots$ over the DMC $T_{Y|X}$ consists of a sequence of mappings

$$(f_1, g_1), (f_2, g_2), \ldots,$$

where

$$f_n : \mathcal{U}^n \to \mathcal{X} \quad \text{and} \quad g_n : \mathcal{Y}^{n+\Delta} \to \mathcal{U}^n.$$

The $n$-th symbol sent over the channel by the transmitter is

$$X_n := f_n(U_{[1,n]}),$$

where we let

$$U_{[1,n]} = (U_1, U_2, \ldots, U_n)$$

denote the first $n$ symbols output by the DMS (we will also employ this notation for other random vectors). The receiver estimates the first $n$ DMS symbols $U_{[1,n]}$ from the first $(n+\Delta)$-channel outputs $Y_{[1,n+\Delta]} = (Y_1, Y_2, \ldots, Y_{n+\Delta})$ by

$$\hat{U}^{(n)}_{[1,n]} := g_n(Y_{[1,n+\Delta]}).$$

[4]Previous works, e.g., [7]–[10], appear to exclusively focus on checking whether or not certain tracking error measures are finite for a given channel.

A key point here is that the receiver initially estimates the first $n$ DMS symbols $U_{[1,n]}$ immediately upon observing the first $(n+\Delta)$-channel outputs. The receiver then revisits and (hopefully) improves this estimate as more channel outputs become available. The speed at which these estimates improve can be partially quantified by the following *AR Hamming* (ARH) distortion function. This distortion function will be our doorway to computable bounds on $\mathsf{ME}_\rho{}^*(\Delta,N)$.

We define the ARH distortion between a source output $u_{[1,n]} \in \mathcal{U}^n$ and reconstruction $\hat{u}_{[1,n]} \in \mathcal{U}^n$ by

$$\mathsf{d}_n(\hat{u}_{[1,n]}, u_{[1,n]}) := \sum_{k=1}^{n} \lambda^{n-k} \mathbb{1}\{\hat{u}_{[1,k]} \neq u_{[1,k]}\},$$

where

$$\mathbb{1}\{\hat{u}_{[1,k]} \neq u_{[1,k]}\} := \begin{cases} 1 & \text{if } \hat{u}_{[1,k]} \neq u_{[1,k]} \\ 0 & \text{otherwise.} \end{cases}$$

For a given $(M,\Delta)$-code, let

$$\mathsf{D}_{\rho,M}(\Delta,N) := \mathbb{E}\ \frac{1}{N}\sum_{n=1}^{N} \Big(\mathsf{d}_n(\hat{U}^{(n)}_{[1,n]}, U_{[1,n]})\Big)^\rho.$$

We say that a distortion $D$ is $(\rho, M, \Delta, N)$-*achievable* if there exists an $(M,\Delta)$-channel code with $\mathsf{D}_{\rho,M}(\Delta,N) \leq D$. Let

$$\mathsf{D}^*_{\rho,M}(\Delta,N) := \min\Big\{D : D \text{ is } (\rho,M,\Delta,N)\text{-achievable}\Big\}$$

denote the optimal ARH distortion for a given $\Delta$ and $N$. The key result of this section is the next lemma, which demonstrates that any achievable bound on $\mathsf{D}^*_M(\Delta,N)$ automatically gives an achievable bound on $\mathsf{ME}^*_\rho(\Delta,N)$.

*Lemma 1:* For every integer $M \geq \max\{\lambda, 2\}$, we have

$$\mathsf{ME}^*_\rho(\Delta,N) \leq \Big(\frac{2v_{\max}}{M-\lambda}\Big)^\rho \Big(\frac{1}{2} + 2^{\rho-1}(M-1)^\rho\ \mathsf{D}^*_{\rho,M}(\Delta,N)\Big).$$

*Proof:* A proof can be found in [4, Appendix A]. ∎

## IV. Random Coding Union Bound on $\mathsf{D}^*_M(\Delta,N)$

We now present an achievable bound on $\mathsf{D}^*_M(\Delta,N)$ that is motivated by the *random coding union* bound for block codes [14, Thm. 17]. We need the following notation. Given a pair of discrete random variables $(A,B)$ on $\mathcal{A}\times\mathcal{B}$ with joint pmf $P_{AB}(a,b)$ and marginals $P_A(a)$ and $P_B(b)$, the *information density* $\imath_{A;B} : \mathcal{A}\times\mathcal{B} \to [-\infty,\infty]$ is

$$\imath_{A;B}(a;b) := \log \frac{P_{AB}(a,b)}{P_A(a)\ P_B(b)}.$$

Let $\mathcal{P}_\mathcal{X}$ denote the set of all pmfs on the channel input alphabet $\mathcal{X}$. For each $k \in \{1,2,\ldots\}$ and $P_X \in \mathcal{P}_\mathcal{X}$, let

$$\tau_M(P_X,k) := \mathbb{E}\Big[\min\Big\{1, M^{k-1}(M-1)\ \zeta_k(X_{[1,k]}, Y_{[1,k]}, \tilde{X}_{[1,k]})\Big\}\Big], \quad (3)$$

where

$$(X_{[1,k]}, Y_{[1,k]}, \tilde{X}_{[1,k]}) = (X_1, Y_1, \tilde{X}_1), \ldots, (X_k, Y_k, \tilde{X}_k)$$

is a string of $k$ iid tuples

$$(X, Y, \tilde{X}) \sim P_X(x)\ T_{Y|X}(y|x)\ P_X(\tilde{x})$$

on $\mathcal{X} \times \mathcal{Y} \times \mathcal{X}$; the expectation in (3) is taken with respect to $(X_{[1,k]}, Y_{[1,k]})$; and

$$\zeta_k(X_{[1,k]}, Y_{[1,k]}, \tilde{X}_{[1,k]}) := \mathbb{P}\Big[\sum_{i=1}^{k} \imath_{X;Y}\big(\tilde{X}_i; Y_i\big) \geq \sum_{i=1}^{k} \imath_{X;Y}\big(X_i; Y_i\big)\Big|\big(X_{[1,k]}, Y_{[1,k]}\big)\Big].$$

*Theorem 2:*

$$\mathsf{D}^*_{\rho,M}(M, \Delta) \leq \mathsf{RCU}^*_{\rho,M}(\Delta, N),$$

where

$$\mathsf{RCU}^*_{\rho,M}(\Delta, N) := \inf_{P_X \in \mathcal{P}_\mathcal{X}} \frac{1}{N} \sum_{n=1}^{N} \left( \sum_{k=1}^{n} \left( \frac{\lambda^{n-k+1} - 1}{\lambda - 1} \right)^{\rho} \min\left\{1, \sum_{i=0}^{k-1} \tau_M(P_X, n + \Delta - i)\right\}\right).$$

*Proof:* A proof can be found in [4, Appendix D]. ∎

Theorem 2 generalises our previous RCU achievable bound in [13, Thm. 1], which is summarised below in Theorem 3, from $\rho = 1$ to arbitrary $\rho \geq 1$.

*Theorem 3:*

$$\mathsf{D}^*_{M}(\Delta, N) \leq \mathsf{RCU}^*_{M}(\Delta, N),$$

where

$$\mathsf{RCU}^*_{M}(\Delta, N) := \inf_{P_X \in \mathcal{P}_\mathcal{X}} \frac{1}{N} \sum_{n=1}^{N} \Big( \sum_{k=1}^{n} \lambda^{n-k} \min\Big\{1, \sum_{i=0}^{k-1} \tau_M(P_X, n + \Delta - i)\Big\}\Big).$$

Theorem 3 is a little stronger than Theorem 2 for $\rho = 1$, so we used this bound to prove Theorem 1 and the propositions in Section II.

## V. Boundedness of $\mathsf{ME}^*_\rho(\Delta)$ for General DMCs

The achievable bounds on $\mathsf{ME}^*_\rho(\Delta)|_{\rho=1}$ presented in Section II-A for the binary erasure and packet erasure channels followed by carefully bounding the function $\tau_M(\cdot)$ in Theorem 3. In this section, we consider arbitrary DMCs and give a sufficient condition for $\mathsf{ME}^*_\rho(\Delta)$ to be finite. To proceed, we first need the following definitions and notation.

Let us denote the capacity of the DMC $T_{Y|X}$ (in nats per channel use) by

$$\mathsf{C} := \max_{P_X \in \mathcal{P}_\mathcal{X}} I(X; Y).$$

For rates $0 \leq R < \mathsf{C}$, the *random-coding exponent* of the DMC $T_{Y|X}$ is [2, p. 139]

$$\mathsf{E}_\mathsf{r}(R) := \max_{\rho \in [0,1]} \max_{P_X \in \mathcal{P}_\mathcal{X}} \Big[\mathsf{E}_\mathsf{o}(\rho, P_X) - \rho R\Big],$$

where

$$\mathsf{E}_\mathsf{o}(\rho, P_X) := -\ln\left( \sum_{y \in \mathcal{Y}} \left( \sum_{x \in \mathcal{X}} P_X(x) \Big(T_{Y|X}(y|x)\Big)^{\frac{1}{1+\rho}} \right)^{1+\rho} \right).$$

*Theorem 4:*

$$\sup_{N \in \{1,2,\ldots\}} \mathsf{RCU}^*_M(\Delta, N) \leq \inf_R \left( \frac{(1 - e^{-R})}{e^{\mathsf{E}_\mathsf{r}(R)}\ (1 - e^{-\mathsf{E}_\mathsf{r}(R)})\ (1 - \lambda e^{-\mathsf{E}_\mathsf{r}(R)})} \right) e^{-\Delta \mathsf{E}_\mathsf{r}(R)},$$

where the infimum is taken over all $R > \ln 2$ such that $\ln \lambda < \mathsf{E}_\mathsf{r}(R)$. If no such $R$ exists, then we take the bound to be infinite.

*Proof:* A proof can be found in [4, Appendix I]. ∎

*Corollary 4.1:* Suppose that the DMC satisfies $\mathsf{E}_\mathsf{r}(R) > 0$ for all $0 \leq R < \mathsf{C}$ and $\mathsf{E}_\mathsf{r}(\mathsf{C}) = 0$. If

$$\ln \lambda < \mathsf{E}_\mathsf{r}(\ln 2) < \mathsf{C},$$

then $\mathsf{ME}^*_\rho(\Delta)|_{\rho=1}$ is finite for all $\Delta \in \{0, 1, \ldots\}$.

*Proof:* A proof can be found in [4, Appendix I]. ∎

## Acknowledgements

This work was supported by the Alexander von Humboldt foundation and the ARC Grant DE12010016.

## References

[1] R. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inform. Theory*, vol. 16, no. 4, pp. 412 – 421, 1970.
[2] R. Gallager, *Information Theory and Reliable Communication*. John Wiley and Sons, Inc. New York, NY, USA, 1968.
[3] J. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inform. Theory*, vol. 29, no. 6, pp. 814 – 820, 1983.
[4] R. Timo, B. N. Vellambi, A. Grant, and K. D. Nguyen, "Tracking unstable autoregressive sources over discrete memoryless channels," *preprint submitted to IEEE Trans. Inform. Theory,* August, 2015. http://roytimo.wordpress.com/pub/
[5] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4017 – 4035, 2006.
[6] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5317 – 5338, 2009.
[7] S. Yüksel and T. Başar, *Stochastic networked control systems*. Birkhauser, 2013.
[8] A. Sahai, "Any-time capacity and a separation theorem for tracking unstable processes," in *IEEE Intl. Symp. Inform. Theory*, Italy, 2000.
[9] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link — Part I: scalar systems," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3369–3395, 2006.
[10] A. S. Matveev and A. V. Savkin, "An analogue of Shannon information theory for detection and stabilization via noisy discrete communication channels," *SIAM J. Control Optim.*, vol. 46, no. 4, 2007.
[11] E. Arikan, "Channel combining and splitting for cutoff rate improvement," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 628–639, 2006.
[12] ——, "A perspective on polar coding," *preprint*, 2015.
[13] R. Timo, A. Grant, and B. N. Vellambi, "Streaming with autoregressive-hamming distortion for ultra short-delay communications," in *IEEE Intl. Symp. Inform. Theory*, Honolulu, USA, 2014.
[14] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," PhD Thesis, Princeton University, 2010.

# Strong Converse for General Compound Channels

Sergey Loyka, Charalambos D. Charalambous

***Abstract*—A general compound channel is considered, where no stationarity, ergodicity or information stability is required. Following the recent result on the capacity of this channel under the full Rx CSI, sufficient and necessary conditions are obtained for the strong converse to hold. In a nutshell, even though no information satiability is required upfront, the conditions imply that there exists a sub-sequence of (bad) channel states (indexed by the blocklength) for which the respective information density rates converge in probability to the compound channel capacity, i.e. this sub-sequence is information stable.**

## I. Introduction

It is well-known that channel state information (CSI) affects significantly system performance and respective channel capacity. It can be rather limited in many scenarios, especially for wireless systems, where low SNR, interference and channel dynamics are significant, and where the feedback (if any) is also limited [1]. A popular approach to model the impact of limited CSI is to assume that the receiver (Rx) and transmitter (Tx) know that the unknown channel is fixed and belongs to a certain class of channels (uncertainty set), which is known as the compound channel model [2]-[6]. The capacity of compound channels has been extensively studied since late 1950s [2]-[5]; see [6] for an extensive literature review up to late 1990s, and [9] for more recent results.

All of these studies assume that each channel in the uncertainty set is information-stable (in the sense of Dobrushin [10] or Pinsker [11]), e.g. stationary and ergodic. However, there are many scenarios (especially in wireless communications) where the channels are not stationary, ergodic or information-stable. This setting was recently studied in [14], where the capacity of general (information-unstable) compound channels was established under the full Rx CSI using the information density (spectrum) approach of [7][8]. The assumption of full Rx CSI is motivated by the fact that channel estimation is done at the Rx so that full Rx CSI may be available if the SNR is high enough but limited (if any) feedback to the Tx makes full Tx CSI unfeasible.

While the channel capacity theorem ensures the achievability of any rate below the capacity with arbitrary low error probability, there exists a hope to achieve higher rates by allowing slightly higher error probability, since the transition from arbitrary low to high error probability may be slow. Strong converse ensures that this transition is very sharp (for any rate above the capacity, the error probability converges to 1) and hence dispels the hope. In this paper, we extend the study in [14] by establishing the sufficient and necessary conditions for the strong converse to hold for the general compound channel. In a nutshell, the conditions require the existence of an information-stable sub-sequence of (bad) channel states (indexed by the blocklength) such that the respective sub-sequence of information densities converges in probability to the compound channel capacity. No assumptions of stationarity, ergodicity or information stability are made for the members of the uncertainty set.

S. Loyka is with the School of Electrical Engineering and Computer Science, University of Ottawa, Ontario, Canada, e-mail: sergey.loyka@ieee.org

C.D. Charalambous is with the ECE Department, University of Cyprus, Nicosia, Cyprus, e-mail: chadcha@ucy.ac.cy

## II. Channel Model

Let us consider a generic discrete-time channel model where $X^n = \{X_1 ... X_n\}$ is a (random) sequence of $n$ input symbols, $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ denotes all such sequences, and $Y^n$ is the corresponding output sequence; $s \in \mathcal{S}$ denotes the channel state (which may also be a sequence) and $\mathcal{S}$ is the (arbitrary) uncertainty set; $p_s(y^n|x^n)$ is the channel transition probability; $p(x^n)$ and $p_s(y^n)$ are the input and output distributions under channel state $s$.

Let us assume that the full CSI is available at the receiver (Rx) but not the transmitter (Tx) (see e.g. [1] for a detailed motivation of this assumption; when the channel is quasi-static, this assumption is not necessary) and that the channel input $\boldsymbol{X}$ and state $s$ are independent of each other. Following the standard approach (see e.g. [1]), we augment the channel output with the state: $Y^n \to (Y^n, s)$. The information density [10]-[13] between the input and output for a given channel state $s$ and a given input distribution $p(x^n)$ is

$$i(x^n; y^n, s) = \ln \frac{p_s(x^n, y^n)}{p(x^n) p_s(y^n)} = i(x^n; y^n|s) \qquad (1)$$

where we have used the fact that the input $X^n$ and channel state $s$ are independent of each other. Note that we make no assumptions of stationarity, ergodicity or information stability in this paper, so that the normalized information density $n^{-1} i(X^n; Y^n|s)$ does not have to converge to the respective mutual information rate as $n \to \infty$. There is no need for the consistency assumption on $p_s(y^n|x^n)$ either (e.g. the channel may behave differently for even and odd $n$).

For future use, we give the formal definitions of information stability following [10]-[12] (with a slight extension to the compound setting).

**Definition 1.** *Two random sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information-stable if*

$$\frac{i(X^n; Y^n|s)}{I(X^n; Y^n|s)} \xrightarrow{\Pr} 1 \text{ as } n \to \infty \qquad (2)$$

*i.e. the normalized information density converges in probability to the respective mutual information rate $\frac{1}{n} I(X^n; Y^n|s)$.*

**Definition 2.** *Channel state $s$ is information stable if there exists an input $\boldsymbol{X}$ such that*

$$\frac{i(X^n;Y^n|s)}{I(X^n;Y^n|s)} \xrightarrow{\Pr} 1, \quad \frac{I(X^n;Y^n|s)}{C_{ns}} \to 1 \text{ as } n \to \infty, \quad (3)$$

*where $C_{ns} = \sup_{p(x^n)} I(X^n;Y^n|s)$ is the information capacity.*

Note that the 2nd definition requires effectively the channel to behave ergodically under the optimal input only, and tells us nothing about its behaviour under other inputs (e.g. a practical code) and, in this sense, is rather limiting. To characterize the channel behaviour under different inputs (not only the optimal one), we will consider the information stability of its input $\boldsymbol{X}$ and the induced output $\boldsymbol{Y}$ following Definition 1. Further note that, for the compound channel, some channel states may be information stable while others are not.

## III. Capacity of the General Compound Channel

We define an ($n$, $r_n$, $\varepsilon_n$)-code for a compound channel in the standard way, where $n$ is the blocklength, $r_n = \ln M_n/n$ is the code rate and $M_n$ is the number of codewords, and $\varepsilon_n$ is the compound error probability,

$$\varepsilon_n = \sup_{s\in\mathcal{S}} \varepsilon_{ns} \quad (4)$$

where $\varepsilon_{ns}$ is the error probability under channel state $s$. Rate $R$ is achievable if $\liminf_{n\to\infty} r_n \geq R$ and $\lim_{n\to\infty} \varepsilon_n = 0$, which ensures arbitrary low error probability for any channel in the uncertainty set for sufficiently large $n$ [1]-[6]. The capacity is the supremum of all achievable rates. Codebooks are required to be independent of the actual channel state $s$ while the decision regions are allowed to depend on $s$ (due to full Rx CSI).

Below, we briefly review the relevant results in [14], which are instrumental for further development here.

**Theorem 1** ([14])**.** *Consider a general compound channel where the channel state $s \in \mathcal{S}$ is known to the receiver but not the transmitter and is independent of the channel input; the transmitter knows the (arbitrary) uncertainty set $\mathcal{S}$. Its compound channel capacity is given by*

$$C_c = \sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) \quad (5)$$

*where the supremum is over all sequences of finite-dimensional input distributions and $\underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y})$ is the compound inf-information rate,*

$$\underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) = \sup_R \left\{ R : \lim_{n\to\infty} \sup_{s\in\mathcal{S}} \Pr\{Z_{ns} \leq R\} = 0 \right\} \quad (6)$$

*where $Z_{ns} = n^{-1} i(X^n;Y^n|s)$ is the normalized information density under channel state $s$.* □

This theorem was proved using the Verdu-Han and Feinstein Lemmas properly extended to the compound channel setting.

**Lemma 1** (Feinstein Lemma for compound channels [14])**.** *For arbitrary input $X^n$ and uncertainty set $\mathcal{S}$ and any $r_n$, there exists a $(n, r_n, \varepsilon_n)$-code (where the codewords are independent of channel state $s$), satisfying the following inequality,*

$$\varepsilon_n \leq \sup_{s\in\mathcal{S}} \Pr\left\{ n^{-1} i(X^n;Y^n|s) \leq r_n + \gamma \right\} + e^{-\gamma n} \quad (7)$$

*for any $\gamma > 0$.* □

**Lemma 2** (Verdu-Han Lemma for compound channels [14])**.** *For any uncertainty set $\mathcal{S}$, every $(n, r_n, \varepsilon_n)$-code satisfies the following inequality,*

$$\varepsilon_n \geq \sup_{s\in\mathcal{S}} \Pr\left\{ n^{-1} i(X^n;Y^n|s) \leq r_n - \gamma \right\} - e^{-\gamma n} \quad (8)$$

*for any $\gamma > 0$, where $X^n$ is uniformly distributed over all codewords and $Y^n$ is the corresponding channel output under channel state $s$.* □

## IV. Strong Converse for the General Compound Channel

Strong converse ensures that slightly larger error probability cannot be traded off for higher data rate (since the transition from arbitrary low to high error probability is sharp).

**Definition 3.** *A compound channel is said to satisfy strong converse if*

$$\lim_{n\to\infty} \varepsilon_n = 1 \quad (9)$$

*for any code satisfying*

$$\liminf_{n\to\infty} r_n > C_c \quad (10)$$

To obtain conditions for strong converse, let $\check{I}(\boldsymbol{X};\boldsymbol{Y})$ be the "worst-case" sup-information rate,

$$\check{I}(\boldsymbol{X};\boldsymbol{Y}) = \inf_R \left\{ R : \lim_{n\to\infty} \inf_{s\in\mathcal{S}} \Pr\{Z_{ns} > R\} = 0 \right\} \quad (11)$$

where $Z_{ns} = n^{-1} i(X^n;Y^n|s)$ is the information density rate, and $I_{ns}(a)$ be the truncated mutual information,

$$I_{ns}(a) = E\{Z_{ns} 1[Z_{ns} \leq a]\}, \; I_{ns} = \lim_{a\to\infty} I_{ns}(a) \quad (12)$$

where $1[\cdot]$ is the indicator function and $I_{ns} = I(X^n;Y^n|s)$ is the mutual information under channel state $s$. The compound sup-information rate $\overline{\overline{I}}(\boldsymbol{X};\boldsymbol{Y})$ and the sup-information rate $\bar{I}(\boldsymbol{X};\boldsymbol{Y}|s)$ under channel state $s$ are defined as

$$\overline{\overline{I}}(\boldsymbol{X};\boldsymbol{Y}) = \inf_R \left\{ R : \lim_{n\to\infty} \sup_{s\in\mathcal{S}} \Pr\{Z_{ns} \geq R\} = 0 \right\} \quad (13)$$

$$\bar{I}(\boldsymbol{X};\boldsymbol{Y}|s) = \inf_R \left\{ R : \lim_{n\to\infty} \Pr\{Z_{ns} \geq R\} = 0 \right\} \quad (14)$$

The following Proposition establishes an ordering of various information rates.

**Proposition 1.** *The following inequalities hold for any input*

$$\begin{aligned} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) &\leq \check{I}(\boldsymbol{X};\boldsymbol{Y}) \\ &\leq \inf_s \bar{I}(\boldsymbol{X};\boldsymbol{Y}|s) \\ &\leq \sup_s \bar{I}(\boldsymbol{X};\boldsymbol{Y}|s) \\ &\leq \overline{\overline{I}}(\boldsymbol{X};\boldsymbol{Y}) \end{aligned} \quad (15)$$

*Proof.* see the Appendix. □

It can be shown, via examples, that all inequalities can be strict. Using this Proposition, sufficient and necessary conditions for the strong converse to hold can be established.

**Theorem 2.** *A sufficient and necessary condition for the general compound channel to satisfy strong converse is*

$$\sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) = \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) \tag{16}$$

*If this holds and the convergence $I_{ns}(a) \to I_{ns}$ is uniform in $n, s$ for any input $\boldsymbol{X}^*$ satisfying $\underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) > C_c - \delta$ for some $\delta > 0$ (i.e. the input $\boldsymbol{X}^*$ is $\delta$-suboptimal), then*

$$C_c = \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) = \liminf_{n\to\infty} \sup_{p(x^n)} \inf_s \frac{1}{n} I(X^n;Y^n|s) \tag{17}$$

*The condition* (16) *is equivalent to:*

*1) for any $\delta > 0$ and any input $\boldsymbol{X}^*$ satisfying $\underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) > C_c - \delta$,*

$$\lim_{n\to\infty} \inf_s \Pr\{|Z^*_{ns} - C_c| > \delta\} = 0 \tag{18}$$

*where $Z^*_{ns} = \frac{1}{n} i(X^{n*};Y^{n*}|s)$ is the normalized information density under input $\boldsymbol{X}^*$.*

*2) for any input $\boldsymbol{X}$ and any $\delta > 0$,*

$$\lim_{n\to\infty} \inf_s \Pr\{Z_{ns} > C_c + \delta\} = 0 \tag{19}$$

*Proof.* see the Appendix. □

**Remark 1.** *In the case of a single-state channel,*

$$\underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) = \underline{I}(\boldsymbol{X};\boldsymbol{Y}),\ \check{I}(\boldsymbol{X};\boldsymbol{Y}) = \overline{I}(\boldsymbol{X};\boldsymbol{Y}) \tag{20}$$

*where $\underline{I}(\boldsymbol{X};\boldsymbol{Y})$, $\overline{I}(\boldsymbol{X};\boldsymbol{Y})$ are inf and sup-information rates for the regular (single-state) channel, and Theorem 2 reduces to the corresponding Theorem in [7][8].*

**Remark 2.** *Note that, under the conditions of Theorem 2 that lead to* (17)*, the compound channel behaves ergodically even though no assumption of ergodicity (or information stability) was made upfront.*

Below, we consider a special case when the supremum in (5) is achieved.

**Corollary 1.** *If the channel satisfies strong converse and the supremum in $\sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y})$ is achieved, i.e.*

$$\exists \boldsymbol{X}^* :\ \underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) = C_c \tag{21}$$

*then $\check{I}(\boldsymbol{X}^*;\boldsymbol{Y}^*) = C_c$ and there exists such sequence of channel states $s(n)$ that the corresponding sequence of normalized information densities $Z^*_{ns(n)}$ (under input $\boldsymbol{X}^*$) converges in probability to the compound channel capacity $C_c$,*

$$\lim_{n\to\infty} \Pr\{|Z^*_{ns(n)} - C_c| > \delta\} = 0\ \forall \delta > 0 \tag{22}$$

*i.e. this sequence (which represents worst-case channels in the uncertainty set) is information-stable.*

*Proof.* Observe that $\underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) = C_c$ implies

$$C_c = \underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) \le \check{I}(\boldsymbol{X}^*;\boldsymbol{Y}^*) \le \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) = C_c \tag{23}$$

so that $\check{I}(\boldsymbol{X}^*;\boldsymbol{Y}^*) = C_c$ follows, which also implies that

$$\lim_{n\to\infty} \inf_s \Pr\{Z^*_{ns} > C_c + \delta\} = 0\ \forall\ \delta > 0 \tag{24}$$

On the other hand, $\underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) = C_c$ implies

$$\lim_{n\to\infty} \sup_s \Pr\{Z^*_{ns} < C_c - \delta\} = 0\ \forall\ \delta > 0 \tag{25}$$

and hence

$$\lim_{n\to\infty} \inf_s \Pr\{|Z^*_{ns} - C_c| > \delta\} = 0\ \forall \delta > 0 \tag{26}$$

follows. Next, we need the following technical Lemma.

**Lemma 3.** *Let $\{x_{ns}\}$ be a non-negative compound sequence such that*

$$\lim_{n\to\infty} \inf_s x_{ns} = 0 \tag{27}$$

*Then, there exists such sequence of states $s(n)$ that*

$$\lim_{n\to\infty} x_{ns(n)} = 0 \tag{28}$$

*Proof.* When $\inf_s$ is achieved, the statement is trivial. To prove it in the general case, observe that, from the definition of $\inf_s$ and for any $n$, there always exists such $s(n)$ that

$$x_{ns(n)} < \inf_s x_{ns} + 1/n \tag{29}$$

so that taking $\lim_{n\to\infty}$ of both sides, one obtains (28)[1]. □

Using this Lemma, (26) implies the existence of a sequence of channel states $s(n)$ such that (22) holds. □

**Remark 3.** *Note that, under the conditions of Corollary 1, the sequence $s(n)$ of worst-case channel states is information-stable even though no assumption of information stability was made upfront.*

**Remark 4.** *In light of Lemma 3, condition* (19) *means that there exists such sequence of (bad) channel states $s(n)$ that the information spectrum of the corresponding sequence of normalized information densities $Z_{ns(n)}$ does not exceed $C_c$ under any input, i.e.*

$$\exists s(n) :\ \lim_{n\to\infty} \Pr\{Z_{ns(n)} > C_c + \delta\} = 0\ \forall \delta > 0 \tag{30}$$

## V. Appendix

### A. Proof of Proposition 1

The 1st inequality is proved by contradiction. Let $\underline{I} = \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y})$, $\check{I} = \check{I}(\boldsymbol{X};\boldsymbol{Y})$, assume $\underline{I} - \check{I} = 2\delta > 0$ and set

$$R = (\underline{I} + \check{I})/2 = \underline{I} - \delta = \check{I} + \delta \tag{31}$$

so that

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} < \underline{I} - \delta\} \\ &= \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} < R\} \\ &= 1 - \lim_{n\to\infty} \inf_s \Pr\{Z_{ns} \ge R\} \\ &= 1 - \lim_{n\to\infty} \inf_s \Pr\{Z_{ns} \ge \check{I} + \delta\} = 1 \end{aligned} \tag{32}$$

[1] this way of proof was suggested by a reviewer.

i.e. a contradiction.

The 2nd inequality is also proved by contradiction. Let $\bar{I} = \inf_s \bar{I}(\boldsymbol{X};\boldsymbol{Y}|s)$, assume $\check{I} - \bar{I} = 2\delta > 0$ and set

$$R = (\bar{I} + \check{I})/2 = \bar{I} + \delta = \check{I} - \delta \tag{33}$$

so that, from the definition of $\check{I}$,

$$\begin{aligned} 0 < \epsilon &= \limsup_{n\to\infty} \inf_s \Pr\{Z_{ns} > \check{I} - \delta\} \\ &\le \inf_s \limsup_{n\to\infty} \Pr\{Z_{ns} > \check{I} - \delta\} \\ &= \inf_s \limsup_{n\to\infty} \Pr\{Z_{ns} > \bar{I} + \delta\} \\ &\le \limsup_{n\to\infty} \Pr\{Z_{ns^*} > \bar{I} + \delta\} \\ &\le \limsup_{n\to\infty} \Pr\{Z_{ns^*} > \bar{I}(\boldsymbol{X};\boldsymbol{Y}|s^*) + \delta/2\} = 0 \end{aligned} \tag{34}$$

i.e. a contradiction, where $s^*$ is such channel state that

$$\bar{I}(\boldsymbol{X};\boldsymbol{Y}|s^*) \le \inf_s \bar{I}(\boldsymbol{X};\boldsymbol{Y}|s) + \delta/2 \tag{35}$$

The last inequality can be proved in a similar way.

### *B. Proof of Theorem 2*

To prove sufficiency, let the equality in (16) to hold and select a code satisfying

$$\liminf_{n\to\infty} r_n = R = C_c + 3\delta \tag{36}$$

for some $\delta > 0$, so that

$$r_n \ge R - \delta = C_c + 2\delta = \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) + 2\delta \tag{37}$$

for sufficiently large $n$. Using Lemma 2 for this code, one obtains:

$$\begin{aligned} \lim_{n\to\infty} \varepsilon_n &\ge \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} \le r_n - \delta\} \\ &\ge \lim_{n\to\infty} \sup_s \Pr\left\{ Z_{ns} \le \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) + \delta \right\} \\ &\ge \lim_{n\to\infty} \sup_s \Pr\left\{ Z_{ns} \le \check{I}(\boldsymbol{X};\boldsymbol{Y}) + \delta \right\} \\ &= 1 - \lim_{n\to\infty} \inf_s \Pr\left\{ Z_{ns} > \check{I}(\boldsymbol{X};\boldsymbol{Y}) + \delta \right\} \\ &= 1 \end{aligned} \tag{38}$$

so that (9) holds, where the last equality is due to

$$\lim_{n\to\infty} \inf_s \Pr\left\{ Z_{ns} > \check{I}(\boldsymbol{X};\boldsymbol{Y}) + \delta \right\} = 0 \tag{39}$$

which follows from (11).

To prove the necessary part, assume that (9) holds and, using Lemma 1, select a code satisfying

$$\lim_{n\to\infty} r_n = R = C_c + \delta \tag{40}$$

for some $\delta > 0$. This implies that

$$r_n \le C_c + 2\delta \tag{41}$$

for any sufficiently large $n$. Applying Lemma 1, one obtains

$$\begin{aligned} 1 = \lim_{n\to\infty} \varepsilon_n &\le \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} \le r_n + \delta\} \\ &\le \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} \le C_c + 3\delta\} \\ &= 1 \end{aligned} \tag{42}$$

from which it follows that

$$\lim_{n\to\infty} \inf_s \Pr\{Z_{ns} > C_c + 3\delta\} = 0 \tag{43}$$

which implies (19) and $\check{I}(\boldsymbol{X};\boldsymbol{Y}) \le C_c$ (under any input) so that, from Proposition 1,

$$C_c = \sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) \le \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) \le C_c \tag{44}$$

from which (16) follows.

To establish the sufficiency of (19), observe that it implies the 2nd inequality in (44) from which (16) follows, which is sufficient.

To establish (18), observe that $C_c = \sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y})$ implies that there exists such input $\boldsymbol{X}^*$ that $\underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) > C_c - 2\delta$ so that, for any such $\boldsymbol{X}^*$,

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_s \Pr\left\{ \frac{1}{n} i(X^{n*};Y^{n*}|s) < \underline{\underline{I}}(\boldsymbol{X}^*;\boldsymbol{Y}^*) - \delta \right\} \\ &\ge \lim_{n\to\infty} \sup_s \Pr\left\{ \frac{1}{n} i(X^{n*};Y^{n*}|s) < C_c - 3\delta \right\} = 0 \end{aligned} \tag{45}$$

Combining this with (43) applied to input $\boldsymbol{X}^*$, one obtains

$$\begin{aligned} \lim_{n\to\infty} \inf_s \Pr\{|Z^*_{ns} - C_c| > 3\delta\} &\le \lim_{n\to\infty} \inf_s \Pr\{Z^*_{ns} > C_c + 3\delta \\ &+ \lim_{n\to\infty} \sup_s \Pr\{Z^*_{ns} < C_c - 3\delta\} = 0 \end{aligned} \tag{46}$$

from which (18) follows.

To establish last equality in (17), let $\check{I} = \check{I}(\boldsymbol{X};\boldsymbol{Y})$ and observe that

$$\begin{aligned} I_{ns}(a) = &\underbrace{E\{Z_{ns} 1[Z_{ns} \le \check{I} + \delta]\}}_{e_1} \\ &+ \underbrace{E\{Z_{ns} 1[\check{I} + \delta < Z_{ns} \le a]\}}_{e_2} \end{aligned} \tag{47}$$

for some $\delta > 0$, where $1[\cdot]$ is the indicator function. The two expectation terms can be upper bounder as

$$\begin{aligned} e_1 &\le (\check{I} + \delta) \Pr\{Z_{ns} \le \check{I} + \delta\} \\ e_2 &\le a \cdot \Pr\{Z_{ns} > \check{I} + \delta\} \end{aligned} \tag{48}$$

so that

$$\begin{aligned} \liminf_{n\to\infty} \inf_s \frac{1}{n} I(X^n;Y^n|s) &= \liminf_{n\to\infty} \inf_s \lim_{a\to\infty} I_{ns}(a) \\ &= \lim_{a\to\infty} \liminf_{n\to\infty} \inf_s I_{ns}(a) \\ &\le \lim_{a\to\infty} \liminf_{n\to\infty} \inf_s ((\check{I} + \delta) \Pr\{Z_{ns} \le \check{I} + \delta\} \\ &\quad + a \cdot \Pr\{Z_{ns} > \check{I} + \delta\}) \\ &\le \lim_{a\to\infty} ((\check{I} + \delta) \limsup_{n\to\infty} \sup_s \Pr\{Z_{ns} \le \check{I} + \delta\} \\ &\quad + a \cdot \liminf_{n\to\infty} \inf_s \Pr\{Z_{ns} > \check{I} + \delta\}) \\ &= \check{I} + \delta \end{aligned} \tag{49}$$

where the 2nd equality is due to uniform convergence and the last equality is due to

$$\liminf_{n\to\infty} \inf_s \Pr\{Z_{ns} > \check{I} + \delta\}) = 0 \tag{50}$$

$$\begin{aligned} &\limsup_{n\to\infty} \sup_s \Pr\{Z_{ns} \le \check{I} + \delta\} \\ &\quad = 1 - \liminf_{n\to\infty} \inf_s \Pr\{Z_{ns} > \check{I} + \delta\}) = 1 \end{aligned} \tag{51}$$

Since (49) holds for arbitrary small $\delta > 0$, it follows that

$$\liminf_{n\to\infty} \inf_s \frac{1}{n} I(X^n; Y^n|s) \le \check{I} \tag{52}$$

for any input. Taking $\sup_{p(\boldsymbol{x})}$ on both sides, one obtains:

$$\begin{aligned} C_c &= \sup_{p(\boldsymbol{x})} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) \\ &\le \liminf_{n\to\infty} \sup_{p(x^n)} \inf_s \frac{1}{n} I(X^n; Y^n|s) \\ &\le \sup_{p(\boldsymbol{x})} \check{I}(\boldsymbol{X};\boldsymbol{Y}) = C_c \end{aligned} \tag{53}$$

from which the desired result follows, where the 1st inequality is due to Proposition 2 below.

**Proposition 2.** *Consider the general compound channel. Its compound inf-information rate is bounded as follows:*

$$\underline{\underline{I}}(\boldsymbol{X},\boldsymbol{Y}) \le \liminf_{n\to\infty} \inf_s \frac{1}{n} I(X^n; Y^n|s) \le \check{I}(\boldsymbol{X};\boldsymbol{Y}) \tag{54}$$

*Proof.* Let $Z_{ns} = \frac{1}{n} i(X^n; Y^n|s)$ and observe that

$$\begin{aligned} &\frac{1}{n} I(X^n; Y^n|s) = E\left\{Z_{ns}\right\} \\ &\quad \ge E\{Z_{ns} 1[Z_{ns} \le 0]\} + E\{Z_{ns} 1[Z_{ns} \ge \underline{\underline{I}} - \delta]\} \end{aligned} \tag{55}$$

for any $0 < \delta < \underline{\underline{I}}$, where $1[\cdot]$ is the indicator function and $\underline{\underline{I}} = \underline{\underline{I}}(\boldsymbol{X},\boldsymbol{Y})$. The 1st term $t_1$ can be lower bounded as follows:

$$\begin{aligned} t_1 &= E\{Z_{ns} 1[Z_{ns} \le 0]\} \\ &= \frac{1}{n} \sum_{x^n, y^n : z_{ns} \le 0} p_s(y^n) p(x^n) w_{ns} \ln w_{ns} \\ &\ge -\frac{1}{ne} \sum_{x^n, y^n : z_{ns} \le 0} p_s(y^n) p_s(x^n) \\ &\ge -\frac{1}{ne} \end{aligned} \tag{56}$$

where $w_{ns} = p_s(y^n|x^n)/p_s(y^n)$ and the 1st inequality follows from $w \ln w \ge -1/e$. The 2nd term $t_2$ can be lower bounded as follows:

$$\begin{aligned} t_2 &= E\{Z_{ns} 1[Z_{ns} \ge \underline{\underline{I}} - \delta]\} \\ &= \sum_{x^n, y^n : z_{ns} \ge \underline{\underline{I}} - \delta} z_{ns} p_s(y^n|x^n) p(x^n) \\ &\ge (\underline{\underline{I}} - \delta) \Pr\{Z_{ns} \ge \underline{\underline{I}} - \delta\} \end{aligned} \tag{57}$$

Combining these two bounds, one obtains:

$$\begin{aligned} &\liminf_{n\to\infty} \inf_s \frac{1}{n} I(X^n; Y^n|s) \\ &\quad \ge (\underline{\underline{I}} - \delta) \liminf_{n\to\infty} \inf_s \Pr\{Z_{ns} \ge \underline{\underline{I}} - \delta\} \\ &\quad = \underline{\underline{I}} - \delta \end{aligned} \tag{58}$$

where the equality follows from

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_s \Pr\{Z_{ns} < \underline{\underline{I}} - \delta\} \\ &= 1 - \lim_{n\to\infty} \inf_s \Pr\{Z_{ns} \ge \underline{\underline{I}} - \delta\} \end{aligned} \tag{59}$$

Since the inequality in (58) holds for each $\delta > 0$, one obtains the 1st inequality in (54) by taking $\delta \to 0$; the 2nd one has been already established in (52). □

## REFERENCES


[1] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Trans. Inform. Theory*, vol. 44, No. 6, pp. 2619-2692, Oct. 1998.

[2] R.L. Dobrushin, "Optimal information Transmission through a channel with unknown parameters," *Radiotekhnika i Electronika*, vol. 4, pp. 1951-1956, 1959.

[3] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, pp. 1229-1241, December 1959.

[4] J. Wolfowitz, "Simultaneous channels," *Arch. Rat. Mech. Anal.*, vol. 4, pp. 371-386, 1960.

[5] W. L. Root, P. P. Varaya, "Capacity of Classes of Gaussian Channels", *SIAM J. Appl. Math.*, vol. 16, no. 6, pp. 1350-1393, Nov. 1968.

[6] A. Lapidoth and P. Narayan, "Reliable Communication Under Channel Uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, No. 6, Oct. 1998.

[7] S. Verdu, T.S. Han, "A General Formula for Channel Capacity", *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147-1157, July 1994.

[8] T. S. Han, Information-Spectrum Method in Information Theory, New York: Springer, 2003.

[9] M. Effros, A. Goldsmith, Y. Liang, "Generalizing Capacity: New Definitions and Capacity Theorems for Composite Channels," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3069-3087, July 2010.

[10] R. L. Dobrushin, "A general formulation of the fundamental theorem of Shannon in information theory", Uspekhi Mat. Nauk, v. 14, no. 6(90), Nov.-Dec. 1959, pp.3–104.

[11] M. S. Pinsker, Information and Information Stability of Random Variables and Processes. San Francisco: Holden-Day, 1964.

[12] H.K. Ting, "On the Information Stability of a Sequence of Channels", Theory of Probability and Its Applications, v.7, N. 3, pp. 258-269, 1962.

[13] R.L. Stratonovich, Information Theory, Moscow: Sovetskoe Radio, 1974.

[14] S. Loyka, C. D. Charalambous, A General Formula for Compound Channel Capacity, IEEE Int. Symp. on Information Theory (ISIT-15), Hong Kong, June 14-19, 2015.

# Refined Error Probability Approximations in Quasi–Static Erasure Channels

Josep Font–Segura
Universitat Pompeu Fabra
josep.font@ieee.org

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

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

***Abstract*—This paper considers the transmission of codewords over a quasi–static binary erasure channel, where the erasure probability changes independently at each transmitted codeword. An approximation to the random–coding union bound suggests that the error probability exceeds the outage probability by a quantity that is inversely proportional to the blocklength.**

## I. Introduction

A quasi–static channel is a good model for delay–constrained communication over slow–varying channels [1]. The outage capacity has been emphasized as the most important information–theoretic measure in quasi–static channels. However, little attention has been given to the error probability. In [2], the performance of the quasi–static fading channel is described by means of Gallager–type random–coding bounds. Malkamäki *et al.* [3] proposed a tighter bound, and showed that the average error probability is asymptotically given by the outage probability in the limit of infinite codeword blocklength [3, Th. 2]. However, for finite codeword blocklength, this tighter bound has to be evaluated numerically, as the optimization of the bound involves the fading coefficients.

This paper considers the random–coding union (RCU) bound [4] to the error probability in the simple quasi–static binary erasure channel (BEC). By writing the RCU bound as a tail probability, we propose two saddlepoint approximations [5] that build upon the techniques of [2], [3]. By inspecting the asymptotic behavior of the saddlepoint with the blocklength, we finally derive an expansion of the RCU bound in inverse powers of the blocklength that suggests that the error probability converges to the outage probability as $\frac{\delta(R)}{n}$, where $n$ is the codeword blocklength, $R$ is the rate of the code, and $\delta(R)$ is a rate–dependent constant.

## II. Preliminaries

Consider the transmission of codewords of blocklength $n$ symbols, where each codeword spans a single BEC with uniformly distributed erasure probability $\varepsilon$, that changes independently from codeword to codeword. Given the erasure probability $\varepsilon$, the transition probability during the transmission of a codeword can be factorized as

$$W_\varepsilon^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} W_\varepsilon(y_i|x_i), \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ are the channel input and channel output sequence, respectively, and $W_\varepsilon(y|x)$ denotes the transition probability of a single BEC of erasure probability $\varepsilon$ [6].

We study the transmission of equiprobable messages $m \in \{1, \ldots, M\}$, where each message is mapped onto a codeword $\boldsymbol{x}(m)$, and the collection of all codewords is a code of rate $R = \frac{1}{n}\log M$. For a fixed erasure probability $\varepsilon$, the average error probability of the code is denoted as $P_e(n, \varepsilon)$. Here, we are mostly interested in the error probability averaged over the erasure probability, i.e.,

$$P_e(n) = \mathbb{E}\left[P_e(n, \varepsilon)\right]. \tag{2}$$

Random–coding arguments show the existence of a code whose error probability is, at least, as good as that of the ensemble average. In this work, we consider such a code.

Two random–coding upper bounds to the error probability for the block–fading channel were reported by Malkamäki *et al.* in [3]. Particularized for the quasi–static BEC, the first bound is based on a conditional Gallager bound [7] given the erasure probability [3, Eq. (16)–(17)], i.e.,

$$P_e(n, \varepsilon) \leq \begin{cases} 1 & \hat{\rho}_\varepsilon < 0 \\ e^{-n(E_0(\hat{\rho}_\varepsilon, \varepsilon) - \hat{\rho}_\varepsilon R)} & 0 \leq \hat{\rho}_\varepsilon \leq 1 \\ e^{-n(E_0(1, \varepsilon) - R)} & \hat{\rho}_\varepsilon > 1. \end{cases} \tag{3}$$

Then, the average over the erasure probability is applied. In (3), $\hat{\rho}_\varepsilon$ is the argument that maximizes $E_0(\rho, \varepsilon) - \rho R$, closely related to (15) later derived in the paper. As $\hat{\rho}_\varepsilon$ is a function of the erasure probability, the expectation of (3) with respect to $\varepsilon$ has to be numerically evaluated for a finite blocklength. Asymptotically, the Gallager bound (3) shows that the error probability converges to the outage probability, denoted as $P_{\text{out}}(R)$ and given as

$$P_{\text{out}}(R) = \mathbb{P}\left[I(\varepsilon) < R\right], \tag{4}$$

where $I(\varepsilon)$ is the mutual information of a single BEC with erasure probability $\varepsilon$ maximized over the input distribution.

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 under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150, TEC2012-38800-C03-03, and FJCI-2014-22747.

For the quasi–static BEC with uniformly distributed error probability, we have that

$$I(\varepsilon) = (1-\varepsilon)\log 2, \tag{5}$$

$$P_{\text{out}}(R) = \frac{R}{\log 2}. \tag{6}$$

A simpler bound was also proposed in [3, Eq. (22)] by first averaging the erasure probability and then optimizing a parameter that does not depend on $\varepsilon$:

$$P_e(n) \leq \begin{cases} 1 & \hat{\rho} < 0 \\ e^{-n(E_0(\hat{\rho})-\hat{\rho}R)} & 0 \leq \hat{\rho} \leq 1 \\ e^{-n(E_0(1)-R)} & \hat{\rho} > 1. \end{cases} \tag{7}$$

Here, $\hat{\rho}$ is the argument that maximizes $E_0(\rho) - \rho R$, closely related to (30) later derived in the paper. As pointed out in [3], (7) is a weaker bound to the error probability, as $\hat{\rho}$ will be only optimal for some realizations of $\varepsilon$.

In summary, Gallager arguments lead to a tighter bound that needs to be numerically evaluated, and a simpler bound that is especially loose in quasi–static channels (see [3, Fig. 2]). In this work, we discuss whether the performance gap between (3) and (7) is a genuine issue of quasi–static channels by studying more refined expressions of the error probability based on the random–coding union bound. We further study the convergence of the error probability to the outage probability. For a different perspective, the dual problem, i.e., the convergence of the achievable rate to the outage capacity, see the recent work by Yang *et al.* [8].

## III. Saddlepoint Approximations

### A. Saddlepoint Approximation of RCU$(n,\varepsilon)$

For a fixed BEC realization of the erasure probability $\varepsilon$, the RCU bound to the average error probability [4] is given by

$$P_e(n,\varepsilon) \leq \mathbb{E}\left[\min\left\{1, M\mathbb{P}\left[W_\varepsilon^n(\boldsymbol{Y}|\overline{\boldsymbol{X}}) \geq W_\varepsilon^n(\boldsymbol{Y}|\boldsymbol{X})|\boldsymbol{X},\boldsymbol{Y}\right]\right\}\right], \tag{8}$$

where $\boldsymbol{X}$, $\boldsymbol{Y}$ are the random variables for channel input and channel output sequences, respectively, and $\overline{\boldsymbol{X}}$ is distributed as $\boldsymbol{X}$ but independent of $\boldsymbol{Y}$. As noted in [9], we can apply Markov's inequality and weaken the RCU bound as

$$P_e(n,\varepsilon) \leq \text{RCU}(n,\varepsilon) \tag{9}$$

where $\text{RCU}(n,\varepsilon)$ is the tail probability

$$\text{RCU}(n,\varepsilon) = \mathbb{P}\left[\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon) \leq 0\right]. \tag{10}$$

In (10), the random variable $\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)$ is

$$\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon) = \sum_{i=1}^{n} i_s(X_i,Y_i,\varepsilon) + \log U - nR, \tag{11}$$

where $U$ is a uniform $(0,1)$ random variable, and the symbol $s$–information density is defined as

$$i_s(X,Y,\varepsilon) = \log \frac{W_\varepsilon(Y|X)^s}{\mathbb{E}[W_\varepsilon(Y|\overline{X})^s|Y]}. \tag{12}$$

For the quasi–static BEC, we note that (12) is independent on $s$, and that the bounds (8) and (10) coincide [9].

As noted in [10], the tail probability (10) can be expressed in terms of the inverse Laplace transformation [11] as

$$\text{RCU}(n,\varepsilon) = \frac{1}{2\pi j}\int_{\nu-j\infty}^{\nu+j\infty} \frac{\mathbb{E}\left[e^{-t\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)}\right]}{t}\,\mathrm{d}t, \tag{13}$$

where we assume that $\nu$ is within the range of convergence, i.e., $\nu \in (0,1)$. The evaluation of the expectation term in (13), using (11) and (12), leads to

$$\mathbb{E}\left[e^{-t\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)}\right] = \frac{e^{\kappa_{n,\varepsilon}(t)}}{1-t}, \tag{14}$$

where $\kappa_{n,\varepsilon}(t)$ is given as

$$\kappa_{n,\varepsilon}(t) = ntR + n\log\left(\frac{1}{2^t}(1-\varepsilon)+\varepsilon\right). \tag{15}$$

We note that the former expression can be written in terms of the Gallager function $E_0(t,\varepsilon)$ that appears in (3) through $\kappa_{n,\varepsilon}(t) = -n(E_0(t,\varepsilon) - tR)$. The critical points of (13) are two poles at $t=0$ and $t=1$, and a saddlepoint at $t = t_{n,\varepsilon}$, the absolute minimum of $\kappa_{n,\varepsilon}(t)$ in the real axis, i.e.,

$$t_{n,\varepsilon} = \underset{-\infty<t<\infty}{\arg\min}\, \kappa_{n,\varepsilon}(t). \tag{16}$$

If $0 \leq t_{n,\varepsilon} \leq 1$, it is safe to set $\nu = t_{n,\varepsilon}$ in (13). Yet, whenever $t_{n,\varepsilon} < 0$ and $t_{n,\varepsilon} > 1$, the poles at $t=0$ and $t=1$ introduce additional terms due to the Cauchy's residue theorem [11].

Since no closed–form solutions to the complex–integration (13) are available in general, we propose a Taylor expansion of $\kappa_{n,\varepsilon}(t)$ around $t_{n,\varepsilon}$, i.e.,

$$\kappa_{n,\varepsilon}(t) \approx \kappa_{n,\varepsilon}(t_{n,\varepsilon}) + \kappa'_{n,\varepsilon}(t_{n,\varepsilon})(t-t_{n,\varepsilon}) + \frac{1}{2}\kappa''_{n,\varepsilon}(t_{n,\varepsilon})(t-t_{n,\varepsilon})^2, \tag{17}$$

where $\kappa'_{n,\varepsilon}(t)$ and $\kappa''_{n,\varepsilon}(t)$ denote, respectively, the first and second derivatives of $\kappa_{n,\varepsilon}(t)$ with respect to $t$. Finally, following the footsteps of [10], we obtain that the RCU bound for the quasi–static BEC can be approximated as

$$\text{RCU}(n,\varepsilon) \approx \gamma_{n,\varepsilon} + \sigma_{n,\varepsilon} e^{\kappa_{n,\varepsilon}(t_{n,\varepsilon})}. \tag{18}$$

Here, the additive term $\gamma_{n,\varepsilon}$ can be expressed as

$$\gamma_{n,\varepsilon} = \begin{cases} 1 & t_{n,\varepsilon} < 0 \\ 0 & 0 \leq t_{n,\varepsilon} \leq 1 \\ e^{\kappa_{n,\varepsilon}(1)} & t_{n,\varepsilon} > 1, \end{cases} \tag{19}$$

whereas the pre–exponential term $\sigma_{n,\varepsilon}$ is given by

$$\sigma_{n,\varepsilon} = \overline{Q}\left(t_{n,\varepsilon}\sqrt{\kappa''_{n,\varepsilon}(t_{n,\varepsilon})}\right) + \overline{Q}\left((1-t_{n,\varepsilon})\sqrt{\kappa''_{n,\varepsilon}(t_{n,\varepsilon})}\right), \tag{20}$$

where

$$\overline{Q}(x) = \text{sign}(x)\frac{1}{2}\text{erfc}\left(\frac{|x|}{\sqrt{2}}\right)e^{\frac{x^2}{2}}. \tag{21}$$

The proposed approximation of the RCU involves determining the saddlepoint of (15), given by

$$t_{n,\varepsilon} = \log_2\left(\frac{1-\varepsilon}{\varepsilon}\frac{\log 2 - R}{R}\right). \tag{22}$$

It is straightforward to show that, asymptotically, the saddlepoint approximation (18) satisfies

$$\lim_{n\to\infty} \mathrm{RCU}(n,\varepsilon) = \mathbb{1}\{I(\varepsilon) < R\}\,, \tag{23}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. As $n \to \infty$, the saddlepoint approximation (18) approaches a Bernoulli random variable with probability of success $P_{\mathrm{out}}(R)$. Since this random variable is bounded, we can apply the Lebesgue dominated convergence theorem [12] to prove that

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{RCU}(n,\varepsilon)\right] = P_{\mathrm{out}}(R). \tag{24}$$

Therefore, the average of the saddlepoint approximation to the RCU given the erasure probability, shows that the error probability converges to the outage probability, but gives no direct information about the rate of this convergence.

### *B. Saddlepoint Approximation of RCU$(n)$*

For symmetry with the work by Malkamäki *et al.* [3], we now study the RCU bound to the error probability averaged over the erasure probability, i.e.,

$$P_e(n) \le \mathbb{E}\left[\min\left\{1, M\mathbb{P}\left[W_\varepsilon^n(\boldsymbol{Y}|\overline{\boldsymbol{X}}) \ge W_\varepsilon^n(\boldsymbol{Y}|\boldsymbol{X})|\boldsymbol{X},\boldsymbol{Y},\varepsilon\right]\right\}\right]. \tag{25}$$

Similarly to (9), we can apply the Markov's inequality and further weaken (25) as

$$P_e(n) \le \mathrm{RCU}(n), \tag{26}$$

where now the erasure probability $\varepsilon$ is treated as a random variable in the evaluation of the tail probability of $\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)$, i.e.,

$$\mathrm{RCU}(n) = \mathbb{P}\left[\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon) \le 0\right]. \tag{27}$$

We can again express the tail probability (27) in terms of the inverse Laplace transformation as

$$\mathrm{RCU}(n) = \frac{1}{2\pi j}\int_{\nu-j\infty}^{\nu+j\infty} \frac{\mathbb{E}\left[e^{-t\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)}\right]}{t}\,\mathrm{d}t, \tag{28}$$

where $\nu$ is within the region of convergence, i.e., $\nu \in (0,1)$. Taking into account the erasure probability $\varepsilon$ in the following expectation

$$\mathbb{E}\left[e^{-t\Phi_n(\boldsymbol{X},\boldsymbol{Y},\varepsilon)}\right] = \frac{e^{\kappa_n(t)}}{1-t}, \tag{29}$$

now $\kappa_n(t)$ is defined as

$$\kappa_n(t) = ntR + \log\frac{2^t - 2^{-nt}}{(2^t-1)(n+1)}. \tag{30}$$

Again, (30) is related to the Gallager function $E_0(t)$ involved in (7) through $\kappa_n(t) = -n\left(E_0(t) - tR\right)$. The saddlepoint to $\mathrm{RCU}(n)$ is defined as the absolute minimum of $\kappa_n(t)$ over the real axis, i.e.,

$$t_n = \underset{-\infty<t<\infty}{\arg\min}\ \kappa_n(t). \tag{31}$$

Similarly to (17), we approximate (28) by expanding $\kappa_n(t)$ around $t_n$, and obtain that the averaged RCU bound for the quasi–static BEC can be approximated as

$$\mathrm{RCU}(n) \approx \gamma_n + \sigma_n e^{\kappa_n(t_n)} \tag{32}$$

where now

$$\gamma_n = \begin{cases} 1 & t_n < 0 \\ 0 & 0 \le t_n \le 1 \\ e^{\kappa_n(1)} & t_n > 1, \end{cases} \tag{33}$$

and

$$\sigma_n = \overline{Q}\left(t_n\sqrt{\kappa_n''(t_n)}\right) + \overline{Q}\left((1-t_n)\sqrt{\kappa_n''(t_n)}\right). \tag{34}$$

Even though closed-form expressions for the saddlepoint $t_n$ are not available in this case, we further investigate its relation to the outage probability by proposing a saddlepoint approximation to the outage probability.

### *C. Saddlepoint Approximation of $P_{\mathrm{out}}(R)$*

We note that the outage probability (4) can be seen as a tail probability of the random variable

$$\Phi_{\mathrm{out}}(R) = I(\varepsilon) - R. \tag{35}$$

Therefore, it is natural to express the outage probability as the inverse Laplace transformation

$$P_{\mathrm{out}}(R) = \frac{1}{2\pi j}\int_{\nu-j\infty}^{\nu+j\infty} \frac{\mathbb{E}\left[e^{-t\Phi_{\mathrm{out}}(R)}\right]}{t}\,\mathrm{d}t, \tag{36}$$

where now the expectation is only with respect to the erasure probability, and $\nu \in (0,\infty)$. This leads to

$$\mathbb{E}\left[e^{-t\Phi_{\mathrm{out}}(R)}\right] = e^{\kappa_{\mathrm{out}}(t)}, \tag{37}$$

where $\kappa_{\mathrm{out}}(t)$ is given as

$$\kappa_{\mathrm{out}}(t) = tR + \log\frac{1-2^{-t}}{t\log 2}. \tag{38}$$

Now, (36) has only one pole at $t = 0$ and a saddlepoint at

$$t_{\mathrm{out}}(R) = \underset{-\infty<t<\infty}{\arg\min}\ \kappa_{\mathrm{out}}(t). \tag{39}$$

Mimicking (17) with (38), we may hence approximate the outage probability as

$$P_{\mathrm{out}}(R) \approx \gamma_{\mathrm{out}}(R) + \sigma_{\mathrm{out}}(R)e^{\kappa_{\mathrm{out}}(t_{\mathrm{out}}(R))}. \tag{40}$$

In this case, we have that the additive term $\gamma_{\mathrm{out}}(R)$ and the pre–exponential term $\sigma_{\mathrm{out}}(R)$ are given, respectively, as

$$\gamma_{\mathrm{out}}(R) = \begin{cases} 1 & t_{\mathrm{out}}(R) < 0 \\ 0 & t_{\mathrm{out}}(R) \ge 0 \end{cases} \tag{41}$$

$$\sigma_{\mathrm{out}}(R) = \overline{Q}\left(t_{\mathrm{out}}(R)\sqrt{\kappa''(t_{\mathrm{out}}(R))}\right). \tag{42}$$

## IV. An Asymptotic Expansion of RCU$(n)$

One advantage of the complex–integration expression of the RCU (28) is that the average with respect to the erasure probability is naturally incorporated in the definition of $\kappa_n(t)$. By further inspecting the behavior of the saddlepoint to the RCU as the codeword blocklength $n \to \infty$, we numerically notice that the saddlepoint $t_n \to 0$. This motivates to study the behavior of the product $nt_n$, illustrated in Fig. 1 for three different rates. Remarkably, $nt_n$ converges to $t_{\text{out}}(R)$. This suggests that it is safe to make the change of variable $nt = \alpha$ and integrate with $\alpha$, i.e.,

$$\text{RCU}(n) = \frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \frac{e^{\kappa_n\left(\frac{\alpha}{n}\right)}}{\alpha\left(1-\frac{\alpha}{n}\right)}\, \mathrm{d}\alpha, \tag{43}$$

where now the region of convergence is $\nu \in (0, n)$. From (30), we note that $\kappa_n\left(\frac{\alpha}{n}\right)$ has the form

$$\kappa_n\Big(\frac{\alpha}{n}\Big) = \alpha R + \log \frac{2^{\frac{\alpha}{n}} - 2^{-\alpha}}{(2^{\frac{\alpha}{n}} - 1)(n+1)}. \tag{44}$$

For sufficiently large codeword blocklength $n$, we derive a Taylor expansion in inverse powers of the codeword blocklength $n$, i.e.,

$$\frac{e^{\kappa_n\left(\frac{\alpha}{n}\right)}}{\alpha\left(1-\frac{\alpha}{n}\right)} = \theta_0(\alpha) + \frac{\theta_1(\alpha)}{n} + O\left(\frac{1}{n^2}\right), \tag{45}$$

where $O\left(\frac{1}{n^2}\right)$ is a term that vanishes at least as fast as $\frac{1}{n^2}$, and the coefficients $\theta_0(\alpha)$ and $\theta_1(\alpha)$ are given by

$$\theta_0(\alpha) = \frac{e^{\alpha R}(1-2^{-\alpha})}{\alpha^2 \log 2}, \tag{46}$$

and

$$\theta_1(\alpha) = \frac{e^{\alpha R}(1-2^{-\alpha})}{\alpha \log 2} - \frac{e^{\alpha R}(1-2^{-\alpha})}{\alpha^2 \log 2} + \frac{e^{\alpha R}(1+2^{-\alpha})}{2\alpha}, \tag{47}$$

respectively. Comparing (38) and (46), we first observe that in fact $\theta_0(\alpha)$ is related to $\kappa_{\text{out}}(t)$ as

$$\theta_0(\alpha) = \frac{e^{\kappa_{\text{out}}(\alpha)}}{\alpha}. \tag{48}$$

Hence, we may identify $\theta_0(\alpha)$ with the evaluation of the outage probability

$$\frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \frac{e^{\alpha R}(1-2^{-\alpha})}{\alpha^2 \log 2}\, \mathrm{d}\alpha = P_{\text{out}}(R). \tag{49}$$

Regarding $\theta_1(\alpha)$, we identify that the first term of $\theta_1(\alpha)$ is actually $e^{\kappa_{\text{out}}(\alpha)}$, and therefore that the complex–integration of this term is the probability density function of $\Phi_{\text{out}}$ (35) evaluated at the origin (see [11]). Since $\varepsilon$ is uniformly distributed, $\Phi_{\text{out}}$ is then uniformly distributed in the interval $[-R, \log 2 - R]$, and we have that

$$\frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \frac{e^{\alpha R}(1-2^{-\alpha})}{\alpha \log 2}\, \mathrm{d}\alpha = \frac{1}{\log 2}. \tag{50}$$

Likewise, we identify the second term of $\theta_1(\alpha)$ as $\theta_0(\alpha)$ in (48), again leading to the outage probability as in (49). Finally,

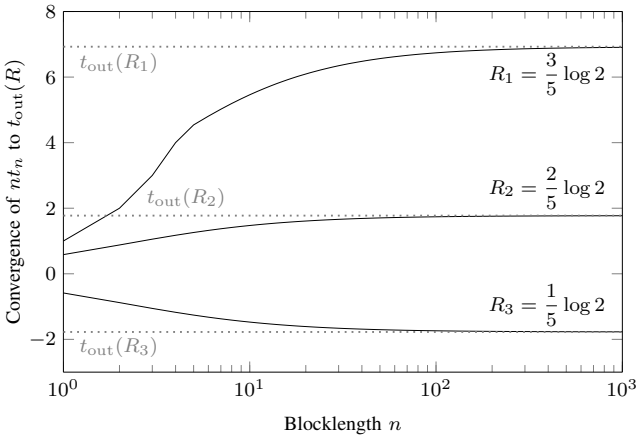



Fig. 1. Convergence of $nt_n$ to $t_{\text{out}}(R)$, versus the blocklength $n$, for several rates.

the last term in (47) can be be split into two additive terms that are identified as tail probabilities of two random variables. The first one is a random variable whose probability density function is a Dirac delta of mass one located at $-R$. Hence, we have that

$$\frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \frac{e^{\alpha R}}{2\alpha}\, \mathrm{d}\alpha = \frac{\mathbb{1}\{R > 0\}}{2}. \tag{51}$$

Similarly, the second one is a random variable whose probability density function is a Dirac delta of mass one located at and $-R + \log 2$ that evaluates as

$$\frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \frac{e^{\alpha R} 2^{-\alpha}}{2\alpha}\, \mathrm{d}\alpha = \frac{\mathbb{1}\{R > \log 2\}}{2}. \tag{52}$$

Defining $\delta(R)$ as

$$\delta(R) = \frac{1}{2\pi j} \int_{\nu-j\infty}^{\nu+j\infty} \theta_1(\alpha)\, \mathrm{d}\alpha, \tag{53}$$

within $0 < R < \log 2$ we have that

$$\delta(R) = \frac{1}{\log 2} - \frac{R}{\log 2} + \frac{1}{2}. \tag{54}$$

As a consequence, the expansion of the RCU is given by

$$\text{RCU}(n) = \frac{R}{\log 2} + \frac{1}{n}\left(\frac{1}{\log 2} - \frac{R}{\log 2} + \frac{1}{2}\right) + O\left(\frac{1}{n^2}\right). \tag{55}$$

The former expansion suggests that the error probability converges to the outage probability as $\frac{\delta(R)}{n}$, where $\delta(R)$ is a monotonically decreasing function of the rate.

## V. Numerical Results

In this section, we compare the proposed error probability approximations with the Gallager bounds (3) and (7), and the simulated RCU (27). More specifically, we numerically evaluate the saddlepoint approximations (18), (32), and (40), as well as the expansion (55).

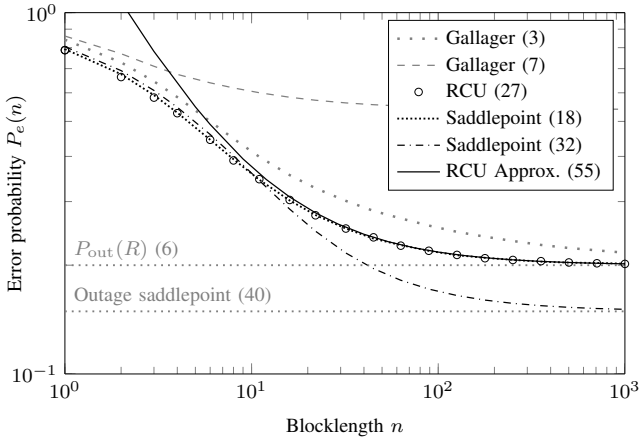



Fig. 2. Error probability bounds and approximations versus blocklength at $R = \frac{1}{5}\log 2$.

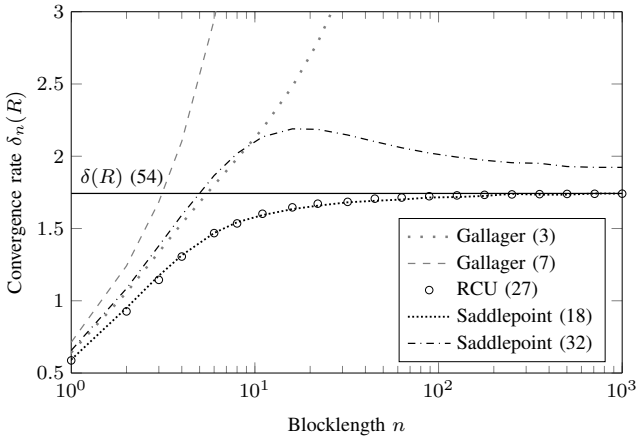



Fig. 3. Rate of convergence of the error probability to the outage probability versus codeword blocklength at $R = \frac{1}{5}\log 2$.

In Fig. 2, we observe that the saddlepoint approximation (18) is an accurate approximation of the RCU. As $n \to \infty$, the error probability converges to the outage probability (4), numerically confirming (24). Comparing the Gallager bound (3) with the saddlepoint approximation (18), and the Gallager bound (7) with the saddlepoint approximation (32), we note that in both cases the additive and the pre–exponential terms of the saddlepoint approximation provide a more refined characterization of the error probability. The contribution of these terms cannot be neglected in the quasi–static channel, since the exponential term of the error probability is not a dominant term when the error probability saturates.

A second observation from Fig. 2 is that, compared to the Gallager bound (7), the saddlepoint approximation (32) is tighter for small codeword blocklength. However, since the randomness of the erasure probability is considered in the approximation of the tail probability (27), this approximation exhibits a misadjustment for large blocklength, as it converges to the saddlepoint approximation of the outage probability (40), rather than to the actual outage probability (6).

Finally, we are interested in the convergence of the error probability to the outage probability. In particular, Fig. 3 depicts the convergence rate $\delta_n(R)$, defined as

$$\delta_n(R) = n\left(P_e(n) - P_{\text{out}}(R)\right), \tag{56}$$

where $P_e(n)$ is a placeholder for the bounds and approximations of Fig. 3. Remarkably, Fig. 2 numerically illustrates that the Taylor expansion of the RCU (55) is a good approximation even for small codeword blocklength. Moreover, Fig. 3 illustrates that the error probability indeed exceeds the outage probability in a quantity that vanishes proportionally to $\frac{1}{n}$. That is,

$$\lim_{n\to\infty} \delta_n(R) = \delta(R). \tag{57}$$

As expected, none of the Gallager bounds provide the convergence in $\frac{1}{n}$, as the bounds are only tight for sufficiently large $n$. Contrarily, the saddlepoint approximation (32) does exhibit, although misadjusted, a convergence coefficient as $\frac{1}{n}$, whereas the saddlepoint approximation (18) leads to the correct convergence rate of the RCU (27).

## VI. Conclusions

In this paper, we have derived refined approximations of the random–coding union bound in quasi–static binary erasure channels with uniformly distributed erasure probability. An expansion of the random–coding union bound in inverse powers of the codeword blocklength suggests that the error probability exceeds the outage probability by a quantity that is inversely proportional to the codeword blocklength.

## References

[1] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[2] G. Kaplan and S. Shamai (Shitz), "Error probabilities for the block–fading Gaussian channel," *Archiv für Elektronik und Übertragungstechnik (AEÜ)*, vol. 49, no. 4, pp. 192–205, 1995.

[3] E. Malkamäki and H. Leib, "Coded diversity on block–fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 771–781, Mar. 1999.

[4] Y. Polyanskiy, H. 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.

[5] J. L. Jensen, *Saddlepoint Approximations*. Oxford University Press, 1995.

[6] P. Elias, "Coding for two noisy channels," in *Proc. Inf. Theory: Third London Symp.*, 1955, pp. 61–74.

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

[8] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi–static multiple–antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[9] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random–coding bounds." in *Proc. Inf. Theory and Applications (ITA) Workshop*, 2011, pp. 257–262.

[10] A. Martinez, J. Scarlett, M. Dalai, and A. Guillén i Fàbregas, "A complex–integration approach to the saddlepoint approximation for random–coding bounds," in *Proc. Int. Symp. Wireless Commun. Systems (ISWCS)*, 2014, pp. 618–621.

[11] G. Doetsch, *Introduction to the Theory and Applications of the Laplace Transformation*. Springer, 1974.

[12] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2010.

# A Deterministic Construction and Density Evolution Analysis for Generalized Product Codes

Christian Häger[†], Henry D. Pfister[‡], Alexandre Graell i Amat[†], Fredrik Brännström[†], and Erik Agrell[†]
[†]Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden
[‡]Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina
{christian.haeger, alexandre.graell, fredrik.brannstrom, agrell}@chalmers.se, henry.pfister@duke.edu

***Abstract*—Generalized product codes (GPCs) are extensions of product codes (PCs) where code symbols are protected by two component codes but not necessarily arranged in a rectangular array. In this tutorial paper, we review a deterministic construction for GPCs that has been previously proposed by the authors together with an accompanying density evolution (DE) analysis. The DE analysis characterizes the asymptotic performance of the resulting GPCs under iterative bounded-distance decoding of the component codes over the binary erasure channel. As an application, we discuss the analysis and design of three different classes of GPCs: spatially-coupled PCs, symmetric GPCs, and GPCs based on component code mixtures.**

## I. Introduction

Several authors have proposed modifications of the classical product code (PC) construction by Elias [1], typically by considering non-rectangular code arrays. These modifications can be regarded as generalized low-density parity-check (GLDPC) codes [2]. In particular, they are GLDPC codes where the underlying Tanner graph consists exclusively of degree-2 variable nodes (VNs) (i.e., each bit is protected by two component codes). We refer to such codes as generalized PCs (GPCs).

In practice, the component codes of a GPC are typically Bose–Chaudhuri–Hocquenghem or Reed–Solomon codes, which can be efficiently decoded via algebraic bounded-distance decoding (BDD). The overall GPC can then be suboptimally decoded using iterative hard-decision decoding, i.e., by iteratively performing BDD of all component codes. This makes GPCs particularly suited for high-speed applications due to their significantly reduced decoding complexity compared to "soft" message-passing decoding of low-density parity-check (LDPC) codes [3]. For example, GPCs have been investigated by many authors as practical solutions for high-speed fiber-optical communications [3]–[7].

A standard tool to analyze the performance of iteratively decoded codes is density evolution (DE) [8], [9], which is based on an ensemble argument. That is, rather than analyzing a particular code directly, one considers a set of codes defined via suitable randomized edge connections in the Tanner graph. While this approach can be applied to GPCs, many classes of GPCs have a very regular Tanner graph structure. Therefore, the performance of such codes is not necessarily well predicted by using an ensemble analysis. In general, it would be desirable to make precise statements about the performance of sequences of deterministic codes without resorting to an ensemble argument.

This work was partially funded by the Swedish Research Council under grant #2011-5961.

In this tutorial paper, we discuss some recent results about the performance of deterministically constructed GPCs over the binary erasure channel (BEC) presented in [10]–[12]. We start in Section II by reviewing the deterministic construction for GPCs proposed in [10]. The resulting GPCs are defined by Tanner graphs that consist of a fixed arrangement of (degree-2) VNs and constraint nodes (CNs). In Section III, it is shown that the asymptotic performance of these GPCs is rigorously characterized by a recursive DE equation. Finally, in Section IV, we present a high-level overview of different results presented in [10]–[12]. In particular, we discuss the analysis and design of spatially-coupled PCs, symmetric GPCs, and GPCs based on component code mixtures.

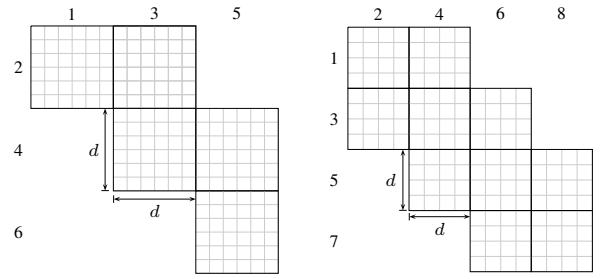
*Notation.* We use boldface to denote column vectors and matrices (e.g., $\boldsymbol{x}$ and $\boldsymbol{A}$). The symbols $\mathbf{0}_m$ and $\mathbf{1}_m$ denote the all-zero and all-one vectors of length $m$, respectively, where the subscript may be omitted. The tail-probability of a Poisson random variable is defined as $\Psi_{\geq t}(x) \triangleq 1 - \sum_{i=0}^{t-1} \frac{x^i}{i!} e^{-x}$. We use boldface to denote the element-wise application of a scalar-valued function to a vector. For example, if $\boldsymbol{x}$ is a vector, then $\boldsymbol{\Psi}_{\geq t}(\boldsymbol{x})$ applies the function to each element. For vectors $\boldsymbol{x} = (x_1, \ldots, x_m)^\intercal$ and $\boldsymbol{y} = (y_1, \ldots, y_m)^\intercal$, we use $\boldsymbol{x} \succeq \boldsymbol{y}$ if $x_i \geq y_i$ for all $i$. We also define $[m] \triangleq \{1, 2, \ldots, m\}$. Lastly, the indicator function is denoted by $\mathbb{1}\{\cdot\}$.

## II. A Deterministic Construction for Generalized Product Codes

### A. Motivation

Recall that a PC is defined as the set of $n \times n$ arrays such that every row and every column is a codeword in some binary linear component code $\mathcal{B}$ of length $n$. The corresponding Tanner graph has a fixed deterministic structure that resembles a complete bipartite graph: There exists two types of CNs ($n$ CNs corresponding to "row codes" and $n$ CNs corresponding to "column codes") and each CN of one type is connected to all CNs of the other type through a VN. This gives rise to exactly $n^2$ VNs, where each VN corresponds to one element in the array. An illustration is shown for example in [2, Fig. 3].

Consider now the code arrays shown in Fig. 1. We will discuss these arrays (and the resulting GPCs) in more detail in the next section. For now, we note that one can almost apply




(a) staircase code (b) block-wise braided code

Fig. 1. Code arrays for $\mathcal{C}_{12}(\boldsymbol{\eta})$, where in (a) $\gamma = 1/2$ and in (b) $\gamma = 1/3$. Numbers indicate the position indices in the code construction.

the exact definition of a PC to these arrays. In particular, fill the array with bits such that every row and every column is a codeword in some component code. The underlying Tanner graph that results from this definition is again easily seen to be very structured. We essentially seek a general and flexible way to directly construct the Tanner graphs corresponding to the GPCs defined by these arrays.

### B. Code Construction

We denote a GPC by $\mathcal{C}_n(\boldsymbol{\eta})$, where $n$ is proportional to the number of CNs in the underlying Tanner graph and $\boldsymbol{\eta}$ is a binary, symmetric $L \times L$ matrix that defines the graph connectivity. Due to the natural representation of GPCs in terms of two-dimensional code arrays, one may alternatively think about $\boldsymbol{\eta}$ as specifying the array shape. We will see in the following that different choices for $\boldsymbol{\eta}$ recover well-known code classes.

Let $\gamma > 0$ be some fixed and arbitrary constant such that $d \triangleq \gamma n$ is an integer. To construct the Tanner graph that defines $\mathcal{C}_n(\boldsymbol{\eta})$, assume that there are $L$ classes of CNs, here called "positions". Then, place $d$ CNs at each position and connect each CN at position $i$ to each CN at position $j$ through a VN if and only if $\eta_{i,j} = 1$.

*Example* 1. A PC is obtained for $L = 2$ and $\boldsymbol{\eta} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The two positions correspond to "row codes" and "column codes". If we choose $\gamma = 1$, then the code array is of size $n \times n$. $\triangle$

*Example* 2. For $L \geq 2$, the matrix $\boldsymbol{\eta}$ describing a staircase code [3] has entries $\eta_{i,i+1} = \eta_{i+1,i} = 1$ for $i \in [L-1]$ and zeros elsewhere. For example, for $L = 6$, we have

$$\boldsymbol{\eta} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{1}$$

The corresponding code array is exactly the one shown in Fig. 1(a), where $n = 12$ and $\gamma = 1/2$. $\triangle$

*Example* 3. For even $L \geq 4$, the matrix $\boldsymbol{\eta}$ for a particular instance of a block-wise braided code [13] has entries $\eta_{i,i+1} = \eta_{i+1,i} = 1$ for $i \in [L-1]$, $\eta_{2i-1,2i+2} = \eta_{2i+2,2i-1} = 1$ for $i \in [L/2-1]$, and zeros elsewhere. For example, we have

$$\boldsymbol{\eta} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \tag{2}$$

for $L = 8$. The corresponding code array is the one shown in Fig. 1(b), where $n = 12$ and $\gamma = 1/3$. $\triangle$

For a fixed $n$, the constant $\gamma$ scales the number of CNs in the graph. This is inconsequential for the asymptotic analysis (where we assume that $n \to \infty$) and $\gamma$ manifests itself in the DE equations merely as a scaling parameter. The scaling parameter $\gamma$ in the previous two examples is chosen such that the component codes have length $n$ in both cases, except at the array boundaries, see Fig. 1.

From the code construction, it follows that the total number of VNs (i.e., the length of the code $\mathcal{C}_n(\boldsymbol{\eta})$) is given by

$$m = \binom{d}{2} \sum_{i=1}^{L} \eta_{i,i} + d^2 \sum_{1 \leq i < j \leq L} \eta_{i,j}. \tag{3}$$

By construction, all of these VNs have degree two. Moreover, the total number of CNs is $dL$. In general, CNs at position $i$ have degree $d \sum_{j \neq i} \eta_{i,j} + \eta_{i,i}(d-1)$, where the second term arises from the fact that we cannot connect a CN to itself if $\eta_{i,i} = 1$. The CN degree specifies the length of the component code associated with the CN. We assume in the following that each CN corresponds to a $t$-erasure correcting component code. This assumption is relaxed in Section IV-C.

## III. Density Evolution Analysis

### A. Iterative Decoding

Suppose that a codeword of $\mathcal{C}_n(\boldsymbol{\eta})$ is transmitted over the BEC with erasure probability $p$. The decoding is performed iteratively assuming $\ell$ iterations of BDD for the component codes associated with all CNs. This means that in each iteration, if the weight of an erasure pattern associated with a CN is less than or equal to $t$, the pattern is corrected. If the weight exceeds $t$, we say that the component code declares a decoding failure in that iteration.

The decoding can be represented by applying the following peeling procedure to the so-called residual graph [4], [14]. The residual graph is obtained by deleting known VNs and their adjacent edges. Furthermore, erased VNs are collapsed into edges between CNs. In each iteration, determine all vertices that have degree at most $t$ and remove them, together with all adjacent edges. The decoding is successful if the resulting graph is empty after (at most) $\ell$ iterations.

### B. Density Evolution

We wish to characterize the decoding performance in the limit as $n \to \infty$. The first important observation is that with the assumptions given so far (in particular the finite erasure-correcting capability of the component codes), this problem is ill-posed for a fixed erasure probability $p$. The reason is that for $n \to \infty$, with high probability there will be a large number of erasures associated with each component code. Even if we choose $p$ very small, eventually, the number of erasures will exceed the (assumed) finite erasure-correcting capability of each component code. In other words, for any fixed $p$ and $n \to \infty$, the decoding will fail with high probability.

In order to allow for a meaningful analysis, the natural choice is to let the erasure probability decay slowly as $p = c/n$ for some $c > 0$. Since now $p \to 0$ as $n \to \infty$, one may (falsely) conclude that the decoding will always be successful in the limit. As we will see, however, the answer depends crucially on the choice of $c$, which may thus be interpreted as the effective channel quality in this regime. Its operational meaning (assuming an appropriate choice for $\gamma$, see [10, Sec. VI-A]) is given in terms of the expected number of initial erasures per component code.

Now, assume that we compute

$$\boldsymbol{z}^{(\ell)} = \boldsymbol{\Psi}_{\geq t+1}(c\boldsymbol{B}\boldsymbol{x}^{(\ell-1)}), \text{with } \boldsymbol{x}^{(\ell)} = \boldsymbol{\Psi}_{\geq t}(c\boldsymbol{B}\boldsymbol{x}^{(\ell-1)}), \quad (4)$$

where $\boldsymbol{x}^{(0)} = \mathbf{1}_L$ and $\boldsymbol{B} \triangleq \gamma\boldsymbol{\eta}$. The main technical result is that the fraction of component codes that declare decoding failures in iteration $\ell$ converges almost surely to $\frac{1}{L}\sum_{i=1}^{L} z_i^{(\ell)}$ as $n \to \infty$. In other words, the code performance concentrates around a deterministic value computed by the recursion (4) for sufficiently large $n$. This result is analogous to the DE analysis for LDPC codes [9, Th. 2]. The proof exploits the above peeling representation of the decoding and is based on a convergence result for so-called inhomogeneous random graphs in [15], see [10] for details.

For notational convenience, we define $h(x) \triangleq \Psi_{\geq t}(cx)$, so that the recursion in (4) can be succinctly written as

$$\boldsymbol{x}^{(\ell)} = \boldsymbol{h}(\boldsymbol{B}\boldsymbol{x}^{(\ell-1)}). \quad (5)$$

Furthermore, the decoding threshold is defined in terms of the effective channel quality as

$$\bar{c} \triangleq \sup\{c \geq 0 \,|\, \boldsymbol{x}^{(\infty)} = \mathbf{0}_L\}. \quad (6)$$

*Remark* 1. For component codes with fixed erasure-correcting capabilities, one can show that the code rate of $\mathcal{C}_n(\boldsymbol{\eta})$ approaches 1 as $n \to \infty$. The studied setup is sometimes also referred to as the high-rate regime or high-rate scaling limit [16]. It turns out that the regime that can be analyzed is also the regime that is relevant in practice: It is at high rates where GPCs are competitive in terms of performance and complexity compared to other code families, e.g., LDPC codes [3]–[5].

## IV. Applications

In this section, we discuss the analysis and design of three different classes of GPCs: spatially-coupled PCs, symmetric GPCs, and GPCs based on component code mixtures. This section is based on results presented in [10]–[12], [17].

### A. Spatially-Coupled Product Codes

Of particular interest are cases where the matrix $\boldsymbol{\eta}$ has a band-diagonal "convolutional-like" structure. The associated GPC can then be classified as a spatially-coupled PC. For example, the GPCs discussed in Examples 2 and 3, i.e., staircase and braided codes, are particular instances of spatially-coupled PCs. The matrix $\boldsymbol{B}$ is referred to as an averaging matrix in this case. Spatially-coupled codes have attracted a lot of attention in the literature due to their outstanding performance under iterative decoding [18], [19].

Spatially-coupled PCs have been previously analyzed using ensemble-based methods in [6], [16]. In [12], we compare the obtained DE recursion in (5) for deterministic spatially-coupled PCs to the DE recursion for the spatially-coupled PC *ensemble* in [16]. Without going into the details, the ensemble performance is described by the recursion (see [16, eq. (9)] and [12, Sec. III])

$$\boldsymbol{x}^{(\ell)} = \boldsymbol{h}(\tilde{\boldsymbol{B}}\boldsymbol{x}^{(\ell-1)}), \quad (7)$$

where $\boldsymbol{x}^{(0)} = \mathbf{1}_L$, $\tilde{\boldsymbol{B}} \triangleq \boldsymbol{A}^{\intercal}\boldsymbol{A}$, and $\mathbf{A}$ is an $L - w + 1 \times L$ matrix with entries $A_{i,j} = w^{-1}\mathbb{1}\{1 \leq j - i + 1 \leq w\}$ for $i \in [L - w + 1]$ and $j \in [L]$. The parameter $w$ is referred to as the coupling width. For example, for $L = 6$, the matrix $\tilde{\boldsymbol{B}}$ for $w = 2$ and $w = 3$ is given by

$$\frac{1}{4}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 2 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 2 & 1 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}, \quad (8)$$

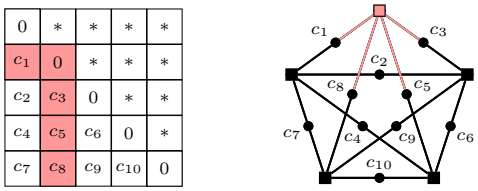
respectively. The ensemble DE recursion (7) has evidently the same form as (5). The difference lies in the averaging due to the matrix $\tilde{\boldsymbol{B}}$. This is illustrated in the following example.

*Example* 4. For the braided codes in Example 3, one can simplify (5) by exploiting the inherent symmetry in the code construction which implies $x_i^{(\ell)} = x_{i+1}^{(\ell)}$ for odd $i$ and any $\ell$. It is then sufficient to retain odd (or even) positions in (5). With this simplification, the effective averaging matrix is given by

$$\boldsymbol{B}' = \frac{1}{3}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad (9)$$

for $L = 12$. The matrix $\boldsymbol{B}'$ may be used to replace $\boldsymbol{B}$ in (5). Moreover, $\boldsymbol{B}'$ differs from both matrices $\tilde{\boldsymbol{B}}$ in (8). In general, the effective averaging matrices for the randomized and deterministic constructions are not the same. △

It is shown in [12] that there exists a different but related family of (deterministic) braided codes that has the same effective averaging matrix as the spatially-coupled PC ensemble,




(a) $5 \times 5$ array (b) Tanner graph

Fig. 2. Illustrations for an HPC with $n = 5$. In the array, "*" means "equal to the transposed element". The highlighted array elements illustrate one particular code constraint, which is also highlighted in the Tanner graph.

i.e., we have $\boldsymbol{B}' = \tilde{\boldsymbol{B}}$. This implies that the resulting DE recursions are identical and certain ensemble-properties proved in [16] (in particular lower bounds on the decoding threshold) also apply to certain deterministically constructed spatially-coupled PCs.

### B. Symmetric Generalized Product Codes

All examples for $\mathcal{C}_n(\boldsymbol{\eta})$ discussed so far share the property that the corresponding matrix $\boldsymbol{\eta}$ does not contain any ones on the diagonal, i.e., $\eta_{i,i} = 0$ for all $i \in [L]$. In this section, we discuss the implications of choosing $\eta_{i,i} = 1$. In other words, we discuss the implications of connecting CNs to other CNs *at the same position* in the deterministic GPC construction.

The simplest case is obtained when there is only one position (i.e., $L = 1$) and we have $\boldsymbol{\eta} = 1$ with $\gamma = 1$. The resulting Tanner graph can be described as a "complete Tanner graph": There exist $n$ CNs in total and each CNs is connected to all other CNs through a VN. All CNs have degree $n - 1$ and the total number of VNs, i.e., the length of the resulting code $\mathcal{C}_n(\boldsymbol{\eta})$, is given by $m = \binom{n}{2}$. Tanner already used such a construction as one of the first examples in [2, Fig. 6].

While the graph structure appears to be appealing due to its simplicity, it is not immediately clear if $\mathcal{C}_n(\boldsymbol{\eta})$ has a corresponding interpretation in terms of a code array. Such an interpretation was later provided by Justesen in [4, Sec. III-B]. In particular, assume that we start with a conventional (square) PC based on a component code with length $n$. Then, form a subcode of this PC by retaining only symmetric codeword arrays (i.e., arrays that are equal to their transpose) with a zero diagonal. After puncturing the diagonal and the upper (or lower) triangular part of the array, one obtains a code of length $m = \binom{n}{2}$. Justesen termed the resulting codes half-product codes (HPCs), emphasizing the fact that they have roughly half the length of the PCs from which they are derived.

*Example* 5. Figs. 2(a) and (b) show the code array and Tanner graph of an HPC for $n = 5$ and $m = 10$. The highlighted array elements show the code symbols participating in the second row constraint, which, due to the enforced symmetry, is also the second column constraint. Effectively, each component code acts on an L-shape in the array, i.e., both a partial row and column, which includes one diagonal element. The degree of each CN is $n-1 = 4$, due to the zeros on the diagonal. △

The definition of an HPC as a (punctured) symmetric subcode of a conventional PC extends without much difficulty to other GPCs. This leads to the class of symmetric GPCs which can be seen as a subclass of GPCs [17]. In general, symmetric GPCs use symmetry to reduce the block length of a GPC while employing the same component code [17].

*Example* 6. Consider again the code array in Fig. 1(b) corresponding to the braided code in Example 3. Similar to an HPC, we can form a half-braided code by enforcing the additional constraint that the array should be equal to its transpose and the array diagonal is zero (see [11, Fig. 1] for an illustration). After puncturing, we find that this GPC is defined by a matrix $\boldsymbol{\eta}$ where $\eta_{i,j} = 1$ if and only if $|i - j| < 3$. For example, if we start with a braided code where $L = 12$, then the matrix $\boldsymbol{\eta}$ for the corresponding half-braided code is given by $\boldsymbol{\eta} = 3\boldsymbol{B}'$, where $\boldsymbol{B}'$ is given in (9). △

An interesting question is how symmetric PCs perform when compared to their nonsymmetric counterparts. Partial answers to this question are given in [17] and [11]. For example, it is shown in [17] that, depending on the parameters, HPCs can have a larger normalized minimum distance than the PC from which they are derived. For the half-braided codes discussed in Example 5, a comparison with staircase codes and regular braided codes can be found in [11]. The comparison is based on the derived DE equations and supplemented with an error floor analysis. It is shown that half-braided codes can outperform both staircase codes and regular braided codes in the waterfall region, at a lower error floor and decoding delay. In general, symmetric PCs appear to be interesting candidates for further theoretical investigation and also implementation in practical communication systems.

### C. Component Code Mixtures

In the construction of $\mathcal{C}_n(\boldsymbol{\eta})$, it is assumed that each CN corresponds to a $t$-erasure correcting component code. More generally, one may wish to assign different erasure-correcting capabilities to the component codes associated with the CNs. One example is given by a PC where the row and column codes can correct a different number of erasures. If the erasure-correcting capabilities also vary across the row (or column) codes, one obtains a so-called irregular PC [20], [21].

In order to formalize this concept in the context of the deterministic GPC construction, assume that $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{t_{\text{max}}})^\intercal$ is a probability vector (i.e., $\mathbf{1}^\intercal\boldsymbol{\tau} = 1$ and $\boldsymbol{\tau} \succeq 0$). We let $\tau_t$ be the fraction of CNs at each position that can correct $t$ erasures, where $t_{\text{max}}$ is the maximum erasure-correcting capability. We further define the average erasure-correcting capability as $\bar{t} \triangleq \sum_{t=1}^{t_{\text{max}}} t\tau_t$. The assignment of the erasure-correcting capabilities to the component codes can be done in different ways. For example, we can do the assignment deterministically if $\tau_t d$ is an integer for all $t$, or independently at random according to the distribution $\boldsymbol{\tau}$. In both cases, $\boldsymbol{\tau}$ manifests itself in the DE equation (5) by changing the function $h$ to $h(x) = \sum_{t=1}^{t_{\text{max}}} \tau_t \Psi_{\geq t}(cx)$, see [10] for details.

The resulting GPCs now depend on $\boldsymbol{\tau}$ and this change is reflected in our notation by writing $\mathcal{C}_n(\boldsymbol{\eta}, \boldsymbol{\tau})$. For a fixed $\boldsymbol{\eta}$, we

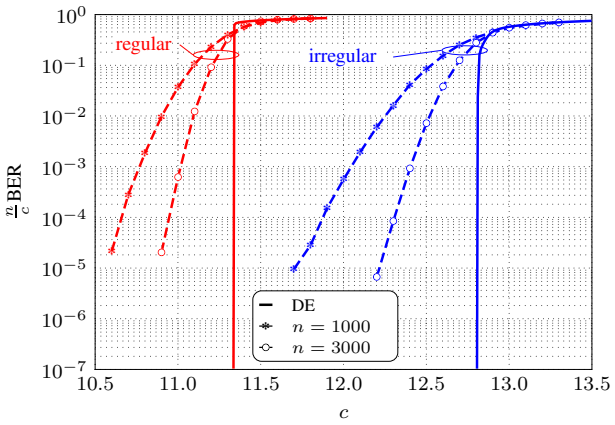



Fig. 3. Simulation results (dashed) for regular and optimized irregular HPCs for two values of $n$ and $\ell = 100$. DE results (solid) are shown for $\ell = 100$.

are interested in finding "good" distributions $\boldsymbol{\tau}$, in the sense that they lead to large decoding thresholds for $\mathcal{C}_n(\boldsymbol{\eta}, \boldsymbol{\tau})$.

*Example* 7. For $L = 1$, $\boldsymbol{\eta} = 1$, and $\gamma = 1$, we refer to the resulting code $\mathcal{C}_n(\boldsymbol{\eta}, \boldsymbol{\tau})$ as an irregular HPC. This case is considered in detail in [10]. It is shown that the performance of HPCs can be improved by employing component codes with different strengths. Using an approach based on linear programming and fixing the average erasure-correcting capability to be $\bar{t} = 7$, we obtain the optimized distribution

$$\tau_4 = 0.495, \quad \tau_9 = 0.029, \quad \tau_{10} = 0.476. \tag{10}$$

The decoding threshold is given by $\bar{c} \approx 12.88$ compared to $\bar{c} \approx 11.34$ for a regular HPC with $\tau_7 = 1$. Fig. 3 shows simulation results for $n = 1000$ and $n = 3000$ together with the DE prediction, where we used $\ell = 100$. The performance gain predicted by DE is similar to what is achieved for finite lengths. Note that the figure shows a scaled bit error rate (BER) plotted against the effective channel quality $c$ in order to better illustrate the convergence of the simulation results towards the asymptotic DE curve for increasing $n$, see [10, Sec. II-D] and [10, Sec. VII-E] for details. △

*Example* 8. For spatially-coupled PCs, one may use the approach described in [19] to study iterative decoding thresholds. In particular, the decoding thresholds for the braided code family mentioned in the last paragraph of Section IV-A coincides with the so-called potential threshold defined in [19], provided that the coupling width is sufficiently large. This result is useful since it is typically easier to characterize the potential threshold (both numerically and theoretically) than the actual decoding threshold. Now, assume that we employ different component codes according to $\boldsymbol{\tau}$. In this case, the potential threshold depends on $\boldsymbol{\tau}$. In [12, Th. 2], it is proved that for a fixed $\bar{t} \in \{2, 3, \dots\}$, the potential threshold is maximized by a regular distribution where $\tau_{\bar{t}} = 1$. From this, we can conclude that employing component code mixtures for spatially-coupled PCs is not beneficial from an asymptotic point of view. △

## V. Conclusion

A deterministic construction of GPCs is reviewed along with a DE analysis for code sequences. As an application, these results are used to design and analyze spatially-coupled PCs, symmetric GPCs, and GPCs with component code mixtures.

## References


[1] P. Elias, "Error-free coding," *IRE Trans. Inf. Theory*, vol. 4, no. 4, pp. 29–37, Apr. 1954.

[2] R. Tanner, "A recursive approach to low complexity codes," *IEEE Trans. Inf. Theory*, vol. 27, no. 5, pp. 533–547, Sep. 1981.

[3] B. P. Smith, A. Farhood, A. Hunt, F. R. Kschischang, and J. Lodge, "Staircase codes: FEC for 100 Gb/s OTN," *J. Lightw. Technol.*, vol. 30, no. 1, pp. 110–117, Jan. 2012.

[4] J. Justesen, "Performance of product codes and related structures with iterated decoding," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 407–415, Feb. 2011.

[5] Y.-Y. Jian, H. D. Pfister, K. R. Narayanan, R. Rao, and R. Mazahreh, "Iterative hard-decision decoding of braided BCH codes for high-speed optical communication," in *Proc. IEEE Glob. Communication Conf. (GLOBECOM)*, Atlanta, GA, 2014.

[6] L. M. Zhang and F. R. Kschischang, "Staircase codes with 6% to 33% overhead," *J. Lightw. Technol.*, vol. 32, no. 10, pp. 1999–2002, May 2014.

[7] C. Häger, A. Graell i Amat, H. D. Pfister, A. Alvarado, F. Brännström, and E. Agrell, "On parameter optimization for staircase codes," in *Proc. Optical Fiber Communication Conf. (OFC)*, Los Angeles, CA, 2015.

[8] M. G. Luby, M. Mitzenmacher, and M. A. Shokrollahi, "Analysis of random processes via and-or tree evaluation," in *Proc. 9th Annual ACM-SIAM Symp. Discrete Algorithms*, San Franscisco, CA, 1998.

[9] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[10] C. Häger, H. D. Pfister, A. Graell i Amat, and F. Brännström, "Density evolution for deterministic generalized product codes on the binary erasure channel," *submitted to IEEE Trans. Inf. Theory*, 2015. [Online]. Available: http://arxiv.org/pdf/1512.00433.pdf

[11] ——, "Density evolution and error floor analysis of staircase and braided codes," in *Proc. Optical Fiber Communication Conf. (OFC)*, Anaheim, CA, 2016.

[12] ——, "Deterministic and ensemble-based spatially-coupled product codes," 2016. [Online]. Available: http://arxiv.org/pdf/1512.09180.pdf

[13] A. J. Feltström, D. Truhachev, M. Lentmaier, and K. S. Zigangirov, "Braided block codes," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2640–2658, Jul. 2009.

[14] J. Justesen and T. Høholdt, "Analysis of iterated hard decision decoding of product codes with Reed-Solomon component codes," in *Proc. IEEE Information Theory Workshop (ITW)*, Tahoe City, CA, 2007.

[15] B. Bollobás, S. Janson, and O. Riordan, "The phase transition in inhomogeneous random graphs," *Random Structures and Algorithms*, vol. 31, no. 1, pp. 3–122, Aug. 2007.

[16] Y.-Y. Jian, H. D. Pfister, and K. R. Narayanan, "Approaching capacity at high rates with iterative hard-decision decoding," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Cambridge, MA, 2012.

[17] H. D. Pfister, S. K. Emmadi, and K. Narayanan, "Symmetric product codes," in *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, 2015.

[18] S. Kudekar, T. Richardson, and R. Urbanke, "Threshold saturation via spatial coupling: Why convolutional LDPC ensembles perform so well over the BEC," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 803–834, Feb. 2011.

[19] A. Yedla, Y.-Y. Jian, P. S. Nguyen, and H. D. Pfister, "A simple proof of Maxwell saturation for coupled scalar recursions," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6943–6965, Nov. 2014.

[20] S. Hirasawa, M. Kasahara, Y. Sugiyama, and T. Namekawa, "Modified product codes," *IEEE Trans. Inf. Theory*, vol. 30, no. 2, pp. 299–306, Mar. 1984.

[21] M. Alipour, O. Etesami, G. Maatouk, and A. Shokrollahi, "Irregular product codes," in *Proc. IEEE Information Theory Workshop (ITW)*, Lausanne, Switzerland, 2012.

# Locally Repairable Codes with Availability and Hierarchy: Matroid Theory via Examples

Ragnar Freij-Hollanti
Department of Communications and Networking
P.O.Box 13000, FI-00076 AALTO, Finland
Email: ragnar.freij@aalto.fi

Thomas Westerbäck and Camilla Hollanti
Department of Mathematics and Systems Analysis
P.O.Box 11100, FI-00076 AALTO, Finland
Emails: {firstname.lastname}@aalto.fi

***Abstract*—Recent research on distributed storage systems (DSSs) has revealed interesting connections between matroid theory and locally repairable codes (LRCs). The goal of this paper is to illustrate these as well as some new — rather technical in nature — results via simple examples. The examples embed all the essential features of LRCs, namely locality, availability, and hierarchy alongside with related generalized Singleton bounds.**

## I. Introduction

The need for large-scale data storage is continuously increasing. Within the past few years, *distributed storage systems* (DSSs) have revolutionized our traditional ways of storing, securing, and accessing data. Storage node failure is a frequent obstacle in large-scale DSSs, making repair efficiency an important objective. A bottle-neck for repair efficiency, measured by the notion of *locality* [1], is the number of contacted nodes needed for repair. The key objects of study in this paper are *locally repairable codes* (LRCs), which are, informally speaking, storage systems where a small number of failing nodes can be recovered by boundedly many other (close-by) nodes. Repair-efficient LRCs are already in use for large-scale DSSs used by, for example, Facebook and Windows Azure Storage [2].

Another desired attribute, measured by the notion of *availability* [3], is the property of having multiple alternative ways to repair nodes. This is particularly relevant for nodes containing so-called hot data that is frequently and simultaneously accessed by many users. Moreover, as failures are often spatially correlated, it is valuable to have each node repairable at several different *scales*. This means that if a node fails simultaneously with the set of nodes that should normally be used for repairing it, then there still exists a larger set of helper nodes that can be used to recover the lost data. This property is captured by the notion of *hierarchy* [4], [5] in the storage system.

In this paper, we consider the hierarchical availability of linear LRCs. Our main mathematical tools for analyzing linear LRCs come from matroid theory. A *matroid* is an abstract structure in algebraic combinatorics. Matroids have been successfully used to solve problems in many areas in mathematics and computer science [6], [7], [8], [9], [10].

This work was partially supported by the Academy of Finland grants #276031, #282938, #283262. The support from the European Science Foundation under the COST Action IC1104 is also gratefully acknowledged.

*a) Related Work:* Network coding techniques for large-scale DSSs were considered in [11]. Since then, a plethora of research on DSSs with a focus on linear LRCs and various localities has been carried out, see [12], [1], [13], [14], [15] among many others. Availability for linear LRCs was defined in [3]. The notion of hierarchical locality was first studied in [4], where bounds for the global minimum distance were also obtained.

Let us denote by $(n, k, d, r, \delta, t)$, respectively, the *code length, dimension, global minimum distance, locality, local minimum distance,* and *availability*. Bold-faced parameters $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$ will be used in the sequel to refer to hierarchical locality and availability. It was shown in [2] that the $(r, \delta = 2)$-locality of a linear LRC is a matroid invariant. The connection between matroid theory and linear LRCs was examined in more detail in [17]. In addition, the parameters $(n, k, d, r, \delta)$ for linear LRCs were generalized to matroids, and new results for both matroids and linear LRCs were given therein. Even more generally, the parameters $(n, k, d, r, \delta, t)$ were generalized to polymatroids, and new results on the parameters $(n, k, d, r, \delta, t)$ for matroids and nonlinear LRCs were derived in [16].

*b) Contributions and Notation:* The main purpose of this paper is to give an overview of the connection between matroid theory and linear LRCs with availability and hierarchy via examples. In particular, we are focusing on how the parameters $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$ of a LRC can be analyzed using the *lattice of cyclic flats* of an associated matroid, and on a construction derived from matroid theory that provides us with linear LRCs. The results reviewed here were mostly derived in [17], [16]. In addition, we provide a glance at our recently submitted work [5]. The following notation will be used throughout the paper:

| | |
|---|---|
| $\mathbb{F}$: | a field; |
| $\mathbb{F}_q$: | the finite field of prime power size $q$; |
| $E$: | a finite set; |
| $G$: | a matrix over $\mathbb{F}$ with columns indexed by $E$; |
| $G(X)$: | the matrix obtained from $G$ by the columns indexed by $X$, where $X \subseteq E$; |
| $\mathrm{C}(G)$: | the vector space generated by the columns of $G$; |
| $\mathrm{R}(G)$: | the vector space generated by the rows of $G$; |
| $C$: | linear code $C = \mathrm{R}(G)$ over $\mathbb{F}$ generated by $G$; |
| $C_X$: | the *punctured* code of $C$ on $X$, *i.e.*, $C_X = \mathrm{R}(G(X))$, where $X \subseteq E$; |
| $2^S$: | the collection of all subsets of a finite set $S$; |
| $[j]$: | the set $\{1, 2, \ldots, j\}$ for an integer $j$. |

The motivation to study punctured codes arises from hierarchy; the locality parameters at the different hierarchy levels correspond to the global parameters of the related punctured codes. This further leads to so-called restricted matroids.

We also point out that $G(E) = G$ and $C_E = C$. We will often index a matrix $G$ by $[n]$, where $n$ is the number of columns in $G$.

*Example 1.1:* Let $E = [7]$ and $C$ the linear code generated by the matrix $G$ over $\mathbb{F}_2$, with columns indexed by $[7]$. Then, for $X = \{5, 6, 7\}$, we have $C_X = \{(000), (110), (101), (011)\}$ with the generator matrix $G(X)$.

$$G = \begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ & 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{array}, \quad G(X) = \begin{array}{ccc} 5 & 6 & 7 \\ \hline 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}$$

## II. Preliminaries

*a) Linear Codes as Distributed Storage Systems:* A linear code $C$ can be used to obtain a DSS, where every coordinate in $C$ represents a storage node in the DSS, and every point in $C$ represents a stored data item. While one often assumes that the data items are field elements in their own right, no such assumption is necessary. However, the alphabet (to which the data items belong) must be acted upon freely by the field, for example as a vector space. Therefore, if the data items are measured in, *e.g.*, kilobytes, then we are restricted to work over fields of size not larger than about $2^{1000}$. Beside this strict upper bound on the field size, the complexity of operations also makes small field sizes — ideally even binary fields — naturally desirable.

*Example 2.1:* Let $C$ be the linear code generated by the following matrix $G$ over $\mathbb{F}_3$:

$$G = \begin{array}{c|ccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 & 2 \\ & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 2 \end{array}$$

Then, $C$ corresponds to a 9 node storage system, storing four files $(a, b, c, d)$, each of which is an element in an alphabet on which $\mathbb{F}_3$ acts freely. In this system, node 1 stores $a$, node 5 stores $a + b$, node 9 stores $a + 2b + 2d$, and so on.

Two very basic properties of any DSS are that every node can be repaired by some other nodes and that every node contains some information. We therefore give the following definition.

*Definition 2.1:* A linear $[n, k, d]$-code $C$ over a field is a non-degenerate storage code if $d \geq 2$ and there is no zero column in a generator matrix of $C$.

*b) Linear LRCs with Hierarchical Availability:* The very broad class of linear LRCs with $h$-hierarchical availability will be defined next.

*Definition 2.2:* Let $G$ be a matrix over $\mathbb{F}$ indexed by $E$ and $C$ the linear code generated by $G$. Then, for $X \subseteq E$, $C_X$ is a linear $[n_X, k_X, d_X]$-code where

$$\begin{aligned} n_X &= |X|, \\ k_X &= \operatorname{rank}(G(X)), \\ d_X &= \min\{|Y| : Y \subseteq X \text{ and } k_{X \setminus Y} < k_X\}. \end{aligned}$$

*Example 2.2:* Consider the storage code $C$ from Example 2.1. Let $Y_1 = \{1, 2, 3, 5, 6, 7\}$, $X_1 = \{1, 2, 5\}$ and $X_2 = \{2, 6, 7\}$. Then $C_{Y_1}$, $C_{X_1}$ and $C_{X_2}$ are storage codes with

$$\begin{aligned} [n_{Y_1}, k_{Y_1}, d_{Y_1}] &= [6, 3, 3], \\ [n_{X_1}, k_{X_1}, d_{X_1}] &= [3, 2, 2], \\ [n_{X_2}, k_{X_2}, d_{X_2}] &= [3, 2, 2]. \end{aligned}$$

The parameter $d_X$ is the minimum (Hamming) distance of $C_X$. We say that $C$ is an $[n, k, d]$-code with $[n, k, d] = [n_E, k_E, d_E]$.

*Definition 2.3:* Let $h \geq 1$ be an integer, and let

$$(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t}) = [(n_1, k_1, d_1, t_1), \ldots, (n_h, k_h, d_h, t_h)]$$

be a $h$-tuple of integer 4-tuples, where $k_i \geq 1$, $n_i, d_i \geq 2$, and $t_i \geq 1$ for $1 \leq i \leq h$. Then, a coordinate $x$ of a linear $[n, k, d] = [n_0, k_0, d_0]$-LRC $C$ indexed by $E$ has *$h$-level hierarchical availability* $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$ if there are $t_1$ coordinate sets $X_1, \ldots, X_{t_1} \subseteq E$ such that

- $(i)$ $x \in X_i$ for $i \in [t_1]$,
- $(ii)$ $i, j \in [t_1]$, $i \neq j \Rightarrow X_i \cap X_j = \{x\}$,
- $(iii)$ $n_{X_i} \leq n_1$, $k_{X_i} = k_1$ and $d_{X_i} \geq d_1$ for the punctured $[n_{X_i}, k_{X_i}, d_{X_i}]$-code $C_{X_i}$, for $i \in [t_1]$,
- $(iv)$ for $i \in [t_1]$, $x$ has $(h-1)$-level hierarchical availability $[(n_2, k_2, d_2, t_2), \ldots, (n_h, k_h, d_h, t_h)]$ in $C_{X_i}$.

The code $C$ above as well as all the related subcodes $C_{X_i}$ should be non-degenerate. For consistency of the definition, we say that any symbol in a non-degenerate storage code has 0-level hierarchical availability.

*Example 2.3:* Let $C$ be the code generated by the matrix $G$ in Example 2.1 and $x = 2$. Then $x$ has 2-level hierarchical availability

$$(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t}) = [(6, 3, 3, 1), (3, 2, 2, 2)].$$

This follows from Example 2.2 where $C_{Y_1}$ implies the $(6, 3, 3, 1)$-availability, and the $(3, 2, 2, 2)$-availability is implied by $C_{X_1}$ and $C_{X_2}$.

*Definition 2.4:* A subset $X \subseteq E$ has *$h$-level hierarchical availability* $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$ in $C$, if every $x \in X$ has $h$-level hierarchical availability $(\boldsymbol{r}, \boldsymbol{\delta}, \boldsymbol{t})$ in $C$.

An *information set* of a linear $[n, k, d]$-code $C$ is defined as a set $X \subseteq E$ such that $k_X = |X| = k$. Hence, $X$ is an information set of $C$ if and only if there is a generator matrix $G$ of $C$ such that $G(X)$ equals the identity matrix, *i.e.*, $C$ is systematic in the coordinate positions indexed by $X$ when generated by $G$. In terms of storage systems, this means that the nodes in $X$ together store all the information of the DSS.

*Example 2.4:* Two information sets of the linear code $C$ generated by $G$ in Example 1.1 are $\{1, 2, 3, 4\}$ and $\{1, 2, 6, 7\}$.

## III. Matroids and linear codes

In this section we give some basics about matroids and their connection to linear codes. For more information on matroids we refer the reader to [18].

*a) Matroid Fundamentals:* Matroids were introduced by Whitney in 1935 [19] in order to capture fundamental properties of independence common to various areas of mathematics.

*Definition 3.1:* A *(finite) matroid* $M = (\mathcal{I}, E)$ is a finite set $E$ and a collection of subsets $\mathcal{I} \subseteq 2^E$ such that

$(I.1)$ $\emptyset \in \mathcal{I}$,
$(I.2)$ $Y \in \mathcal{I}$, $X \subseteq Y \Rightarrow X \in \mathcal{I}$,
$(I.3)$ For all pairs $X, Y \in \mathcal{I}$ with $|X| < |Y|$, there exists $y \in Y \setminus X$ such that $X \cup \{y\} \in \mathcal{I}$.

The subsets in $\mathcal{I}$ are the *independent sets* of the matroid.

Any matrix $G$ is associated with a matroid $M[G] = (\mathcal{I}, E)$, where the columns of $G$ are indexed by $E$ and a subset $X$ of $E$ is independent if and only if the column vectors indexed by $X$ in $G$ are linearly independent.

*Example 3.1:* Let $G$ be the matrix given in Example 2.1. Then, when $M[G] = (\mathcal{I}, [9])$, we have

$$\{3, 4, 6\}, \{1, 2, 3, 8\}, \{2, 3, 4, 6\} \in \mathcal{I},$$

and $\{1, 2, 3, 7\} \notin \mathcal{I}$.

A matroid $M$ is *linear* if there exists a matrix $G$ such that $M = M[G]$. If a matroid $M$ can be represented by a matrix over a specific field $\mathbb{F}$ then $M$ is $\mathbb{F}$-linear.

*Example 3.2:* The matroid that arises from the matrix $G$ in Example 2.1 is linear. In particular, it is $\mathbb{F}$-linear for any field $\mathbb{F}$ with characteristic $\neq 2$, ensuring that $2 \neq 0$ in $\mathbb{F}$.

The $\mathbb{F}$-linearity of a matroid may depend on the field $\mathbb{F}$. Further, many matroids are not linear over any field, and it is strongly believed (but not proven) that this is true for asymptotically almost all matroids.

An alternative, but equivalent definition of a matroid is the following.

*Definition 3.2:* A *(finite) matroid* $M = (\rho, E)$ is a finite set $E$ together with a function $\rho : 2^E \to \mathbb{Z}$ such that for all subsets $X, Y \subseteq E$

$(R.1)$ $0 \leq \rho(X) \leq |X|$,
$(R.2)$ $X \subseteq Y \quad \Rightarrow \quad \rho(X) \leq \rho(Y)$,
$(R.3)$ $\rho(X) + \rho(Y) \geq \rho(X \cup Y) + \rho(X \cap Y)$.

The rank function $\rho$ and the independents sets $\mathcal{I}$ of a matroid on a ground set $E$ are linked as follows: For $X \subseteq E$,

$$\rho(X) = \max\{|Y| : Y \subseteq X \text{ and } Y \in \mathcal{I}\},$$

and $X \in \mathcal{I}$ if and only if $\rho(X) = |X|$.

For a linear matroid $M[G] = (\rho, E)$, the rank $\rho(X)$ equals the rank of $G(X)$ over the ground field $\mathbb{F}$.

*Example 3.3:* Let $G$ be the matrix given in Example 1.1. Then, $\rho(3, 4, 6) = 3$, $\rho(\{3, 4, 5\}) = 2$ and $\rho([7]) = 4$ for the linear matroid $M[G] = (\rho, [7])$.

The *restriction* of $M = (\rho, E)$ to a subset $X$ of $E$ is the matroid $M|X = (\rho_{|X}, X)$, where

$$\rho_{|X}(Y) = \rho(Y), \text{ for } Y \subseteq X. \tag{1}$$

Obviously, $M|E = M$.

For any linear matroid $M[G] = (\rho, E)$ and subset $X \subseteq E$, the restriction of $M[G]$ to $X$ equals the linear matroid on $G(X)$, *i.e.*,

$$M[G]|X = M[G(X)].$$

An important property of $M[G]|X$ is that

$$M[G]|X = M[G_X]$$

for every matrix $G_X$ whose row space equals the row space of $G(X)$.

*Example 3.4:* Let $G$ be the matrix given in Example 1.1. Then, for $X = \{5, 6, 7\}$, $M[G]|X = M[G_X]$ where

$$G_X = \begin{array}{|c|c|c|}\hline 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline \end{array}.$$

Indeed, observe that the row $(0, 1, 1)$ is obtainable as the sum of these two rows, as $G$ is a matrix over $\mathbb{F}_2$.

We need a few more concepts from matroid theory. Let $M = (\rho, E)$ be a matroid and $X$ a subset of $E$. The subset $X$ is a *circuit* if it is dependent and all proper subsets of $X$ are independent, *i.e.*, $\rho(X) = |X| - 1$ and $\rho(Y) = |Y|$ for all $Y \subsetneq X$. A *cyclic* set is a (possibly empty) union of circuits. Equivalently, $X$ is *cyclic* if for every $x \in X$

$$\rho(X \setminus \{x\}) = \rho(X).$$

Let us define the operation $\mathrm{cyc} : 2^E \to 2^E$ by

$$\mathrm{cyc}(X) = X \setminus \{x \in X : \rho(X \setminus \{x\}) < \rho(X)\}.$$

Then $X$ is cyclic if and only if $\mathrm{cyc}(X) = X$.

Dually, we define the *closure* of $X$ to be

$$\mathrm{cl}(X) = X \cup \{y \in E \setminus X : \rho(X \cup \{y\}) = \rho(X)\}$$

and say that $X$ is a *flat* if $\mathrm{cl}(X) = X$. Therefore, $X$ is a *cyclic flat* if

$$\rho(X \setminus \{x\}) = \rho(X) \text{ and } \rho(X \cup \{y\}) > \rho(X)$$

for all $x \in X$ and $y \in E \setminus X$. The set of circuits, cycles and cyclic flats of $M$ are denoted by $\mathcal{C}(M)$, $\mathcal{U}(M)$ and $\mathcal{Z}(M)$, respectively. For the ease of notation, we will also use the nullity function $\eta : 2^E \to \mathbb{Z}$ on $M$, where

$$\eta(X) = |X| - \rho(X) \text{ for } X \subseteq E.$$

A maximal independent set of $M$ is called a *basis*, *i.e.*, $X$ is a basis if $\rho(X) = |X| = \rho(E)$.

Let $M[G] = (\rho, E)$ be a linear matroid. Then $X \subseteq E$ is a cyclic flat if and only if the following two conditions are satisfied

$(i)$ $\mathrm{C}(G(X)) \cap \mathrm{C}(G(E \setminus X)) = \mathbf{0}$, where $\mathbf{0}$ is the zero column,
$(ii)$ $x \in X \Rightarrow \mathrm{C}(G(X \setminus \{x\})) = \mathrm{C}(G(X))$.

*Example 3.5:* Let $G$ be the matrix given in Example 1.1. Then, with $\mathcal{C} = \mathcal{C}(M[G])$, $\mathcal{U} = \mathcal{U}(M[G])$ and $\mathcal{Z} = \mathcal{Z}(M[G])$,

$\{3, 4, 5\}, \{3, 4, 6, 7\}, \{5, 6, 7\} \in \mathcal{C}$, $\{3, 4, 5, 6, 7\} \notin \mathcal{C}$,
$\{3, 4, 5\}, \{3, 4, 6, 7\}, \{5, 6, 7\}, \{3, 4, 5, 6, 7\} \in \mathcal{U}$,
$\{3, 4, 5\}, \{5, 6, 7\}, \{3, 4, 5, 6, 7\} \in \mathcal{Z}$, $\{3, 4, 6, 7\} \notin \mathcal{Z}$.

*b) Linear Matroids and Codes:* There is a straightforward connection between linear codes and matroids. Indeed, let $C$ be a linear code generated by a matrix $G$. Then $C$ is associated with the matroid $M[G] = (\rho, E)$. As two different generator matrices of $C$ have the same row space, they will generate the same matroid. Therefore, without any inconsistency, we denote the associated linear matroid of $C$ by $M_C = (\rho_C, E)$. In general, there are many different codes $C \neq C'$ with the same matroid structure $M_C = M_{C'}$.

A property of linear codes that depends only on the matroid structure of the code is called *matroid invariant*. For example, the collection of information sets and the parameters $[n, k, d]$ of a code are matroid invariant properties.

In addition to the parameters $[n, k, d]$ of a linear code $C$, we are also interested in the length, rank and minimum distance of the punctured codes, since these corresponds to the locality parameters at the different hierarchy levels. A punctured code can be analyzed using matroid restrictions, since $M_{C|X} = M_C|X$ for every coordinate subset $X$. Thus, the parameters $[n_X, k_X, d_X]$ of $C_X$ are also matroid invariant properties for$C$.

*Proposition 3.1:* Let $C$ be a linear $[n, k, d]$-code and $X \subseteq E$. Then for $M_C = (\rho_C, E)$,

$(i)$ $n_X = |X|$,
$(ii)$ $k_X = \rho_C(X)$,
$(iii)$ $d_X = \min\{|Y| : Y \subseteq X,\ \rho_C(X \setminus Y) < \rho_C(X)\}$,
$(iv)$ $X$ is an information set of $C \iff$
$X$ is a basis of $M_C \iff \rho(X) = |X| = k$.

*Example 3.6:* Let $C$ denote the $[n, k, d]$-code generated by the matrix $G$ given in Example 2.1. Then $[n, k, d] = [9, 4, 3]$, where the value of $d$ arises from the fact that $\rho_C([9]\setminus\{i, j\}) = 4$ for $i, j = 1, 2, \ldots, 7$, and $\rho_C([9] \setminus \{4, 8, 9\}) = 3$. Two information sets of $C$ are $\{1, 2, 3, 4\}$ and $\{1, 2, 6, 8\}$.

Not every property of a linear code is matroid invariant, an important counter-example being the covering radius [20].

## IV. Cyclic Flats and Linear $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$-LRCs

Our main matroid theoretical tool in this paper for analyzing linear LRCs is the lattice of cyclic flats, together with the rank function restricted to this lattice.

*a) The Lattice of Cyclic Flats:* A collection of sets $\mathcal{P} \subseteq 2^E$ ordered by inclusion defines a poset $(\mathcal{P}, \subseteq)$. Let $X$ and $Y$ denote two elements of $\mathcal{P}$. The elements $X$ and $Y$ have a *join* if there is an element $Z \in \mathcal{P}$, denoted by $X \vee Y$, such that $X \subseteq Z$, $Y \subseteq Z$, and if $W \in \mathcal{P}$, $X \subseteq W$, $Y \subseteq W$, then $Z \subseteq W$.

Dually, the elements $X$ and $Y$ have a *meet* if there is an element $Z \in \mathcal{P}$, denoted by $X \wedge Y$, such that $Z \subseteq X$, $Z \subseteq Y$, and if $W \in \mathcal{P}$, $W \subseteq X$, $W \subseteq Y$, then $W \subseteq Z$.

The poset $(\mathcal{P}, \subseteq)$ is a lattice if every pair of elements in $\mathcal{P}$ has a join and a meet. The bottom and top elements of a finite lattice $(\mathcal{P}, \subseteq)$ always exist, and are denoted by $1_{\mathcal{P}} = \bigvee_{X \in \mathcal{P}} X$ and $0_{\mathcal{P}} = \bigwedge_{X \in \mathcal{P}} X$, respectively.

Now, recall that for a matroid $M = (\rho, E)$ and $X \subseteq E$, the collection of cyclic flats of $M$ is denoted by $\mathcal{Z}(M)$, and consists of all $X \subseteq E$ such that $\rho(X \setminus \{x\}) = \rho(X)$ for all $x \in X$ and $\rho(X \cup \{y\}) > \rho(X)$ for all $y \in E \setminus X$.

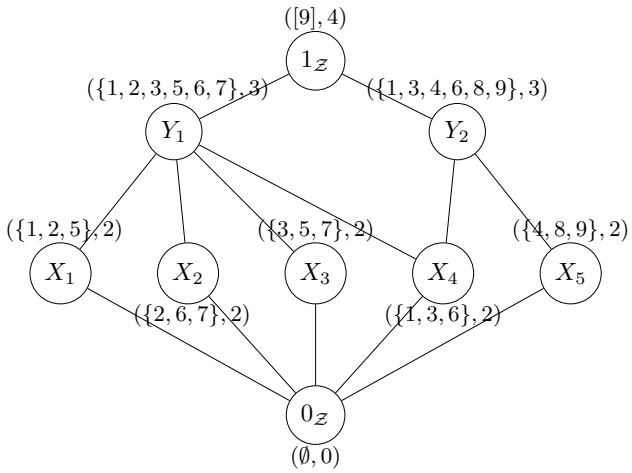
Two basic properties of the cyclic flats of a matroid are given in the following proposition.

*Proposition 4.1 ([21]):* Let $M = (\rho, E)$ be a matroid and $\mathcal{Z}$ the collection of cyclic flats of $M$. Then,

(i) $\rho(X) = \min\{\rho(F) + |X \setminus F| : F \in \mathcal{Z}\}$, for $X \subseteq E$,
(ii) $(\mathcal{Z}, \subseteq)$ is a lattice, $X \vee Y = \mathrm{cl}(X \cup Y)$ and $X \wedge Y = \mathrm{cyc}(X \cap Y)$ for $X, Y \in \mathcal{Z}$.

Proposition 4.1 (i) shows that a matroid is uniquely determined by its cyclic flats and their ranks.

*Example 4.1:* Let $M_C = (\rho_C, E)$ be the matroid associated to the linear code $C$ generated by the matrix $G$ given in Example 2.2. The lattice of cyclic flats $(\mathcal{Z}, \subseteq)$ of $M_C$ is given in the figure below, where the cyclic flat and its rank are given at each node.




In [21], Theorem 3.2 gives an axiom scheme for matroids via cyclic flats and their ranks. This gives a compact way to construct matroids with prescribed local parameters, which we have exploited in [17].

*b) Properties of Linear LRCs via the Lattice of Cyclic Flats:* The results given in this section can be found in [17], [5].

For a linear $[n, k, d]$-code $C$ with $M_C = (\rho_C, E)$ and $\mathcal{Z} = \mathcal{Z}(M_C)$, and for a coordinate $x$, we have

$(i)$ $d \geq 2 \iff 1_{\mathcal{Z}} = E$,
$(ii)$ $C_{\{x\}} \neq \{0_{\mathbb{F}}\}$ for every $x \in E \iff 0_{\mathcal{Z}} = \emptyset$.

Hence, by Definition 2.1, the following propositions are straightforward.

*Proposition 4.2:* Let $C$ be a linear $[n, k, d]$-code and $\mathcal{Z}$ denote the collection of cyclic flats of the matroid $M_C = (\rho_C, E)$. Then $C$ is a non-degenerate storage code if and only if $0_{\mathcal{Z}} = \emptyset$ and $1_{\mathcal{Z}} = E$.

*Proposition 4.3:* Let $C$ be a non-degenerate storage code and $M_C = (\rho_C, E)$. Then, for $X \subseteq E$, $C_X$ is a non-degenerate storage code if and only if $X$ is a cyclic set of$M_C$.

If $X$ is a cyclic flat[1] of a matroid $M$, then $\mathcal{Z}(M|X) = \{F \in \mathcal{Z}(M) : F \subseteq X\}$. Therefore, studying the parameters $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$ of the punctured codes $C_{X_i}$ amounts to studying the *order ideals* $\{F \in \mathcal{Z}(M) : F \subseteq X_i\}$ in the lattice of cyclic flats $\mathcal{Z}(M_C)$.

*c) Constructions of Linear $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$-LRCs:* In [17], a construction of a broad class of linear LRCs is given via matroid theory. This is generalized in [16] and[5] to account for availability and hierarchy, respectively. Here, we only sketch the key elements of the construction.

First, let $F_1, \ldots, F_m$ be a collection of subsets of a finite set $E$. Assign a function $\rho : \{F_i\} \cup \{E\} \to \mathbb{Z}$ satisfying

$$\begin{array}{ll} (i) & 0 < \rho(F_i) < |F_i| \text{ for } i \in [m], \\ (ii) & \rho(E) \leq |F_{[m]}| + \sum_{i \in [m]} (\rho(F_i) - |F_i|), \\ (iii) & j \in [m] \Rightarrow |F_{[m]\setminus\{j\}} \cap F_j| < \rho(F_j). \end{array}$$

Now, denoting $F_I = \cup_{i \in I} F_i$, we can extend $\rho$ to $\{F_I\} \to \mathbb{Z}$, by

$$\rho(F_I) = \min(|F_I| + \sum_{i \in i} \rho(F_i) - |F_i|, \rho(E)).$$

Ignoring the sets $F_I$ with $I \neq [m]$ and $\rho(F_I) = \rho(F_I) = \rho(E)$, we have thus constructed a lattice of cyclic flats with prescribed parameters, following the axiomatic scheme from [21].

*d) Some Classes of Linear LRCs:* The class of linear $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$-LRCs is a very general class. Almost all existing literature on linear LRCs focuses on the case where one or more of the parameters are specialized. Two well studied subclasses are 1-level linear $(n_1 = r+1, k_1 = r, d_1 = \delta = 2, t_1)$-LRCs and $(n_1 = r + \delta - 1, k_1 = r, d_1 = \delta, t_1 = 1)$-LRCs over an information set or over all code symbols. Cases with $t \geq 2$ or $h \geq 2$ are not as well studied as of yet. However, matroid theory and especially the lattice of cyclic flats seem to provide the required tools for the whole class of linear $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$-LRCs.

In particular, the class of 1-level linear $(n, k, d, r, \delta, t = 1)$-LRCs was examined in [17], and $h$-level linear $(n, k, d, r, \delta, t = 1)$-LRCs in [4]. The $h$-level case was later generalized to matroids in [5], and generalised Singleton bounds were given for matroids. This implies, as special cases, the same bounds on linear LRCs and other objects related to matroids, *e.g.* graphs, almost affine LRCs, and matchings. By generalized Singleton bounds we mean results that upper bound the minimum distances $d_i$ in terms of the other parameters $(\boldsymbol{n}, \boldsymbol{k}, \boldsymbol{d}, \boldsymbol{t})$. The most general Singleton bound for matroids with hierarchy in the case $\boldsymbol{t} = \boldsymbol{1}$ are the following given for linear codes in [4] and for matroids in [5]:

$$d_i(M) \leq n_i - k_i + 1 - \sum_{j>i} (d_j - d_{j+1}) \left( \left\lceil \frac{k_i}{k_j} \right\rceil - 1 \right),$$

where we say $d_{h+1} = 1$.

Moreover, results on nonlinear 1-level $(n_1, k_1, d_1, t_1)$-LRCs over arbitrary alphabets is given in [16], where generalizations of matroids, in particular polymatroids, are used to derive corresponding results for matroids and linear LRCs. The most general Singleton bound in the regime $t \neq 1$, $h = 1$, with all-symbol locality and information-symbol availability is

$$d_1 \leq n - k + 1 - \left( \left\lceil \frac{t_1(k-1)+1}{t_1(r_1-1)+1} \right\rceil - 1 \right) (\delta_1 - 1),$$

also given in [16].

As a natural next step, the notions of hierarchy and availability should be studied further from the matroid theoretic perspective, and the related level-specific generalized Singleton bounds should follow. Moreover, adaptions of our methods to account for field size should be studied.

[1] In addition, looking at cyclic sets $X'$ within a cyclic flat $X$ can give us repair groups with $n'_X \leq n_X, r'_X = r_X, d'_X = d_X$. This will be of use when looking at information set locality instead of all-symbol locality.

## REFERENCES


[1] D. S. Papailiopoulos, and A. G. Dimakis, "Locally repairable codes," *2012 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2771–2775.

[2] I. Tamo, D. S. Papailiopoulos, A. G. Dimakis, "Optimal locally repairable codes and connections to matroid theory," *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1814–1818.

[3] A. S. Rawat, D. S. Papailiopoulos, A. G. Dimakis, and S. Vishwanath "Locality and availability in distributed storage", *arXiv: 1402.2011v1*, 2014.

[4] B. Sasidharan, G. K. Agarwal, and P. W. Kumar "Codes with hierarchical locality", *arXiv: 1501.06683v1*, 2015.

[5] R. Freij-Hollanti, T. Westerbäck, and C. Hollanti "Weight enumeration of codes with hierarchical locality and availability", *arXiv: 16xx.xxxxx* (precise entry not yet available at the time of submission), 2016.

[6] J. Edmonds, *Matroids and the greedy algorithm*, Mathematical Programming **1** (1971), 127–136.

[7] N. Kashyap, "A decomposition theory for binary linear codes", *IEEE Trans. Inf. Theory*, 54, pp. 3035–3058, 2008.

[8] R. Dougherty, C. Freiling, and K. Zeger, "Networks, matroids, and non-Shannon information inequalities", *IEEE Trans. Inf. Theory*, 53(6), pp. 1949–1969, 2007.

[9] J. Martí-Farré and C. Padró, "On secret sharing schemes, matroids and polymatroids", In S. Vadhan edt., *4th Theory of Crypt. Conf. TCC 2007, Lecture Notes in Computer Science*, vol. 4392, pp. 253–272, 2007.

[10] S. El Rouayheb, A.Sprintson, and C. Georghiades, 'On the index coding problem and its relation to network coding and matroid theory", *IEEE Trans. Inf. Theory*, 56(7), pp. 3187–3195, 2010.

[11] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, 56(9), pp. 4539–4551, 2010.

[12] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory*, 58(11), pp. 6925–6934, 2012.

[13] N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar, "Optimal linear codes with a local-error-correction property," *2012 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2776 – 2780, 2012.

[14] N. Silberstein, A. S. Rawat, O. O. Koyluoglu, and S. Vishwanath, "Optimal locally repairable codes via rank-metric codes," *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1819–1823.

[15] I. Tamo and A. Barg, "A family of optimal locally recoverable codes" *IEEE Trans. Inf. Theory*, 60(8), pp. 4661–4676, 2014.

[16] T. Westerbäck, R. Freij, C. Hollanti "Applications of polymatroid theory to distributed storage systems", *arXiv: 1510.02499*, 2015.

[17] T. Westerbäck, R. Freij, T. Ernvall, and C. Hollanti "On the combinatorics of locally repairable codes via matroid theory", *arXiv: 1501.00153*, 2015.

[18] J. Oxley, "Matroid Theory" *2:ed, Oxford Graduate Texts in Mathematics, 21.* OxfordUniversity Press, 2011.

[19] H. Whitney, 'On the abstract properties of linear dependence", *Amer. J. Math*, 57, pp. 509–533, 1935.

[20] T. Britz and C. G. Rutherford, "Covering radii are not matroid invariants", *Discrete Math.*, 296, pp. 117–120, 2005.

[21] J. E. Bonin and A. de Mier, "The lattice of cyclic flats of a matroid", *Annals of Combinatorics*, 12, pp. 155–170, 2008.

# Reed-Muller Codes: Thresholds and Weight Distribution

M. Mondelli, S. Kudekar, S. Kumar, H. D. Pfister, E. Şaşoğlu, and R. Urbanke

**Abstract**

We describe a new method to compare the bit-MAP and block-MAP decoding thresholds of Reed-Muller (RM) codes for transmission over a binary memoryless symmetric channel.

The question whether RM codes are capacity-achieving is a long-standing open problem in coding theory and it has recently been answered in the affirmative for transmission over erasure channels. Remarkably, the proof does not rely on specific properties of RM codes, apart from their symmetry. Indeed, the main technical result consists in showing that any sequence of linear codes, with doubly-transitive permutation groups, achieves capacity on the memoryless erasure channel under bit-MAP decoding.

A natural question is what happens under block-MAP decoding. If the minimum distance of the code family is close to linear (e.g., of order $N/\log(N)$), then one can combine an upper bound on the bit-MAP error probability with a lower bound on the minimum distance to show that the code family is also capacity-achieving under block-MAP decoding. This strategy is successful for BCH codes. Unfortunately, the minimum distance of RM codes scales only as $\sqrt{N}$, which does not suffice to obtain the desired result. Then, one can exploit further symmetries of RM codes to show that the bit-MAP threshold is sharp enough so that the block erasure probability also tends to 0. However, this technique relies heavily on the fact that the transmission is over an erasure channel.

We present an alternative approach to strengthen results regarding the bit-MAP threshold to block-MAP thresholds. This approach is based on a careful analysis of the weight distribution of RM codes. In particular, the flavor of the main result is the following: assume that the bit-MAP error probability decays as $N^{-\delta}$, for some $\delta > 0$. Then, the block-MAP error probability also converges to 0. This technique applies to the transmission over any binary memoryless symmetric channel. Thus, it can be thought of as a first step in extending the proof that RM codes are capacity-achieving to the general case.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# From LDPC Codes to LDA Lattices: A Capacity Result

Nicola di Pietro
Texas A&M University at Qatar
c/o Qatar Foundation, Education City
P.O. Box 23874, Doha, Qatar
nicola.ndp@gmail.com

Gilles Zémor
IMB, Université de Bordeaux
351, cours de la Libération
F-33405, Talence Cedex, France
zemor@math.u-bordeaux.fr

Joseph J. Boutros
Texas A&M University at Qatar
c/o Qatar Foundation, Education City
P.O. Box 23874, Doha, Qatar
boutros@ieee.org

***Abstract*—This paper contains a summary of the arguments used to show how to achieve capacity of the AWGN channel with Voronoi constellations of LDA lattices under lattice decoding. No dithering is required in the tranmission scheme and capacity is achievable with LDA lattices whose parity-check matrices have constant row and column degrees. Although most of the technical details of the proof cannot be treated here, the reader is introduced to the fundamentals and novelties of the authors' approach to the problem. The random capacity-achieving LDA ensemble is presented and the definition of $D$-goodness of a bipartite graph is given. As an example of the power of this tool for investigating LDA lattices, a lemma about their minimum Hamming distance is provided.**

## I. Introduction

This paper addresses the problem of communication over the Additive White Gaussian Noise (AWGN) channel with lattice codes and *lattice decoding*. This decoding strategy is suboptimal with respect to the maximum likelihood (ML) decoder, but its easier algorithmic nature makes it appealing for both theoretical analysis and practical implementation.

Erez and Zamir [11], [18] were the first to provide a full proof that capacity can be achieved in this context. Their solution is based on the Modulo-Lattice Additive Noise (MLAN) channel and Voronoi constellations with Construction A lattices. More recently, Belfiore and Ling [15] proposed a solution that involves a non-uniform distribution on the channel inputs and a probabilistically finite codebook.

Once the theoretical problem of non-constructively achieving capacity was solved, it left the place to the challenge of designing some constructive families of lattices adapted to iterative decoding with close-to-capacity performance. Most of the proposed families are inspired by LDPC and turbo codes [1], [21]–[23] and an interesting work about lattices based on polar codes exists [25]; the latter are also shown to be capacity-achieving.

The authors of this paper have contributed to this research domain with the introduction of two lattice families: the most recent are the *Generalized Low-Density (GLD) lattices* [3], [4]. They show great performance under iterative decoding and numerical simulations have been run in remarkably high dimensions (up to one million). Moreover, [10] provides a theoretical analysis about the possibility of achieving the so called Poltyrev capacity with infinite GLD-lattice constellations.

The second family is the one of *Low-Density Construction A (LDA) lattices*, to which this paper is entirely devoted. LDA lattices put together the strength of Construction A and LDPC codes, and their corresponding parity-check matrix is sparse. This is the key idea to reconduct their decoding to well-performing, implementable LDPC decoding algorithms. LDA lattices were referred to with this name by di Pietro *et al.* [6], who also proposed an efficient iterative decoding algorithm which yields very good performance. A theoretical analysis of the Poltyrev-capacity-achieving qualities of infinite LDA constellations was carried on by the same authors [7], [8], whereas the "goodness" properties of LDA lattices are studied in [24]. The problem of attaining capacity of the AWGN channel with finite LDA constellations was approached and solved in [9]. The main purpose of this work is to recall and partially improve the latter result. Defoliated of all technical hypotheses, our main accomplishment can be stated as follows:

**Theorem 1.** *For every* $\mathrm{SNR} > 1$*, there exists a random ensemble of LDA lattices that achieves capacity of the AWGN channel under lattice encoding and decoding.*

Notice that the restriction $\mathrm{SNR} \leq 1$ is not very constraining: for very small $\mathrm{SNR}$ there is no need of using lattice constellations for communications over the AWGN channel and classical coded binary modulations are already known to work in a more than satisfactory way [20].

For lack of space, this paper cannot contain the technical proofs that lead to our result. Its aim is only to depict the strategies and the theoretical tools that underlie them. A longer and detailed version of this paper will be published soon and a substantial part of this work is contained in [9].

### A. Structure of the paper

Section II recalls some definitions about lattice constellations. Section III presents the $D$-goodness of bipartite graphs. Our LDA ensemble is depicted in Section IV, which also describes the information transmission scheme. Section V is a summary of the main features of the proof that LDA lattices are capacity-achieving. It also contains a lemma on their minimum Hamming distance.

### B. Notation

A crucial parameter of our analysis is the prime number $p$ that underlies Construction A. We are interested in describing its growth as a function of the lattice dimension $n$. For this reason, $p$ is defined as $p = n^\lambda$ for some positive constant $\lambda$. Clearly, this is a slight abuse of notation that means, without any undesired consequence, that $p = p(\lambda)$ is the closest prime number to $n^\lambda$.

## II. Lattice constellations for the AWGN channel

We assume that the reader is familiar with lattices as mathematical objects and constellations for the transmission of information; excellent references are [5], [26]. We repeat here some definitions, mainly for fixing our notation.

We exclusively deal with real lattices, i.e. discrete additive subgroups of the Euclidean vector space $\mathbb{R}^n$. Also, we suppose that they are always full-rank and $n$ indicates both the lattice dimension and the dimension of the Euclidean space. The *Voronoi region* of a point $\mathbf{x}$ of a lattice $\Lambda$ is the set

$$\mathcal{V}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| \leq \|\mathbf{y} - \mathbf{z}\|,\ \forall \mathbf{z} \in \Lambda \setminus \{\mathbf{x}\}\}.$$

We call Voronoi region of the lattice, and denote it $\mathcal{V}(\Lambda)$, the Voronoi region of $\mathbf{0}$. The *volume* of $\Lambda$ is $\mathrm{Vol}(\Lambda) = \mathrm{Vol}\left(\mathcal{V}(\Lambda)\right)$ and its *effective radius* is the radius of the ball whose volume is equal to $\mathrm{Vol}(\Lambda)$. Consider two lattices $\Lambda$ and $\Lambda_f$; we say that they are *nested* if $\Lambda \subseteq \Lambda_f$. We call *Voronoi constellation* [12] of two nested lattices the lattice code $\mathcal{C} = \Lambda_f \cap \mathcal{V}(\Lambda)$. In this context, $\Lambda$ is often called the *shaping lattice* and $\Lambda_f$ the *fine lattice*. We can deduce that the Voronoi constellation has cardinality $\mathrm{Vol}(\Lambda)/\mathrm{Vol}(\Lambda_f)$; its elements are the represantatives of the congruence classes of $\Lambda_f/\Lambda$ with minimum norm.

**Definition 1.** *Let $C = C[n,k]_p \subseteq \mathbb{F}_p^n$ be a $p$-ary linear code of length $n$ and dimension $k$ and let us naturally embed $C$ into $\mathbb{Z}^n$. If $H$ is a parity-check matrix of $C$, we say that the lattice $\Lambda \subseteq \mathbb{R}^n$ is built with* Construction A *from $C$ when*

$$\Lambda = C + p\mathbb{Z}^n = \{\mathbf{x} \in \mathbb{Z}^n : H\mathbf{x}^T \equiv \mathbf{0}^T \bmod p\}.$$

*$H$ is called a* parity-check matrix *of $\Lambda$ as well. $\Lambda$ is called a* Low-Density Construction A *(or briefly* LDA*) lattice if it is built with Construction A from an LDPC code.*

We recall that LDPC codes are linear codes whose parity-check matrix has a great majority of zero entries [14], [20].

**Definition 2.** *Let $\mathbf{C}$ be the capacity of our channel. A family of lattice codes is* capacity-achieving *if for every $\delta > 0$ and for every $\varepsilon > 0$ there exists a lattice code in the family with rate at least $\mathbf{C} - \delta$ and decoding error probability at most $\varepsilon$.*

Let $\mathbf{x}$ be the AWGN channel input and let $\mathbf{y} = \mathbf{x} + \mathbf{w}$ be its random output, then the *Wiener coefficient* is $\alpha = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}[\|\mathbf{x} - \beta\mathbf{y}\|^2]$. The minimum in the previous formula is usually called *Minimum Mean Squared Error* and the Wiener coefficient is also called *MMSE coefficient*. It is well known that, if $\mathbb{E}[\|\mathbf{x}\|^2] = nP$ and $w_i \sim \mathcal{N}(0, \sigma^2)$ for every $i$, then $\alpha = \frac{P}{P+\sigma^2}$. We denote $Q_\Lambda(\cdot)$ the *quantizer* of a lattice $\Lambda$ associated with $\mathcal{V}(\Lambda)$: $Q_\Lambda(\mathbf{y}) = \arg\min_{\mathbf{x} \in \Lambda} \|\mathbf{y} - \mathbf{x}\|$.

**Definition 3.** *A* MMSE lattice decoder *returns $\hat{\mathbf{x}} = Q_\Lambda(\alpha\mathbf{y})$ as the channel input guess.*

Multiplication by $\alpha$ is essential for us to achieve capacity with a lattice decoder, as it was for Erez and Zamir [11], [18]. We will give a geometrical explanation of this in Section V.

## III. Expansion properties of bipartite graphs

Let $\mathcal{G} = (V_L, V_R, E)$ be an undirected bipartite graph; $V_L \cup V_R$ is its set of (left and right) vertices and $E$ its set of edges. Let $|V_L| = n$ and $|V_R| = fn$, for some constant non-zero fraction $f \in \mathbb{Q}$ (that can be bigger than 1). If $S$ is a subset of vertices of a graph $\mathcal{G}$, its *neighborhood* $N(S)$ is defined as the set of vertices of the graph that are incident to a vertex of $S$. In a bipartite graph, $N(S) \subseteq V_R$ for every $S \subseteq V_L$ and, vice versa, $N(T) \subseteq V_L$ for every $T \subseteq V_R$. We will consider only *biregular* graphs: the neighborhood of any vertex of $V_R$ (resp. $V_L$) has cardinality exactly $\Delta$ (resp. $f\Delta$). Let us denote by $\mathcal{F}(n, f, \Delta)$ the family of graphs just defined.

**Definition 4.** *Let $D$ be a positive constant. A graph of $\mathcal{F}(n, f, \Delta)$ is* $D$-good from left to right *if*

$$\forall S \subseteq V_L \text{ s.t. } |S| \leq \frac{n}{D+1},\ \text{then } |N(S)| \geq fD|S|. \quad (1)$$

*Analogously, it is* $D$-good from right to left *if*

$$\forall T \subseteq V_R \text{ s.t. } |T| \leq \frac{fn}{D+1},\ \text{then } |N(T)| \geq \frac{D|T|}{f}.$$

*We say that a graph of $\mathcal{F}(n, f, \Delta)$ is* $D$-good *if it is good both from left to right and from right to left.*

**Lemma 1.** *Let $\mathcal{G}$ be a graph chosen uniformly at random in $\mathcal{F}(n, f, \Delta)$, and let $h(\cdot)$ be the binary entropy function. If $D \geq 1$ and*

$$\Delta > \max\left\{ \left(1 + \frac{1}{f}\right)\left(1 - \frac{Dh\left(\frac{1}{D}\right)}{(D+1)h\left(\frac{1}{D+1}\right)}\right)^{-1}, D^2 + \frac{1}{f}, \frac{D^2}{f} + 1 \right\},$$

*then $\lim_{n\to\infty} \mathcal{P}\{\mathcal{G} \text{ is } D\text{-good}\} = 1$.*

The proof of the previous lemma uses the same main ideas that Bassalygo applies in [2]. The reader may also be interested in comparing this lemma with Theorem 8.7 of [20, p. 431] and reading therein about the construction of *expander codes*.

The $D$-goodness of the Tanner graphs [20] associated with LDA lattices plays an essential role in the proof of Lemma 3 and Theorem 2. The way it is exploited to adapt some random-coding arguments to the LDA case is definitely one of the most novel tools of this work.

## IV. The random LDA ensemble and the transmission scheme

Our lattice codes are given by Voronoi constellations of nested LDA lattices. First, let us fix two constants $R$ and $R_f$ such that $0 < R < R_f < 1$. Also, let us fix the constant

$\Delta_P$, which is the number of non-zero entries per row of the LDPC parity-check matrices. Our random shaping lattice $\Lambda$ is the LDA lattice generated by the following $p$-ary parity-check matrix of dimension $n(1-R)\times n$:

$$\mathbf{H}=\left(\frac{\mathbf{H}'}{\mathbf{H}_f}\right).$$

Its lower submatrix $\mathbf{H}_f$, formed by its last $n(1-R_f)$ rows, is the parity-check matrix of the random LDA fine lattice $\Lambda_f$. By construction, we impose that $\mathbf{H}$ has exactly $\Delta_P$ random entries per row and $\Delta_V=\Delta_P(1-R)$ random entries per column. Also, each column of $\mathbf{H}_f$ has exactly $\Delta_P(1-R_f)$ random entries per column. All the other entries are deterministically fixed to 0 and their position is fixed once for all, as well. The random entries of $\mathbf{H}$ are i.i.d. random variables with equiprobable values in $\mathbb{F}_p$. Of course, $\Lambda\subseteq\Lambda_f$ and the random Voronoi constellation is given by $\Lambda_f/\Lambda$. Lemma 1 guarantees that the Tanner graph associated with the fine LDA lattice $\Lambda_f$ is $D$-good for every $D\geq 1$ such that:

$$\Delta_P>\max\left\{\frac{2-R_f}{1-R_f}\left(1-\frac{Dh\left(\frac{1}{D}\right)}{(D+1)h\left(\frac{1}{D+1}\right)}\right)^{-1},\ \frac{D^2}{1-R_f}+1\right\}. \tag{2}$$

It can be shown that (2) suffices to claim that the graph associated with the shaping LDA lattice $\Lambda$ is $D$-good, too.

The points of the LDA-lattice constellation are indexed by the $p^{n(R_f-R)}$ different syndromes of the form $(s_1,s_2,\ldots,s_{n(R_f-R)},0,\ldots,0)$ associated with the matrix $\mathbf{H}$, with $s_i\in\mathbb{F}_p$. More explicitly, let $\mathbb{F}_p^{(R_f-R)}$ be the set of the messages; the bijection

$$\begin{aligned}\varphi\colon \Lambda_f\cap\mathcal{V}(\Lambda)&\to\mathbb{F}_p^{n(R_f-R)}\\ \mathbf{x}&\mapsto\mathbf{H}'\mathbf{x}^T \bmod p\end{aligned}$$

makes a constructive encoding possible. Our transmission scheme works as follows: the sender pairs up a message and a syndrome and transmits $\mathbf{x}$, the corresponding constellation point obtained via $\varphi^{-1}$, over the AWGN channel. The receiver gets the channel output $\mathbf{y}=\mathbf{x}+\mathbf{w}$; by MMSE lattice decoding of $\mathbf{y}$, he gets $\hat{\mathbf{x}}=Q_{\Lambda_f}(\alpha\mathbf{y})$. The decoded message is the one associated with $\varphi(\hat{\mathbf{x}})$. For every $\mathbf{s}'\in\mathbb{F}_p^{n(R_f-R)}$, let $\mathbf{x}\in\Lambda_f$ be any solution of the linear system $H'\mathbf{x}^T\equiv\mathbf{s}'^T \bmod p$. Then, $\varphi^{-1}(\mathbf{s}')=\mathbf{x}-Q_\Lambda(\mathbf{x})$ and encoding can be done substantially thanks to a lattice decoder, too.

Notice that our scheme differs from the others traditionally proposed in the literature about lattices. We do not transform the AWGN into a MLAN channel [11], [18] and, in particular, we do not assume that the sender and the receiver share the common randomness known as *dither*. The possibility of avoiding dithering in this context had already been pointed out by Forney [13], but no proof had ever been provided, to the best of our knowledge. Furthermore, we keep an a priori uniform disitribution on the lattice constellations and do not introduce the random Gaussian coding proposed in [15], [25].

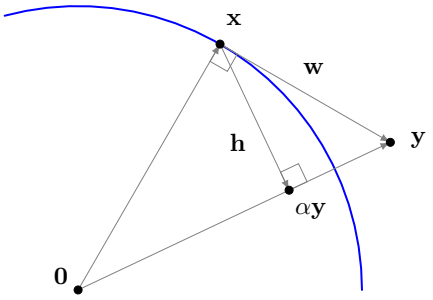



Figure 1. Geometric interpretation: $\mathbf{x}$ is the channel input; $\|\mathbf{x}\|^2=nP$. The AWG noise is $\mathbf{w}$, with norm $\|\mathbf{w}\|^2=n\sigma^2$. The channel output is $\mathbf{y}=\mathbf{x}+\mathbf{w}$. The Wiener coefficient $\alpha=\frac{P}{P+\sigma^2}$ is used for the MMSE scaling of $\mathbf{y}$ and $\alpha\mathbf{y}$ is the lattice decoder input. $\mathbf{h}$ is the effective noise after MMSE scaling.

## V. Overview and discussion on our proof

We give here a general description of our proof, by the means of a heuristic argument that does not take into account all the probabilistic and asymptotic aspects of the rigorous demonstration. With the use of the adverb "typically", we will mean "with probability tending to 1 when $n$ tends to infinity".

Our result is based on the following facts: first, the points of the LDA constellation typically lie very close to the surface of a sphere whose radius is essentially the effective radius of the shaping LDA lattice. Then, the AWG noise is typically almost orthogonal to the sent vector, in the sense that, if $\mathbf{x}$ is our transmitted constellation point and $\mathbf{w}$ is the noise, then $|\mathbf{x}\mathbf{w}^T|$ is "small enough". Furthermore, the "effective noise" due to MMSE scaling and the sent point are not decorrelated. Consequently, it is not possible to show that MMSE lattice decoding works independently of the sent point. Nevertheless, Theorem 2 is based on the fact that the number of points for which this does not happen is not big enough to perturb the average error probability of the family. Finally, we look for lattice points inside a sphere centered at the MMSE-scaled channel output with a very specific radius. Basically, there will be no decoding error if the only lattice point in this *decoding sphere* is the transmitted one.

Now, let us try to understand the geometric sense of the elements that we have just listed. So, suppose that the channel input is a point $\mathbf{x}$ whose norm is fixed to be $\|\mathbf{x}\|=\sqrt{nP}$, for some $P>0$ (Lemma 3 specifies this value). Suppose also that $\mathbf{x}\mathbf{w}^T=0$ (this is a stronger hypothesis than what the actual noise allows to assume, but it helps to understand the more general scenario); if $\mathbf{y}=\mathbf{x}+\mathbf{w}$ is the channel output, then $\|\mathbf{y}\|^2=\|\mathbf{x}\|^2+\|\mathbf{w}\|^2$. We call $\sigma^2$ the Gaussian noise variance per dimension. Basic Euclidean geometry (see Figure 1) tells us that multiplying $\mathbf{y}$ by the Wiener coefficient $\alpha$ helps in bringing the decoder input closer to the sent point.

The receiver decodes $\alpha\mathbf{y}$ and there is no decoding error if the closest lattice point to $\alpha\mathbf{y}$ is $\mathbf{x}$. We can show that this typically happens if $\mathrm{SNR}=\frac{P}{\sigma_2}>1$ and $\|\alpha\mathbf{y}-\mathbf{x}\|^2<np^{2(1-R_f)}/(2\pi e)$. Notice that the latter bound defines what we called the *decoding sphere* before. It concretely means

that our constellation tolerates an effective noise after MMSE scaling whose variance per dimension is less than $\sigma^2_{\mathrm{Pol}} = p^{2(1-R_f)}/(2\pi e)$. This value is far from being fortuitous: it is precisely the so called *Poltyrev limit* or *Poltyrev capacity* of the random infinite constellation $\Lambda_f$ [9], [16], [19]. We intuitively understand that this is the good condition on the maximum bearable noise, admitting that no problem comes from the fact that the effective noise and the sent point $\mathbf{x}$ are not decorrelated (this would be the case if we used dithering).

The condition on the signal-to-noise ratio can be simply understood with the following argument: let us call $\mathbf{h} = \alpha\mathbf{y} - \mathbf{x}$ and suppose that it takes the maximum value allowed by the Poltyrev limit: $\|\mathbf{h}\|^2 = n\sigma^2_{\mathrm{Pol}}$ (which also can be shown to correspond to the rate of the constellation that equals capacity). If we want good decoding, we need $\alpha\mathbf{y}$ to be closer to $\mathbf{x}$ than to $\mathbf{0}$, because the latter deterministically belongs to any lattice; in other terms, it is necessary that $\|\alpha\mathbf{y}\|^2 > \|\mathbf{h}\|^2$. An easy computation based on Figure 1 shows that this holds true if and only if $P > \sigma^2$ or, equivalently, $\mathrm{SNR} > 1$. This gives a first explanation why we do not treat the case $\mathrm{SNR} \leq 1$.

We prove that MMSE decoding works by a probabilistic approach, showing that almost always the only lattice point inside the *decoding sphere* $\mathcal{B}$ centered at $\alpha\mathbf{y}$ is the sent point $\mathbf{x}$. The average argument that we apply leads to the estimation of (a more elaborated version of) the following sum: $\sum_{\mathbf{z}\in\mathcal{B}\smallsetminus\{\mathbf{x}\}} \mathcal{P}\{\mathbf{z} \in \Lambda_f \mid \mathbf{x} \in \Lambda_f\}$. Decoding without errors corresponds to a sum which converges to 0. The easiest situation to deal with is when the two events $\{\mathbf{z} \in \Lambda_f\}$ and $\{\mathbf{x} \in \Lambda_f\}$ are independent, but they may not be, because the multiplication by $\alpha$ adds some correlation between $\mathbf{x}$ and the effective noise $\alpha\mathbf{y} - \mathbf{x}$. Erez and Zamir's dithering technique is a method to eliminate this correlation. In our case, there is a priori some $\mathbf{x}$ for which the probability in the previous sum turns out to be "bigger" than desired, while at the same time we need to show that the whole sum is "small". The originality of our analysis consists of deducing that the proportion of this kind of points in the constellation is very small.

Some considerable difficulties in estimating $\mathcal{P}\{\mathbf{z} \in \Lambda_f \mid \mathbf{x} \in \Lambda_f\}$ arise because the parity-check matrices of LDA lattices are sparse. These difficulties have to be treated with much care and the $D$-goodness of the associated Tanner graphs is of great help. As an example of the techniques used in the proofs of Lemma 3 and Theorem 2, we propose the following lemma:

**Lemma 2.** *Let $\Lambda_f$ be our random $n$-dimensional LDA fine lattice with $p = n^\lambda$, $D > (1-R_f)^{-1}$, and $\lambda > (D(1-R_f) - 1)^{-1}$. Suppose also that* (2) *holds true. For every $\mathbf{x} \in \Lambda_f$, let $w(\mathbf{x}) = |\{i : x_i \neq 0\}|$. Then, for every constant $\delta < D(1-R_f)/(D+1)$,*

$$\lim_{n\to\infty} \mathcal{P}\left\{\mathbf{x} \in \Lambda_f \smallsetminus p\mathbb{Z}^n \middle| w(\mathbf{x}) \leq \delta n\right\} = 0.$$

*Hence, the minimum Hamming distance of the LDPC code underlying $\Lambda_f$ is typically lower bounded by $\frac{D(1-R_f)}{(D+1)}n - o(1)$.*

*Proof:* Let $\Lambda_f = C_f + p\mathbb{Z}^n$, where $C_f$ is the random LDPC code defined by $\mathbf{H}_f$. For $\mathbf{x} \in \mathbb{F}_p^n \smallsetminus \{\mathbf{0}\}$, consider the random variables

$$X_{\mathbf{x}} = \begin{cases} 1, & \text{if } \mathbf{x} \in C_f \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad X = \sum_{\substack{\mathbf{x}\in\mathbb{F}_p^n \\ 1\leq w(\mathbf{x})\leq \delta n}} X_{\mathbf{x}}.$$

Thus, $X$ counts the number of points of $C_f$ of Hamming weight $1 \leq w(\mathbf{x}) \leq \delta n$. To conclude, it suffices to prove that

$$\lim_{n\to\infty} \mathbb{E}[X] = \lim_{n\to\infty} \sum_{\substack{\mathbf{x}\in\mathbb{F}_p^n \\ 1\leq w(\mathbf{x})\leq \delta n}} \mathcal{P}\{\mathbf{x} \in C_f\} = 0.$$

We will split the previous sum into two smaller sums and show that both of them converge to 0.

**Case 1:** $w(\mathbf{x}) \leq n/(D+1)$**.** If $\mathrm{Supp}(\mathbf{x}) = \{x_j \neq 0\}$ and $N(\mathrm{Supp}(\mathbf{x}))$ is its neighborhood in the Tanner graph associated with $\mathbf{H}_f$, notice that

$$\mathcal{P}\{\mathbf{x} \in C_f\} = \mathcal{P}\left\{\mathbf{H}_f\mathbf{x}^T \equiv \mathbf{0}^T \bmod p\right\} \overset{(a)}{\leq} \left(\frac{1}{p}\right)^{|N(\mathrm{Supp}(\mathbf{x}))|} \overset{(b)}{\leq} \left(\frac{1}{p}\right)^{D(1-R_f)|\,\mathrm{Supp}(\mathbf{x})|};$$

(a) comes from the fact that for every parity-check equation $\mathbf{h}_i$ with $i = 1, 2, \ldots, n(1-R_f)$, the events $\{\mathbf{h}_i\mathbf{x}^T \equiv \mathbf{0}^T \bmod p\}_i$ are independent; moreover, parity-check equations connected to only-0 variables are trivially satisfied. (b) is a consequence of the $D$-goodness of the Tanner graph: simply apply (1) to $S = \mathrm{Supp}(\mathbf{x})$ with $f = 1 - R_f$. Therefore,

$$\begin{aligned} \sum_{\substack{\mathbf{x}\in\mathbb{F}_p^n \\ 1\leq w(\mathbf{x})\leq n/(D+1)}} & \mathcal{P}\{\mathbf{x} \in C_f\} \\ &\leq \sum_{w=1}^{\lfloor n/(D+1)\rfloor} \sum_{\substack{\mathbf{x}\in\mathbb{F}_p^n \\ w(\mathbf{x})=w}} \left(\frac{1}{p}\right)^{D(1-R_f)w} \\ &\leq \sum_{w=1}^{\lfloor n/(D+1)\rfloor} \binom{n}{w} \left(\frac{p-1}{p^{D(1-R_f)}}\right)^w \\ &\leq \sum_{w=1}^{\lfloor n/(D+1)\rfloor} \left(n^{1-\lambda(D(1-R_f)-1)}\right)^w \to 0, \end{aligned}$$

because of the conditions on $\lambda$ and $D$.

**Case 2:** $n/(D+1) < w(\mathbf{x}) \leq \delta n$**.** Applying (1) to any $S \subseteq \mathrm{Supp}(\mathbf{x})$ of size $n/(D+1)$, the $D$-goodness of the Tanner graph implies that $|N(\mathrm{Supp}(\mathbf{x}))| \geq \frac{D(1-R_f)}{D+1}n$. Therefore,

$$\begin{aligned} \sum_{\substack{\mathbf{x}\in\mathbb{F}_p^n \\ n/(D+1)<w(\mathbf{x})\leq\delta n}} & \mathcal{P}\{\mathbf{x} \in C_f\} \\ &\leq \sum_{w=\lfloor n/(D+1)\rfloor+1}^{\lfloor \delta n\rfloor} \binom{n}{w}(p-1)^w \left(\frac{1}{p}\right)^{\frac{D(1-R_f)n}{(D+1)}} \\ &\leq n2^n p^{n\left(\delta - \frac{D(1-R_f)}{(D+1)}\right)} \to 0, \end{aligned}$$

because $\delta < D(1-R_f)/(D+1)$ by hypothesis. ∎

### A. Our two main results

The next lemma formally states that our Voronoi LDA constellation points have a very precise typical norm or, similarly, that our LDA shaping lattice has a "spherical" Voronoi region.

**Lemma 3.** *Consider a non-zero syndrome* $\mathbf{s}$ *associated with a constellation point:* $\mathbf{s} = (s_1, s_2, \ldots, s_{n(R_f-R)}, 0, \ldots, 0)$. *Suppose that* $p = n^\lambda$ *for some* $\lambda > 0$ *and let* $0 < \omega < 1$. *Fix the constant* $D$ *to be* $D > \max\left\{(1-R_f)^{-1}, 2\right\}$ *and suppose that* (2) *holds true. Let* $\rho_{\text{eff}}$ *denote the effective radius of the shaping LDA lattice* $\Lambda$. *If* $\mathbf{x}$ *is the random LDA constellation point whose syndrome is* $\mathbf{s}$ *and if*

$$\lambda > \max\left\{\frac{1}{D(1-R_f)-1}, \frac{1}{2R}, \frac{1}{1-R}, \frac{1}{D-2}, \left(1-\frac{1}{D^2-1}-\frac{1}{D(1-R)}\right)^{-1}\right\},$$

*then*

$$\lim_{n\to\infty} \mathcal{P}\left\{\rho_{\text{eff}}\left(1-\frac{1}{n^\omega}\right) \le \|\mathbf{x}\| \le \rho_{\text{eff}}\left(1+\frac{1}{n^\omega}\right)\right\} = 1.$$

**Theorem 2.** *Suppose that* $1 > R_f > R > \frac{1}{2}$. *Fix* $D > (1-R_f)^{-1}$ *and* $\Delta_P$ *that satisfies* (2). *If* $p = n^\lambda$, *with*

$$\lambda > \max\left\{\frac{1}{D(1-R_f)-1}, \frac{1}{1-R_f}, \left(1-\frac{1}{D^2-1}-\frac{1}{D(1-R_f)}\right)^{-1}\right\},$$

*then the random ensemble of nested LDA lattices presented in Section IV achieves capacity of the AWGN channel under MMSE lattice decoding, when* $\text{SNR} > 1$.

We emphasize the fact that $\Delta_P, D, R$, and $R_f$ are constant, therefore the parity-check matrices associated with our LDA lattices have constant row and column degree. For binary LDPC codes to achieve capacity of the binary symmetric channel, logarithmic row degrees are required [14], [17]. Surprisingly, in our LDA scenario this hypothesis can be relaxed.

## VI. Conclusion

We have stated the capacity-achieving properties of a particular ensemble of LDA lattices based on non-binary LDPC lattices. Our solution is innovative because it does not require the tools of the MLAN channel and of dithering. Furthermore, it is based on Voronoi lattice constellations and we do not need to introduce Gaussian coding, keeping an a priori uniform distribution over the lattice constellation.

Also, the row and column degree of the parity-check matrices that underlie our construction are reasonably small constants. The Tanner graphs associated with these matrices have some particular expansion properties that, qualitatively speaking, say that all "small enough" sets of nodes have "big enough" neighborhoods. These properties turn out to be one of the most important theoretical pillars of our analysis.

## Acknowledgment

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

## References

[1] I.-J. Baik and S.-Y. Chung, "Irregular low-density parity-check lattices," in *Proc. ISIT*, Toronto, Canada, 2008, pp. 2479-2483.
[2] L. A. Bassalygo, "Asymptotically optimal switching circuits," *Problems of Inf. Transmission*, vol. 17, no. 3, pp. 206-211, 1981.
[3] J. J. Boutros, N. di Pietro, and N. Basha, "Generalised low-density (GLD) lattices," in *Proc. ITW*, Hobart, Australia, 2014, pp.15-19.
[4] J. J. Boutros, N. di Pietro, Y.-C. Huang, "Spectral thinning in GLD lattices," in *Proc. ITA Workshop*, La Jolla (CA), USA, 2015, pp.1-9.
[5] J. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. 3rd ed., New York (NY), USA: Springer-Verlag, 1999.
[6] N. di Pietro, J. J. Boutros, G. Zémor, and L. Brunel, "Integer low-density latices based on Construction A," in *Proc. ITW*, Lausanne, Switzerland, 2012, pp.422-426.
[7] N. di Pietro, J. J. Boutros, G. Zémor, and L. Brunel, "New results in low-density integer lattices," in *Proc. ITA Workshop*, San Diego (CA), USA, 2013, pp.1-6.
[8] N. di Pietro, J. J. Boutros, G. Zémor "New results on Construction A lattices based on very sparse parity-check matrices," in *Proc. ISIT*, 2013, Istanbul, Turkey, pp.1675-1679.
[9] N. di Pietro, "On infinite and finite lattice constellations for the additive white Gaussian noise channel," Ph.D. dissertation, Inst. de Math., Univ. de Bordeaux, Bordeaux, France, 2014.
[10] N. di Pietro, N. Basha, and J. J. Boutros, "Non-binary GLD codes and their lattices," in *Proc. ITW*, Jerusalem, Israel, 2015, pp.1-5.
[11] U. Erez and R. Zamir, "Achieving $\frac{1}{2}\log(1+\text{SNR})$ on the AWGN channel with lattice encoding and decoding," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2293-2314, Oct. 2004.
[12] G. D. Forney, Jr., "Multidimensional constellations. II. Voronoi constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 941-958, Aug. 1989.
[13] G. D. Forney, Jr., "On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener," in *Proc. Commun., Control, and Computing, 2003 41st Annu. Allerton Conf. on*, Monticello (IL), USA, 2003, pp. 1-14.
[14] R. G. Gallager, *Low-density parity-check codes*. Cambridge (MA), USA: MIT Press, 1963.
[15] C. Ling and J.-C. Belfiore, "Achieving AWGN channel capacity with lattice Gaussian coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5918-5929, Oct. 2014.
[16] H.-A. Loeliger, "Averaging bounds for lattices and linear codes," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1767-1773, Nov. 1997.
[17] D. J. C. MacKay, "Good error correcting codes based on very sparse matrices," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 399-431, Mar. 1999.
[18] O. Ordentlich and U. Erez, "A simple proof for the existence of "good" pairs of nested lattices," in *Proc. IEEEI*, Eilat, Israel, 2012, pp. 1-12.
[19] G. Poltyrev "On coding without restrictions for the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 409-417, Mar. 1994.
[20] T. Richardson and R. Urbanke, *Modern coding theory*. New York, USA: Cambridge University Press, 2008
[21] M.-R. Sadeghi, A. H. Banihashemi, and D. Panario, "Low-density parity-check lattices: construction and decoding analysis," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4481-4495, Oct. 2006.
[22] A. Sakzad, M.-R. Sadeghi, and D. Panario, "Turbo lattices: construction and error decoding performance," Aug. 2011. Available: http://arxiv.org/abs/1108.1873
[23] N. Sommer, M. Feder, and O. Shalvi, "Low-density lattice codes," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1561-1585, Apr. 2008.
[24] S. Vatedka and N. Kashyap, "Some "goodness" properties of LDA lattices," in *Proc. ITW*, Jerusalem, Israel, 2015, pp. 1-5.
[25] Y. Yan, L. Liu, C. Ling, and X. Wu, "Construction of capacity-achieving lattice codes: polar lattices," Nov. 2014. Available: http://arxiv.org/abs/1411.0187
[26] R. Zamir, *Lattice coding for signals and networks*. Cambridge, United Kingdom: Cambridge University Press, 2014.

# Multilevel Coded Modulation and Lattice Construction D

Rami Zamir
Tel Aviv University
Faculty of Engineering
Tel Aviv, Israel
Email: zamir@eng.tau.ac.il

**Abstract**

Construction C (also known as Forney's multi-level code formula) forms an Euclidean code for the additive white Gaussian noise (AWGN) channel from $L$ binary code components. If the component codes are linear, then the minimum distance and kissing number are the same for all the points. However, while in the single level ($L = 1$) case it reduces to lattice Construction A, a multi-level Construction C is in general not a lattice.

We show that a two-level ($L = 2$) Construction C satisfies Forney's definition for a geometrically uniform constellation. Specifically, every point sees the same configuration of neighbors, up to a reflection of the coordinates in which the lower level code is equal to 1. In contrast, for three levels and up ($L \geq 3$), we construct examples where the distance spectrum varies between the points, hence the constellation is not geometrically uniform.

Joint work with Maiara Bollauf.

# Design Considerations for Downlink Broadcast Frame with Short Data Packets

Kasper Fløe Trillingsgaard and Petar Popovski
Department of Electronic Systems, Aalborg University
9220 Aalborg, Denmark

***Abstract*—It is almost an axiom that every cellular wireless system, including the upcoming 5G systems, should be based on data transmissions organized in frames. The frame design is based on heuristics, consisting of a frame header and data part. The frame header contains control information that specifies the sizes of the data packets and provides pointers to their location within the data part. In this paper we show that this design heuristics is suboptimal when the messages in the data part are short. We consider a downlink scenario represented by an AWGN broadcast channel with $K$ users, while the sizes of the messages to the users are random variables. Each data packet encodes a message to one user. However, if the message sizes are small, there is a significant overhead caused by the header and the data packets can not be encoded efficiently. This calls for revision of the established heuristics for framing control information and data. We show that grouping messages of multiple users allows more efficient encoding from a transmitter perspective. On the other hand, it has the undesirable implication that it requires each user to decode the messages for a whole group of users. We assume that the power spend by each user is proportional to the number of channel uses it needs to decode. Using recent results in finite blocklength analysis, we investigate the trade-offs between total transmission time from the transmitter perspective and the average power spend at each user. Our approach shows that the space of feasible protocols is significantly enlarged and thereby allows the designer to trade-off between average total transmission time and the average power spend by each user.**

## I. Introduction

Modern high-speed wireless networks heavily depend on reliable and efficient transmission of large data packets through the use of coding and information theory. The advent of machine-to-machine (M2M), vehicular-to-vehicular (V2V), and various streaming systems has spawned a renewed interest in developing information theoretical bounds and codes for communication of short packets [1][2]. Additionally, these applications often have tight reliability and latency constraints compared to a typical wireless systems today. Communication at shorter blocklengths introduces several new challenges which are not present when considering communication of larger data packets. For example, the overhead caused by control signals and header data is insignificant if large data packets are sent, and hence this overhead is often neglected in the analysis of protocols. However, more stringent latency requirements lead to shortened blocklengths for transmission, such that the size of the control information and header data may approach, or even exceed, the size of the actual data in the packet. This is especially true for multiuser systems such as broadcast channels, two-way channels, or multiple access channels, where the header data must include information about the packet structure, security, and user address information for identification purposes.

The fundamentals of communication of short packets have recently been addressed by Polyanskiy, Poor, and Verdú (2010) [3]. Here, it was shown that the maximal coding rate of a fixed-length block code in a traditional point-to-point setting is tightly approximated by

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right) \qquad (1)$$

where $C$ is the Shannon capacity, $V$ is the channel dispersion, $n$ is the blocklength, $\epsilon$ is the desired probability of error, and $Q^{-1}(\cdot)$ denotes the inverse Q-function. Approximations such as (1) are useful in the design of modern communication problems because the specifics of code selection can be neglected in the optimization of protocol parameters.

In this paper, we consider downlink transmission with an AWGN broadcast channel that consists of a transmitter and $K$ users. There is a message from the transmitter to the $k$-th user with a certain probability $q$ (in this case user $k$ is *active*). The size of the message is itself a random variable which implies that the transmitter needs to convey information about which users are active, the structure of the transmission, and sizes of the messages. An interesting observation from (1) is that larger data packets are encoded more efficiently. This introduces an interesting trade-off with two extremes: (1) in a broadcast setting one can either encode all messages in one large packet which is efficiently encoded or (2) one can encode each message separately as is the norm in modern wireless protocols. In (1) , the average total transmission time seen from the transmitter is minimized. However, all users need to receive for the whole period to be able to decode their message, which is undesirable for devices that are power-constrained. The latter approach (2), depicted on Fig. 1, uses codes which are less efficient, and thus the average total transmission time is larger. On the other hand, each user only needs to decode the information intended for that user. The key point, however, is that these design considerations enlarge the design space and enable the designer to trade-off between transmitter resources and user resources. Despite this, practically all wireless systems solely use the approach (2). The purpose of this paper is to explore this design trade-off. Specifically, by grouping multiple users together, we encode larger amount of information bits jointly, which implies that the rate at which the information bits of the groups can be encoded is larger. The disadvantage of grouping users is that each user

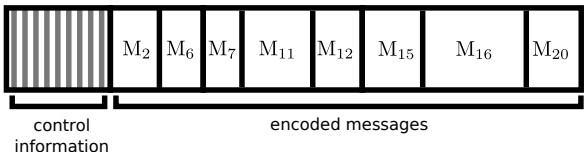



Fig. 1. Conventional approach to downlink broadcasting. An initial packet contains control information that defines the structure of the remaining part of the transmission. Each message are encoded separately.

needs receive for larger proportion of the total transmission time.

This paper is organized as follows. The next section describes the system model. Section III briefly addresses the approximations of the finite blocklength bounds, Section IV discusses design considerations for protocol design for downlink broadcast for short packets and describes our proposed protocol. Finally, we evaluate the proposed protocol in Section V and conclude the paper in Section VI.

## II. System model

We consider an AWGN broadcast channel with one transmitter and $K$ users. In the $t$-th time slot, the $k$-th user receive

$$Y_{k,t} \triangleq \sqrt{\gamma_k} X_t + Z_{k,t}. \tag{2}$$

where $Z_{k,t} \sim \mathcal{N}(0,1)$ and $X_t \in \mathbb{R}$ is the channel input. Throughout the paper, we assume $\gamma_k = \gamma$. The message $\mathrm{M}_k$ destined to the $k$-th user is nonempty with probability $q \in (0,1)$, and we say that the $k$-th user is *active* if there is a message destined to that user. The size of the message $\mathrm{M}_k$ in bits is denoted by $D_k \in \mathbb{Z}_+$, which is a discrete random variable distributed i.i.d. according to the probability mass function $P_D(\cdot)$ given by

$$P_D(d) = \begin{cases} q & \text{if } d = 0 \\ \frac{1-q}{S} & \text{if } d \in \{\alpha, \cdots, \alpha S\} \end{cases}. \tag{3}$$

for some $\alpha \in \mathbb{N}$ and $S \in \mathbb{N}$. The average message size is therefore $\mathbb{E}[D_k] = \alpha(S+1)/2$.

Based on the message sizes $D_k$, the transmitter computes the total transmission time $T$ which is also a random variable. The transmitter encodes the message $\{\mathrm{M}_k\}$ into a sequence of channel inputs using the encoder function $f_t(\mathrm{M}_1, \cdots, \mathrm{M}_K)$ such that

$$X_t \triangleq f_t(\mathrm{M}_1, \cdots, \mathrm{M}_K) \tag{4}$$

for $t \in \{1, \cdots, T\}$ and $X_t = 0$ for $t \in \{T+1, \cdots\}$.

At user $k$, we define the ON-OFF function $g_{k,t} : (\mathbb{R} \cup \{\mathrm{e}\})^{t-1} \to \{0,1\}$ that in turn defines the sequence

$$\bar{Y}_{k,t} \triangleq \begin{cases} Y_{k,t}, & g_{k,t}(\bar{Y}_k^{t-1}) = 1 \\ \mathrm{e}, & \text{otherwise} \end{cases}. \tag{5}$$

The ON-OFF function defines stopping times $T_k \triangleq \min\{n \geq 1 : \forall t > n, g_{k,t}(\bar{Y}_k^{t-1}) = 0\}$ for which we require $T_k < \infty$. Additionally, we define the decoding function $h_{k,t}(\bar{Y}_k^t)$ which estimates the message $\mathrm{M}_k$ based on $\bar{Y}_k^t$. The intuition is that a certain user can only use the channel outputs if the corresponding user is ON. This is modeled by the ON-OFF function which replaces $t$-th channel output with an erasure if the user is OFF at that time. The ON-OFF functions are causal in the sense that the decision of whether the users are ON at time $t$ depends on previous channel outputs, $\bar{Y}_k^{t-1}$. The stopping times $T_k$ represent the time index of the last nonerasure channel output in the sequence $\bar{Y}_{k,t}$. For our and all practical applications, the stopping times $T_k$ are less than or equal $T$. We merely define $T_k$ to emphasize that $T$ is a random variable which is not known by the users, and hence the users need to obtain this information through the sequence $\bar{Y}_{k,t}$. In a conventional system, control information in the initial packet defines the structure of the remaining transmission. Hence, after decoding the control information in the initial packet successfully, the user knows $T_k$ and when to be ON and OFF to receive the message intended for that user.

The ON-OFF function also defines the average power consumption of the $k$-th user which we define by

$$P_k \triangleq \mathbb{E}\left[\sum_{i=1}^{T_k} \mathbb{1}\left\{g_{k,i}(\bar{Y}_k^{i-1}) = 1\right\}\right] \tag{6}$$

where $\mathbb{1}\{\text{condition}\}$ denotes the indicator function. Note that $\mathbb{E}[P_1] = \mathbb{E}[P_k]$, for $k \in \{1, \cdots, K\}$, since the message sizes $D_k$ are distributed identically. Finally, the active users need to decode the their messages with reliability larger than or equal $1-\epsilon$ such that

$$\mathrm{P}\left[h_{k,T_k}(\bar{Y}_k^{T_k}) \neq \mathrm{M}_k | D_k > 0\right] \leq \epsilon \tag{7}$$

for $k \in \{1, \cdots, K\}$ and $\epsilon \in (0,1)$.

Our objective is to explore trade-offs between the competing goals of minimizing $\mathbb{E}[T]$ and $\mathbb{E}[P_k]$. We do this by investigating a class of feasible protocols.

## III. Finite blocklength approximation

In the analysis of the proposed protocol, we apply recent results in finite blocklength information theory. Polyanskiy, Poor, and Verdú [3] showed that the maximal achievable coding rate of a code with fixed blocklength $n$ and reliability $1-\epsilon' \in (0,1)$ over an AWGN channel is tightly approximated by

$$R^*(n,\epsilon') \approx C - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon') + \frac{1}{2}\log_2 n \tag{8}$$

where the channel capacity $C$ and the channel dispersion $V$ are given by

$$C \triangleq \frac{1}{2}\log_2(1+P) \tag{9}$$

$$V \triangleq \frac{P(P+2)}{2(P+1)^2}\log_2(\exp(1))^2 \tag{10}$$

respectively. One can obtain tight upper and lower bounds for $R^*(n,\epsilon')$ using the achievability and converse bounds in [3].

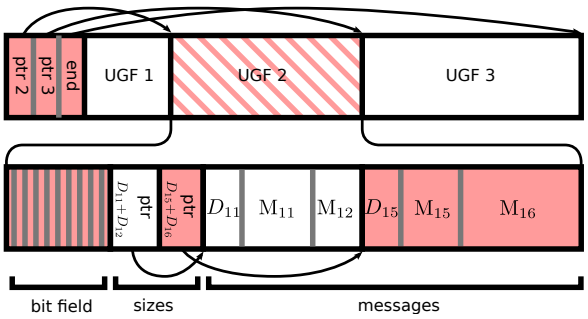



Fig. 2. Proposed protocol for $K = 30$, $L_B = 10$, and $L = 2$. In this case $\{11, 12, 15, 16\} \subseteq \mathcal{U}$ are among the active users. Grey separators means that data on both sides are encoded jointly. The red shaded regions correspond to the packets that the users 15 and 16 need to decode.

Implementations of these bounds are accessible in the short-packet communication toolbox SPECTRE for MATLAB [4]. In present paper, however, we resort to the approximation (8). An important property is that (8) is concave which implies that long packets are encoded more efficient than short packets.

We assume that a user is able to detect whether an error occurs during decoding of a packet. This assumption is suggested by the results in [5] (although this result only applies to discrete memoryless channels), and it is crucial to ensure the integrity of our protocols. In practical systems, one can use CRC checks to ensure integrity of packets. We also assume that a user needs to receive all channel uses of a packet to allow for decoding. In other words, if $k$ bits are encoded into $n$ channel uses, the user needs to receive all $n$ channel uses to decode any of the $k$ bits.

In the design of our protocol, we rely on the approximation of $R^*(n, \epsilon')$. Specifically, the transmitter divides the total transmission time $T$ into several smaller packets, each encoded separately at the maximal coding rate approximated by (8). We also define $N(k, \epsilon') \triangleq \min\{n \geq 0 : nR^*(n, \epsilon') \geq k\}$ for $k \geq 1$ and $N(0, \epsilon') \triangleq 0$, which is smallest number of channel uses that allows the encoding of $k$ bits with reliability $1 - \epsilon'$.

One can easily obtain lower bounds on $\mathbb{E}[T]$ and $\mathbb{E}[P_k]$ by assuming the messages sizes are large:

$$\mathbb{E}[T] \geq \frac{1}{C}\mathbb{E}\left[\sum_{k=1}^{K} D_k\right] = \frac{K\alpha(1-q)(S+1)}{2C} \qquad (11)$$

and

$$\mathbb{E}[P_k] \geq \frac{1}{C}\mathbb{E}[D_k] = \frac{\alpha(1-q)(S+1)}{2C}. \qquad (12)$$

For sufficiently large $\alpha$, the control information becomes negligible, and hence for the conventional approach both $\mathbb{E}[T]$ and $\mathbb{E}[P_k]$ simultanously approach the lower bounds in (11) and (12).

## IV. Protocol design

There are various ways in which the messages $\{M_k\}$ can be conveyed to the respective users. Our approach is to design a protocol in which the transmitter forms multiple packets which are encoded separately. For each of these packets, we apply the finite blocklength approximation in (8) to find the optimal rate at which they can be encoded. We assume that the users are not provided with any control information such as the active users and $\{D_k\}$. Thus, the transmitter needs to encode packets about which users are active, the packet sizes of $\mathrm{M}_k$, $D_k$, and the structure of the transmission. Clearly, this leaves us with a large space of feasible protocols. Here we introduce one class of protocols.

We first discuss what information, the transmitter needs to convey:

1) *Messages* $\{\mathrm{M}_k\}$: The message $\mathrm{M}_k$ only needs to be received by the $k$-th user, but as discussed previously, messages can be grouped and encoded jointly.
2) *Message sizes* $\{D_k\}$: The $k$-th user needs to know the message size $D_k$ before attempting to decode the actual message $\mathrm{M}_k$ (otherwise, the user does not know how many channel uses the message $\mathrm{M}_k$ takes).
3) *Reciever activity* $\mathcal{U}$: It is necessary to convey whether the $k$-th user is active. In total, it requires $K$ information bits to convey this information to all users.[1] As $K$ information bits may represent a significant overhead, it may be beneficial to encode user activity bits in multiple packets such that each user only needs to decode one such packet.

In the proposed protocol, depicted in Fig. 2, users are grouped into $\lceil K/L_B \rceil$ *user groups* with at most $L_B$ users in each user group. User acitivity, messages, and message sizes associated to each of these user groups are conveyed sequentially in *user group frames* (UGF). A transmission is initiated by a packet that jointly encodes the total transmission time (equivalent to the an end of transmission pointer) along with the $\lceil K/L_B \rceil - 1$ time indices that points to the time indices where the 2-th, 3-th, ..., and $\lceil K/L_B \rceil$-th UGF begin. This packet is transmitted with a reliability $1 - \epsilon_4$. The first UGF trivially begins after the initiating packet.

Let $\mathcal{K} \triangleq \{1, \cdots, K\}$ and let the users in the $u$-th user group be $\mathcal{K}_u \subseteq \mathcal{K}$. Then, the UGF for the $u$-th user group is constructed as follows. Initially, the transmitter divides the active users $\mathcal{U}_u \subseteq \mathcal{K}_u$ of the $u$-th user group into subgroups $\mathcal{U}_{u,i} \subseteq \mathcal{U}_u$, $i \in \{1, \cdots, \lceil |\mathcal{U}_u|/L \rceil\}$ of at most $L$ users. The transmitter and users can agree on how to partition the users into subgroups for every set $\mathcal{U}_u$. The set of users $\mathcal{U}_{u,i} \subseteq \mathcal{U}_u$ is referred to as the $i$-th subgroup of the $u$-th user group. The main idea of our protocol is to jointly encode each of the subgroups.

A UGF consists of the following types of packets

1) *Bit field packet*: A bit field, encoding the the information $\{\mathbb{1}\{D_k = 0\}\}_{k \in \mathcal{K}_u}$. Hence, the packet consists of $|\mathcal{K}_u|$ information bits which are encoded with reliability $\epsilon_1$.
2) *Size packets*: After grouping the active users of the $u$-th user group, $\mathcal{U}_u$, into $\lceil |\mathcal{U}_u|/L \rceil$ subgroups, the transmitter constructs a packet for each subgroup. For the $i$-th subgroup, the transmitter conveys a packet consisting of $\sum_{k \in \mathcal{U}_{u,i}} D_k$ along with a pointer to the packet that jointly encodes $\{\mathrm{M}_k\}_{k \in \mathcal{U}_{u,i}}$. Since $\sum_{k \in \mathcal{U}_{u,i}} D_k$ can take at most

[1] For the case $q \neq 1/2$, one can apply compression to reduce the number of information bits. This is, however, left for future work.

$L(S-1)+1$ distinct values, the size packet for the $i$-th subgroup needs to convey $\lceil \log_2(L(S-1)+1) \rceil + \text{ptr}$ information bits which are encoded with reliablity $\epsilon_2$. Here, ptr denotes the number of bits needed to convey a pointer to a time index. The size packets are transmitted sequentially.

3) *Message packets*: Next, the transmitter encodes the messages of each subgroup, $\{\mathrm{M}_k\}_{k\in\mathcal{U}_{u,i}}$, along with the messages sizes of $|\mathcal{U}_{u,i}|-1$ of the messages. We only need $|\mathcal{U}_{u,i}|-1$, since the sum of the sizes, $\sum_{k\in\mathcal{U}_{u,i}} D_k$, is already successfully received in the size packet described above. This requires $\sum_{k\in\mathcal{U}_i} D_k + (|\mathcal{U}_i|-1)\lceil \log_2 S \rceil$ information bits. These information bits are encoded with reliability $\epsilon_3$.

In order to decode the packet destined to user $k$, it needs to decode four packets successfully. If one or more of these packets are not successfully decoded, the user can not decode the packet containing the message destined to that user. Thus, the reliabilities need to be chosen such that $(1-\epsilon_1)(1-\epsilon_2)(1-\epsilon_3)(1-\epsilon_4)$ is kept above or equal to $1-\epsilon$ to fulfill the reliability constraint in (7). If the $k$-th user is inactive, it only needs to decode the initial packet containing pointers to the UGFs and the bit field packet. It thereby achieves a reliability of $(1-\epsilon_4)(1-\epsilon_1)$. We also point out that the described protocol reduces to a variant of the conventional protocol when $L=1$.

We remark that the protocol specified above is one class among a large space of feasible protocols. For small $q$ is may be beneficial to use a different approach for conveying user activity. For example, one could encode the number of active users in an initial packet and encode an additional packet with the user identification numbers. Regarding the size packets, one can also encode all size packets jointly in each UGF jointly to enhance encoding efficiency at the expense of higher average power consumption at the users.

Assuming that $L_B$ divides $K$, we may sum up the blocklengths of all the packets

$$\begin{aligned} T_{L,L_B} &= \frac{K}{L_B} N(L_B,\epsilon_1) + N\left(\frac{K}{L_B}\text{ptr},\epsilon_4\right) \\ &\quad + \frac{K}{L_B}\sum_{i=1}^{\lceil \mathcal{U}_i/L \rceil}\Bigg(N(\lceil \log_2(L(S-1)+1)\rceil + \text{ptr},\epsilon_2) \\ &\quad + N\Bigg(\sum_{k\in\mathcal{U}_i} D_k + (|\mathcal{U}_i|-1)\lceil \log_2 S\rceil,\epsilon_3\Bigg)\Bigg) \end{aligned} \tag{13}$$

For the expected power, we obtain

$$\begin{aligned} P_{L,L_B} &= N(L_B,\epsilon_1) + N\left(\frac{K}{L_B}\text{ptr},\epsilon_4\right) \\ &\quad + \sum_{i=1}^{\lceil \mathcal{U}_i/L \rceil}\Bigg(N(\lceil \log_2(L(S-1)+1)\rceil + \text{ptr},\epsilon_2) \\ &\quad + N\Bigg(\sum_{k\in\mathcal{U}_i} D_k + (|\mathcal{U}_i|-1)\lceil \log_2 S\rceil,\epsilon_3\Bigg)\Bigg). \end{aligned} \tag{14}$$

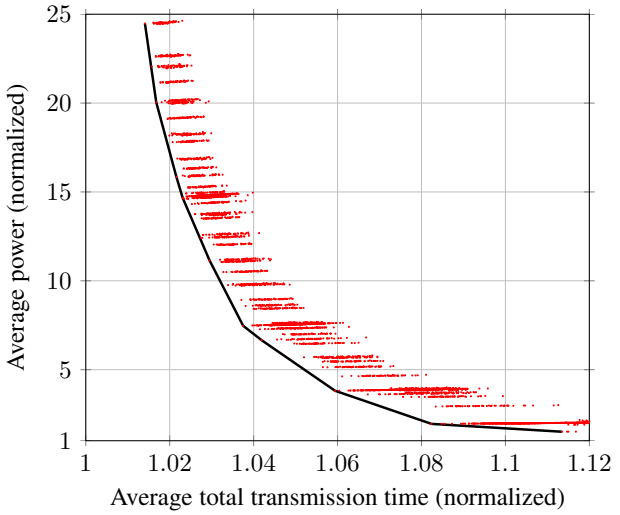



Fig. 3. Trade-off between average transmission time and average power for the case $K=64$, $q=0.5$, $\alpha=1000$, and $S=2$. Red dots are simulation points.

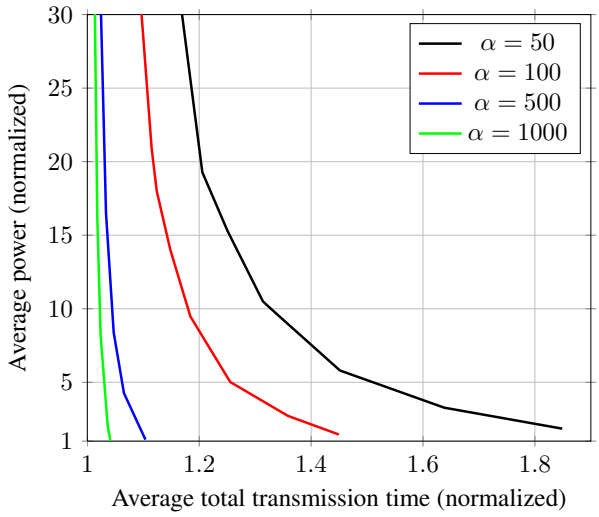



Fig. 4. Trade-off between average total transmission time and average power for the parameters are $N=128$, $q=0.5$ and $S=4$.

The specified protocol leaves the parameters $\epsilon_1$, $\epsilon_2$, $\epsilon_3$, $\epsilon_4$, $L$, and $L_B$ to be specified. Now, we can trace the optimal trade-off between $T_{L,L_B}$ and $P_{L,L_B}$ by solving the optimization problem

$$\min_{\substack{L,L_B,\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4:\\ \prod_{j=1}^4(1-\epsilon_j)\geq 1-\epsilon}} \mathbb{E}[T_{L,L_B}] + \beta\mathbb{E}[P_{L,L_B}]\,. \tag{15}$$

for a range of values of $\beta \geq 0$. The optimization problem is clearly not convex, and hence we find an approximate solution in the next section using a grid search for practical values of $K$, $\epsilon$, $q$, and $P_D$.

## V. Numerical results

In order to solve the optimization problem in (15), we compute $\mathbb{E}[T_{L,L_B}]$ and $\mathbb{E}[P_{L,L_B}]$ using 5000 Monte Carlo simula-

tions of the protocol for $L \in \{1, 2, \cdots, K\}$ and $L_B$ equal to all powers of two between 1 and $K$. We evaluate $\epsilon_1, \cdots, \epsilon_4$ over the four-dimensional grid $10 \times 10 \times 10 \times 10$ grid. The average total transmission time $\mathbb{E}[T_{L,L_B}]$ and average power $\mathbb{E}[P_{L,L_B}]$ are normalized according to the lower bounds in (11) and (12), respectively. The normalization implies that any simulation point must be in the square $[1, \infty) \times [1, \infty)$. The trade-off between average total transmission and average power is computed as the lower convex envelope of the simulation points. This is depicted in Fig. 3, where the simulations points are shown as red dots and the lower convex envelope is the black curve. For the computation, we use $\mathrm{ptr} = 16$ bits. Although the lower convex envelope is not directly achievable using our protocol, it can be achieved by time sharing between two sets of protocol parameters. Note that the lower-most point of the trade-off curve corresponds to the conventional extreme case where the messages of each user are encoded separately. The gap to 1 is thus due overhead from control information.

Our results are depicted in Fig. 4 for the parameters $K = 128$, $q = 0.5$, $S = 4$, and $\alpha \in \{50, 100, 500, 1000\}$. We observe that one can reduce the average total transmission time by grouping users as proposed. Smaller values of $\alpha$ implies that messages are encoded less efficient, and hence grouping becomes an interesting option.

## VI. Conclusions

In this paper, we have addressed the problem of downlink transmission of short packets to $K$ users. Our main objective has been to highlight some of the challenges faced when the messages are small. Specifically, we used recent finite blocklength approximations to visualize the trade-offs between the average power of the each user and the average total transmission time seen from the transmitter. To show this trade-off, we have designed a practical protocol that groups messages and thereby achieves more efficient coding rates. The key element in the protocol design is the encoding of control information.

## References


[1] P. Popovski, "Ultra-reliable communication in 5g wireless systems," in *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014, pp. 146–151.

[2] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless: The art of sending short packets," pp. 1–12, 2015.

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

[4] G. Durisi, J. Östman, Y. Polyanskiy, I. Tal, and W. Yang. (2014, Dec.) SPECTRE: short-packet communication toolbox, v0.2. [Online]. Available: https://github.com/yp-mit/spectre

[5] V. Y. F. Tan and P. Moulin, "Fixed error asymptotics for erasure and list decoding," *arXiv*, pp. 1–18, Feb. 2013. [Online]. Available: http://arxiv.org/abs/1301.7464v2

# The Dispersion of Nearest-Neighbor Decoding for Additive Non-Gaussian Channels

Jonathan Scarlett
Laboratory for Information and Inference Systems,
École Polytechnique Fédérale de Lausanne,
Email: jmscarlett@gmail.com

Vincent Y. F. Tan
Department of ECE,
National University of Singapore
Email: vtan@nus.edu.sg

Giuseppe Durisi
Department of Signals and Systems,
Chalmers University of Technology
Email: durisi@chalmers.se

***Abstract*—We study the second-order asymptotics of information transmission using random Gaussian codebooks and nearest neighbor (NN) decoding over a power-limited additive stationary memoryless non-Gaussian channel. We show that the dispersion term depends on the non-Gaussian noise only through its second and fourth moments. We also characterize the second-order performance of point-to-point codes over Gaussian interference networks. Specifically, we assume that each user's codebook is Gaussian and that NN decoding is employed, i.e., that interference from unintended users is treated as noise at each decoder.**

## I. System Model

Consider the point-to-point additive-noise channel

$$Y^n = X^n + Z^n, \tag{1}$$

where $X^n$ is the input and $Z^n$ is the noise over $n$ scalar channel uses. Throughout, we shall focus exclusively on Gaussian codebooks. More precisely, we consider *shell codes* for which $X^n$ is uniformly distributed on the sphere

$$X^n \sim \mathrm{f}_{X^n}^{(\mathrm{shell})}(\mathbf{x}) := \delta(\|\mathbf{x}\|^2 - nP)/S_n(\sqrt{nP}). \tag{2}$$

Here, $\delta(\cdot)$ is the Dirac delta and $S_n(r) = 2\pi^{n/2}r^{n-1}/\Gamma(n/2)$ is the surface area of a radius-$r$ sphere in $\mathbb{R}^n$. The noise $Z^n$ is assumed to be a stationary and memoryless process that does not depend on the channel input: $Z^n \sim \mathrm{P}_{Z^n}(\mathbf{z}) = \prod_{i=1}^n \mathrm{P}_Z(z_i)$. The distribution $\mathrm{P}_Z$ is non-Gaussian; the only assumptions are:

$$\mathbb{E}[Z^2] = 1, \qquad \xi := \mathbb{E}[Z^4] < \infty, \qquad \mathbb{E}[Z^6] < \infty. \tag{3}$$

Given a shell code consisting of $M \in \mathbb{N}$ random codewords $\mathcal{C} := \{X^n(1), \dots, X^n(M)\}$, we consider an nearest neighbor decoder that returns the message $\hat{W}$ whose corresponding codeword is closest in Euclidean distance to $Y^n$, i.e.,

$$\hat{W} := \operatorname*{arg\,min}_{w \in [1:M]} \|Y^n - X^n(w)\|. \tag{4}$$

This decoder is optimal if the noise is Gaussian, but may not be so in the more general setup considered here.

We define the *average probability of error* as $\bar{p}_{\mathrm{e},n} := \Pr[\hat{W} \neq W]$. This probability is averaged over the uniformly distributed message $W$, the random codebook $\mathcal{C}$ and the channel noise $Z^n$. Note that in traditional channel-coding analyses [1], [2], the probability of error is averaged only over $W$ and $Z^n$. Similar to [3], the additional averaging over the codebook $\mathcal{C}$ is required here to establish an ensemble converse for the class of Gaussian codebooks considered in this paper.

Let $M^*_{\mathrm{shell}}(n, \varepsilon, P; \mathrm{P}_Z)$ be the maximum number of messages that can be transmitted using a shell codebook over the channel (1) with average error probability no larger than $\varepsilon \in (0, 1)$, when the noise is distributed according to $\mathrm{P}_Z$ Lapidoth [3] showed that for all $\varepsilon \in (0, 1)$,

$$\lim_{n\to\infty} \frac{1}{n} \log M^*_{\mathrm{shell}}(n, \varepsilon, P; \mathrm{P}_Z) = \mathsf{C}(P). \tag{5}$$

independent of $\mathrm{P}_Z$.

In Theorem 1 below, we provide the second-order term in the asymptotic expansion of $\log M^*_{\mathrm{shell}}(n, \varepsilon, P; \mathrm{P}_Z)$.

**Theorem 1.** *Consider a noise distribution with statistics as in* (3)*. For shell codes,*

$$\log M^*_{\mathrm{shell}}(n, \varepsilon, P; \mathrm{P}_Z) = n\mathsf{C}(P) - \sqrt{n\mathsf{V}_{\mathrm{shell}}(P, \xi)}\mathrm{Q}^{-1}(\varepsilon) + O(\log n), \tag{6}$$

*where the shell dispersion is*

$$\mathsf{V}_{\mathrm{shell}}(P, \xi) := \big(P^2(\xi - 1) + 4P\big)/\big(4(P + 1)^2\big). \tag{7}$$

The proof together with an extension to Gaussian interference networks can be found in [4]. One of the main tools in our second-order analysis is the *Berry-Esseen theorem for functions of random vectors* (see, e.g., [5, Prop. 1]). The second-order term in the asymptotic expansions of $\log M^*_{\mathrm{shell}}(n, \varepsilon, P; \mathrm{P}_Z)$ depends on the distribution $\mathrm{P}_Z$ only through its second and fourth moments. If $Z$ is standard Gaussian, then the fourth moment $\xi = 3$ and we recover from (7) the Gaussian dispersion [2, Eq. (293)]. Comparing (7) with [2, Eq. (293)] we see that noise distributions $\mathrm{P}_Z$ with higher fourth moments than Gaussian (e.g., Laplace) result in a slower convergence to $\mathsf{C}(P)$.

## References

[1] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. on Inf. Th.*, 55(11):4947–4966, 2009.
[2] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. on Inf. Th.*, 56(5):2307–2359, 2010.
[3] A. Lapidoth. Nearest neighbor decoding for additive non-Gaussian noise channels. *IEEE Trans. on Inf. Th.*, 42(5):1520–1529, 1996.
[4] J. Scarlett, V. Y. F. Tan, and G. Durisi. The dispersion of nearest-neighbor decoding for additive non-Gaussian channels. `arXiv:1512.06618 [cs.IT]`, Dec 2015.
[5] N. Iri and O. Kosut. Third-order coding rate for universal compression of Markov sources. In *Proc. of ISIT*, pages 1996–2000, 2015.

# Resource-Aware Incremental Redundancy in Feedback and Broadcast

Richard D. Wesel, Kasra Vakilinia, Sudarsan V. S. Ranganathan, Tong Mu
University of California, Los Angeles, CA 90095
Email: {wesel,vakiliniak, sudarsanvsr tongmu}@ucla.edu

Dariush Divsalar
Jet Propulsion Laboratory
Cal. Inst. of Tech., Pasadena, CA 91109
Email: Dariush.Divsalar@jpl.nasa.gov

***Abstract*—This paper reviews recent results from the UCLA Communication Systems Laboratory on the use of incremental redundancy. For channels with ACK/NACK feedback, this paper reviews how the transmission lengths used for communicating incremental redundancy should be optimized under the constraint of a limited number of incremental redundancy transmissions. For broadcast channels, this paper reviews optimization of the trade-off between packet-level erasure coding and physical-layer channel coding in the context of block fading with diversity that grows with bocklength.**

## I. Introduction

This invited talk reviews two results [1], [2] optimizing the use of incremental redundancy. In systems with feedback, incremental redundancy adapts the coding rate to the the accumulated information density (the "instantaneous capacity") of the channel allowing the Shannon limit to be approached at much shorter average blocklengths than those required for the accumulated information density to concentrate around the Shannon capacity [3], [4], [5], [6], [7], [8]. In systems without feedback, incremental redundancy can provide a "fountain" of information from which a receiver need only "drink" what is needed to reliably identify the desired message [9], [10], [11]. For each of these two scenarios, this paper reviews optimization techniques that improve performance.

## II. Transmission Lengths for ACK Feedback

ACK/NACK feedback is non-active in the sense that the feedback does not change what is transmitted but rather only indicates whether additional transmissions are needed. For channels with ACK/NACK feedback, the sequential differential optimization (SDO) approach of [1] optimizes the transmission lengths used to communicate incremental redundancy. This optimization maximizes throughput under the constraint of a limited number of incremental redundancy transmissions.

### A. The Normal Approximation

SDO utilizes the power of the normal approximation introduced in [3] that characterizes the behavior of the rate that a channel can support at finite blocklength. Following [3], define information density $i(X,Y)$ as

$$i(X,Y) = \log_2 \frac{f_{Y|X}(y|x)}{f_Y(y)}. \tag{1}$$

The expected value of $i(X,Y)$ is the capacity of the channel. For the example of a BI-AWGN channel with noise $z_k$, $i(X,Y) = 1 - \log_2(1 + e^{-2(z_k+1)/\sigma^2}) = i(z_k)$. The accumulated information density $I_n$ at the receiver after $n$ symbols is

$$I_n = \sum_{k=1}^{n} i(z_k). \tag{2}$$

As pointed out by [3], (2) is a sum of independent random variables that will converge quickly to a normal distribution according to the central limit theorem, leading to the normal approximation of [3].

A key result of [1] is that a normal approximation also accurately describes the rate at which actual variable-length codes with incremental redundancy will successfully decode. Fig. 1 shows that for the NB-LDPC code used in [1] the empirical complementary cumulative distribution function on the rate at which decoding is successful is very closely approximated by a normal distribution for this example of the BI-AWGN channel with SNR of 2 dB. We have similarly confirmed the accuracy of the normal approximation to predict the rate at which decoding is successful for NB-LDPC codes in higher-SNR AWGN channels that require larger constellations and in fading channels with channel state information known at the receiver.

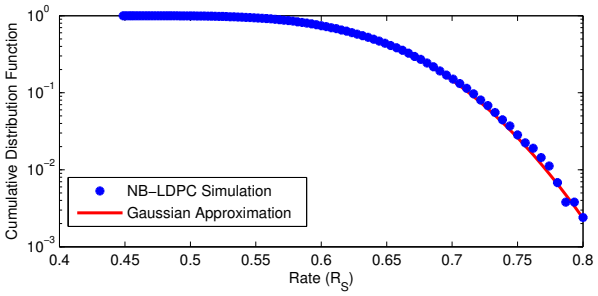



Fig. 1. Empirical complementary cumulative distribution function (c.d.f.) and the Gaussian approximation (Q-function) corresponding to the rate $R_S$ at which the NB-LDPC code of [1] was able to decode successfully.

This material is based upon work supported by the National Science Foundation under Grant Numbers 1162501 and 1161822. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA JPL Task Plan 82-17473.

Fig. 1 shows that $R_S$ is well-approximated by a Gaussian with mean $\mu_S = \mathrm{E}(R_S)$ and variance $\sigma_S^2 = \mathrm{Var}(R_S)$:

$$f_{R_S}(r) = \frac{1}{\sqrt{2\pi\sigma_S^2}} e^{-\frac{(r-\mu_S)^2}{2\sigma_S^2}}. \quad (3)$$

The c.d.f. of the blocklength $N_S$ at which decoding is successful is $F_{N_S}(n) = P(N_S \leq n) = 1 - F_{R_S}(k/n)$. Taking the derivative of $F_{N_S}$ using the Gaussian approximation of $F_{R_S}$ produces the following "reciprocal-Gaussian" approximation for p.d.f. of $N_S$:

$$f_{N_S}(n) = \frac{k}{n^2\sqrt{2\pi\sigma_S^2}} e^{-\frac{\left(\frac{k}{n}-\mu_S\right)^2}{2\sigma_S^2}}. \quad (4)$$

### *B. Sequential Differential Optimization*

SDO uses the tight Gaussian approximation discussed above to optimize the sequence of blocklengths $\{N_1, N_2, \ldots, N_m\}$ to maximize the throughput. Suppose that the number of incremental transmissions is limited to $m$. An accumulation cycle (AC) is a set of $m$ or fewer transmissions and decoding attempts ending when decoding is successful or when the $m^{th}$ decoding attempt fails. If decoding is not successful after the $m^{th}$ decoding attempt, the accumulated transmissions are forgotten and the process starts over with a new transmission of the first block of $N_1$ symbols. From a strict optimality perspective, neglecting the symbols from the previous failed AC is sub-optimal. However, the probability of an AC failure is sufficiently small that the performance degradation is negligible. Neglecting these symbols greatly simplifies analysis.

The cumulative blocklength $N_j$ at the $j$th stage is simply the sum of the first $j$ increment lengths. Using the p.d.f. of $N_S$ from (4) we can compute the probability that the decoder will need a particular incremental transmission. For $N_j < N_{j+1}$, the probability of a successful decoding attempt at blocklength $N_{j+1}$ but not at $N_j$ is

$$\int_{N_j}^{N_{j+1}} f_{N_S}(n)dn = \int_{N_j}^{N_{j+1}} \frac{k}{n^2\sqrt{2\pi\sigma_S^2}} e^{-\frac{\left(\frac{k}{n}-\mu_S\right)^2}{2\sigma_S^2}} dn \quad (5)$$

$$= Q\left(\frac{r_{j+1}-\mu_S}{\sigma_S}\right) - Q\left(\frac{r_j-\mu_S}{\sigma_S}\right), \quad (6)$$

where $r_j = k/N_j$.

Define the throughput as $R_T = \frac{E[K]}{E[N]}$, where $E[N]$ represents the expected number of channel uses and $E[K]$ is the effective number of information bits transferred correctly over the channel. The expression for $E[N]$ is

$$E[N] = N_1 Q\left(\frac{\frac{k}{N_1}-\mu_S}{\sigma_S}\right) \quad (7)$$

$$+ \sum_{j=2}^{m} N_j \left[Q\left(\frac{\frac{k}{N_j}-\mu_S}{\sigma_S}\right) - Q\left(\frac{\frac{k}{N_{j-1}}-\mu_S}{\sigma_S}\right)\right] \quad (8)$$

$$+ N_m \left[1 - Q\left(\frac{\frac{k}{N_m}-\mu_S}{\sigma_S}\right)\right]. \quad (9)$$

The first term (7) shows the contribution to the expected blocklength from successful decoding on the first attempt. $Q\left(\frac{\frac{k}{N_1}-\mu_S}{\sigma_S}\right)$ is the probability of decoding successfully with the initial block of $N_1$. Similarly, the terms in the summation of (8) are the contributions to the expected blocklength from decoding that is first successful at total blocklength $N_j$ for $j \geq 2$ (at the $j^{\text{th}}$ decoding attempt). Finally, the contribution to expected blocklength from not being able to decode even at $N_m$ is (9). Even when the decoding has not been successful at $N_m$, the channel has been used for $N_m$ channel symbols. The expected number of successfully transferred information bits $E[K]$ is

$$E[K] = kQ\left(\frac{\frac{k}{N_m}-\mu_S}{\sigma_S}\right), \quad (10)$$

where $Q\left(\frac{\frac{k}{N_m}-\mu_S}{\sigma_S}\right)$ is the probability of successful decoding. Note that $E[K]$ depends only on $k$ and $N_m$. In fact, $E[K] \approx k$ and is not sensitive to the specific choice of $N_m$ for reasonably large values of $N_m$.

The initial blocklength is $N_1$ and we seek the optimal blocklengths $\{N_1, N_2, \ldots, N_m\}$ to maximize the throughput. Over a range of possible $N_1$ values, the SDO technique introduced in [1] selects $\{N_2, \ldots, N_m\}$ to minimize $E[N]$ for each fixed value of $N_1$ by setting derivatives to zero as follows:

$$\frac{\partial E[N]}{\partial N_j} = 0, \ \forall j = 1, \ldots, m-1. \quad (11)$$

For each $j \in \{2, \ldots, m\}$, the optimal value of $N_j$ is found by setting $\frac{\partial E[N]}{\partial N_{j-1}} = 0$, yielding a sequence of relatively simple computations. In other words, we select the $N_j$ that makes our previous choice of $N_{j-1}$ optimal in retrospect.

For $j > 2$, $\frac{\partial E[N]}{\partial N_{j-1}} = 0$ depends only on $\{N_{j-2}, N_{j-1}, N_j\}$ as follows:

$$\frac{\partial E[N]}{\partial N_{j-1}} = Q\left(\frac{\frac{k}{N_{j-1}}-\mu}{\sigma}\right) + (N_{j-1}-N_j)Q'\left(\frac{\frac{k}{N_{j-1}}-\mu}{\sigma}\right) - Q\left(\frac{\frac{k}{N_{j-2}}-\mu}{\sigma}\right).$$

Thus we can solve for $N_j$ as

$$N_j = \frac{Q\left(\frac{\frac{k}{N_{j-1}}-\mu}{\sigma}\right) + N_{j-1}Q'\left(\frac{\frac{k}{N_{j-1}}-\mu}{\sigma}\right) - Q\left(\frac{\frac{k}{N_{j-2}}-\mu}{\sigma}\right)}{Q'\left(\frac{\frac{k}{N_{j-1}}-\mu}{\sigma}\right)}. \quad (12)$$

For each possible value of $N_1$, SDO can be used to produce an infinite sequence of $N_j$ values that solve (11) for any choice of $m$. The sequence does not depend on $m$, only $N_1$. Each such sequence is an optimal sequence of increment lengths for a given density of decoding attempts on the time axis. As $N_1$ increases, the density of decoding attempts decreases, lowering system complexity. Using SDO to compute the optimal $m$ decoding points is equivalent to selecting the most dense SDO-optimal sequence that when truncated to $m$ points still meets the frame-error-rate target.

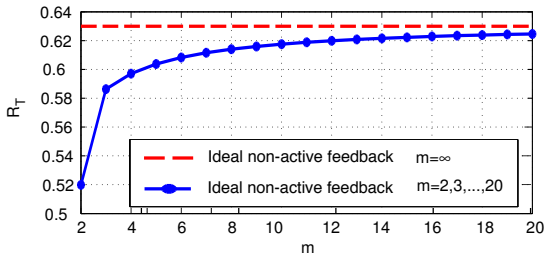



Fig. 2. Throughput as a function of the number $m$ of incremental transmissions permitted.

### C. Approaching Capacity at Short Blocklengths with Feedback

Fig. 2 shows the resulting throughputs obtained by using SDO to find the optimal increment lengths for values of $m$ in the range of $2 < m < 20$ for the target FER of $10^{-3}$ for the NB-LDPC code of [1] for $k = 96$ message bits. Fig. 2 illustrates that with $m = 10$ decoding points, a system can closely approach the performance of a system that has $m = \infty$, which is the limiting case where decoding is attempted and feedback of an ACK/NACK is required after every received symbol.

Using SDO, variable-length codes with average blocklengths of around 500 symbols can closely approach capacity in theory and in practice as demonstrated in [1]. Fig. 3 illustrates the example of a binary-input (BI) additive white Gaussian noise (AWGN) channel with frame error rate (FER) required to be less than $10^{-3}$. For a system transmitting $k$ symbols at an average blocklength of $\lambda$, the throughput $R_t$ is defined by $R_t = k/\lambda$. For reference, Fig. 3 shows the curves of possible throughput $R_t$ as a function of $\lambda$ for some values of $k$. The performance characterization for fixed-blocklength codes is from [3] and is based on the normal approximation, which is shown in [3] to be accurate for blocklengths as small as 100 symbols. The computation of the random coding lower bound on the performance of variable-length codes with feedback is based on the analysis in [4].

Fig. 3 shows curves from [1], [5], [6] that show simulation results that approach or exceed the performance promised by [4] in the range of average blocklengths below 500 bits. For values of $k = 16$, $k = 32$, $k = 64$, and $k = 89$ these throughput results exceed Polyanskiy's random coding lower bound. As the average blocklength becomes larger, the random coding lower bound is more predictive.

Note that variable-length codes with feedback approach capacity at very short blocklengths. In Fig. 3, the random-coding lower bound for a system with feedback is 0.27 dB from the Shannon limit for $k = 280$ with a blocklength of less than 500 bits. Looking at implemented codes for $k = 280$ in Fig. 3, the $m = \infty$ non-binary LDPC (NB-LDPC) code is 0.53 dB from Shannon limit. Using SDO, the NB-LDPC non-active feedback system in Fig. 3 that uses ten rounds of single-bit feedback to operate within 0.65 dB of the Shannon limit with an average blocklength of less than 500 bits.

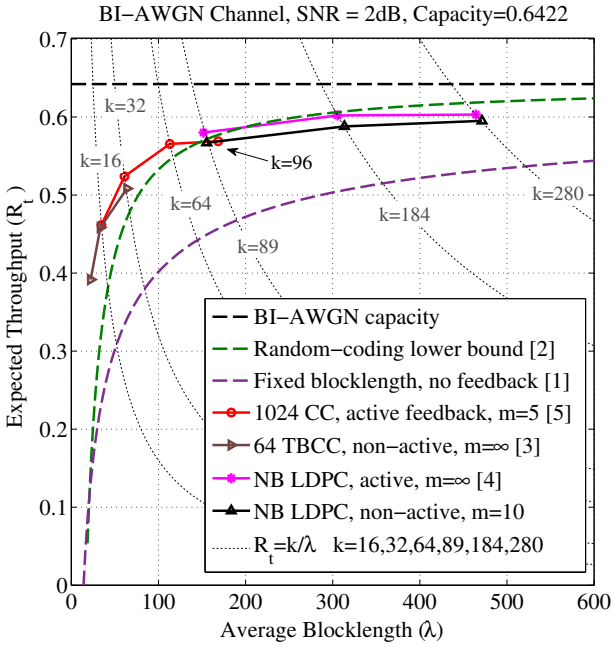



Fig. 3. Approaching capacity at short blocklengths using feedback.

## III. Packet-Level vs. Physical Layer Redundancy

Consider a broadcast setting that uses a hybrid of packet-level erasure coding and physical layer coding to provide a stream of information with the goal of each receiver decoding the desired message at the earliest opportunity. There is a trade-off between using available redundancy for additional packets in a packet-level erasure code or simply for additional physical-layer code symbols.

As the amount of available redundancy grows, the work of [12] shows that in several different block fading scenarios the physical layer coding rate decreases ultimately to zero while the packet-level erasure coding rate does not. This indicates that at some point incremental redundancy should be directed to the physical layer rather than additional packet-level erasure coding. This paper reviews the recent work [2] that studies the this hybrid coding approach using a proportional diversity block fading model (in which diversity increases linearly with blocklength).

### A. The Channel Model and Optimization Problem

Consider a transmitter and a receiver communicating over a fading channel [13]. The one-dimensional channel is modeled as $Y = HX + Z$ where $X$ is the transmitted symbol, $Y$ is the received symbol, $H$ is the fading coefficient, and $Z$ is i.i.d. additive white Gaussian noise (AWGN) with variance $\sigma^2$ and mean 0. We assume the channel is Rayleigh with $\mathbb{E}\left[H^2\right] = 1$, $Z$ has unit variance, i.e. $\sigma^2 = 1$. Let the average transmit power be $\mathbb{E}\left[X^2\right] = P$. Then, the instantaneous *signal-to-noise ratio (SNR)* when $H = h$ is $h^2P$. For this Rayleigh fading channel, SNR (denoted $\gamma$) is exponentially distributed

with parameter $\frac{1}{P}$ that depends only on the average transmit power. Note that, $\gamma$ has a mean of $P$.

A message consisting of $m$ packets with $k$ nats of information per packet is to be transmitted with a low *probability of message error* $q$; this is the probability that the receiver fails to recover all the $m$ packets. The transmitter uses the channel for $T$ units of time for an overall code rate of $\frac{mk}{T}$. It performs erasure coding across the $m$ packets at a rate $R_E$ and codes each resultant packet at a channel-coding rate $R_C$ such that

$$\frac{mk}{T} = R_E R_C. \tag{13}$$

That is, the $m$ packets are first coded using an erasure code at rate $R_E$ to yield $\frac{m}{R_E}$ packets. Note that, for erasure coding, $R_E$ has to satisfy $R_E \leq 1$. To transmit each packet, the transmitter uses a channel code at rate $R_C$ [nats/channel-use] so that the resultant codeword block-length of each packet is $\frac{k}{R_C}$. For a fixed average transmit power, our objective is to pick the value of $R_C$ (and thus $R_E$) that optimizes an objective function. The unit of channel-coding rate is "nats/channel-use" for convenience. The receiver is assumed to know the fading coefficient $H$ while the transmitter does not.

The proportional diversity (PD) model introduce a parameter $l_f$, which describes the length of a fade. With the block-length being $\frac{k}{R_C}$, the number of block fades $F_P$ in a transmitted codeword of a system with PD block fading of fade lengths $l_f$ is

$$F_P = \left\lceil \frac{k}{R_C l_f} \right\rceil. \tag{14}$$

With PD block fading, long codewords benefit from an inherent increase in diversity. For this work, we assume that each block-fading event is independent, i.e. $H$ assumes i.i.d. values across different block fades.

The receiver sees $\frac{m}{R_E} = R_C T k^{-1}$ packets from the channel. The number of packets that the decoder of the erasure code requires to recover the message, denoted $\hat{m} \geq m$, depends upon the erasure code. For Reed-Solomon erasure codes, $\hat{m} = m$; for fountain codes such as a Raptor code, $\hat{m} > m$ typically. Thus, the probability of message error $q$ can be written using the binomial distribution as

$$q = \sum_{i=0}^{\hat{m}-1} \binom{R_C T k^{-1}}{i} (1-p_e)^i \, p_e^{\left(R_C T k^{-1} - i\right)}. \tag{15}$$

In the above expression, $p_e$ denotes the probability that a packet is not decoded successfully (and declared an erasure) upon reception from the channel; this is called the *probability of packet erasure*. Owing to our assumption that the channel codes in the system operate close to capacity with zero block-error probability when the Shannon capacity exceeds the attempted rate, $p_e$ constitutes only one event: *fading outage* [14].

The binomial sum in (15) can be computed numerically only for small values of $R_C T k^{-1}$. Hence, we approximate the random variable that denotes the number of packets successfully decoded by the channel decoder using the Central Limit Theorem (CLT), and obtain the Gaussian approximation for $q$ [12] as

$$q \approx \Phi\left[\frac{(\hat{m}-1) - R_C T k^{-1}(1-p_e)}{\sqrt{R_C T k^{-1} p_e (1-p_e)}}\right], \tag{16}$$

where $\Phi(x)$ is the value of the c.d.f. of the standard normal random variable at $x \in \mathbb{R}$.

To summarize, the objective is to minimize the message-error probability $q$ in (15) via (16), where $p_e$ is also a function of $R_C$. Writing the minimization problem in terms of $R_C$, $R_E$ can be obtained as $R_E = \frac{mk}{TR_C}$. Hence, the optimization problem is as follows:

$$\begin{aligned} \min_{R_C} \quad & \Phi\left[\frac{(\hat{m}-1) - R_C T k^{-1}(1-p_e)}{\sqrt{R_C T k^{-1} p_e (1-p_e)}}\right], \\ \text{s.t.} \quad & p_e(R_C) = \mathbb{P}\left[\frac{1}{\frac{k}{R_C l_f}} \sum_{i=1}^{\left\lfloor \frac{k}{R_C l_f} \right\rfloor} C(\gamma_i) + \right. \\ & \left. \frac{\frac{k}{R_C l_f} - \left\lfloor \frac{k}{R_C l_f} \right\rfloor}{\frac{k}{R_C l_f}} C(\gamma_{\text{last}}) < (1+\epsilon) R_C \right], \\ & \frac{k\hat{m}}{T} \leq R_C \leq \frac{k}{l_f}, \; R_C T k^{-1} \in \mathbb{N}. \end{aligned} \tag{17}$$

Note that, minimizing $\Phi(\cdot)$ is equivalent to minimizing its argument, and the value of $q$ need not be explicitly computed. We have specified the dependence of $p_e$ on $R_C$ here for clarity.

As noted in [12], [15], and many previous works, the evaluation of $p_e$ for the block-Rayleigh fading channel (or for its PD version) is not a straightforward task. One can use [15] or similar works for the block-Rayleigh fading channel to compute the outage probability $p_e$ with a minuscule error. But, our fading model complicates it further as we have a sum of two random variables that are not identically distributed in the expression for $p_e$ in (17). We first expand and rearrange the terms in $p_e$ for our one-dimensional PD block-Rayleigh fading channel with capacity-achieving codes to obtain

$$p_e = \mathbb{P}\left[\sum_{i=1}^{\left\lfloor \frac{k}{R_C l_f} \right\rfloor} W_i + \left(\frac{k}{R_C l_f} - \left\lfloor \frac{k}{R_C l_f} \right\rfloor\right) W_{\text{last}} < \frac{ck}{l_f}\right], \tag{18}$$

where $c = 2(1+\epsilon)$, $W_i = \log(1+\gamma_i)$, $W_{\text{last}} = \log(1+\gamma_{\text{last}})$.

### *B. Gaussian Approximations of the Optimization Problem*

Based on Gaussian approximations of $p_e$ in (18), as inspired by [12], [2] presents four approximations to the optimization problem (17). For numerical-search based results, [2] uses a very low value of the margin, say $\epsilon = 0.05$, to obtain $c$.

*1) Gaussian Approximation 1 (Approx. 1):* Ignoring the contribution of $W_{\text{last}}$ in (18), we get

$$p_e = \mathbb{P}\left[\sum_{i=1}^{\left\lfloor \frac{k}{R_C l_f} \right\rfloor} W_i < \frac{ck}{l_f}\right]. \tag{19}$$

The above can be approximated using the Gaussian CDF as

$$p_e = \Phi\left[\frac{\frac{ck}{l_f} - \left\lfloor \frac{k}{R_C l_f} \right\rfloor \mu(P)}{\sqrt{\left\lfloor \frac{k}{R_C l_f} \right\rfloor \mathrm{Var}(P)}}\right]. \tag{20}$$

The values of $\mu(P)$ and $\mathrm{Var}(P)$, which denote the mean and variance of $\log(1+\gamma)$ with $\gamma \sim$ Exponential $\left(\frac{1}{P}\right)$, can be computed as stated in [12]. By ignoring the flooring function, we get *Gaussian approximation 1 (Approx. 1)*, which is an adaptation of (19) in [12] to PD block-Rayleigh fading:

$$p_e = \Phi\left[\sqrt{\frac{k}{R_C l_f}}\frac{cR_C - \mu(P)}{\sqrt{\mathrm{Var}(P)}}\right]. \tag{21}$$

*2) Gaussian Approximation 2 (Approx. 2):* For *Approx. 2*, we evaluate (20) directly. The approximation to $p_e$ that is being made here is imprecise in the sense that, (20) evaluates to the same value for a range of $R_C$ values; the reason being the presence of the flooring function.

*3) Gaussian Approximation 3 (Approx. 3):* This approximation is the evaluation of (18) with a constrained search space that limits $R_C$ such that both $\frac{m}{R_E}$ and $\frac{k}{R_C l_f}$ are positive integers.

*4) Gaussian Approximation 4 (Approx. 4):* The Gaussian approximation that we make here considers both the terms in (18), making it the most appropriate. Once we find out $\mu(P)$ and $\mathrm{Var}(P)$, we assume that $\sum_{i=1}^{\left\lfloor \frac{k}{R_C l_f} \right\rfloor} W_i$ is Gaussian and also that $\left(\frac{k}{R_C l_f} - \left\lfloor \frac{k}{R_C l_f} \right\rfloor\right) W_{\text{last}}$ is Gaussian. Thus, their linear sum is another Gaussian random variable denoted $W_G$, which stands for *Gaussian approximation of weighted average mutual information*, with

$$\mathrm{mean}(W_G) = \frac{k}{R_C l_f}\mu(P),$$
$$\mathrm{Var}(W_G) = \mathrm{Var}(P)\left[\left\lfloor \frac{k}{R_C l_f} \right\rfloor + \left(\frac{k}{R_C l_f} - \left\lfloor \frac{k}{R_C l_f} \right\rfloor\right)^2\right]. \tag{22}$$

Thus, $p_e$ for this approximation *(Approx. 4)* is

$$p_e = \Phi\left[\frac{\frac{ck}{l_f} - \mathrm{mean}(W_G)}{\sqrt{\mathrm{Var}(W_G)}}\right]. \tag{23}$$

### C. Results and conclusions

Fig. 4 shows an example of the optimal values of $R_C$ and $R_E$ obtained from Approximations 1 and 4 as the overall code rate $\frac{mk}{T}$ goes to 0. As observed by Courtade and Wesel [12] for the (fixed diversity) block-fading channel, for the PD block-fading model that the optimal channel-coding rate goes to 0. However, where the optimal value of $R_E$ approached a non-zero constant less than 1 for fixed diversity in [12], under PD block-fading it is approaching 1. With sufficient overall redundancy, packet-level erasure coding is unnecessary in a block fading channel with proportional diversity. Note that rate-compatibility in this scenario is challenging because the rate $R_E$ increases for a sufficiently low overall rate.

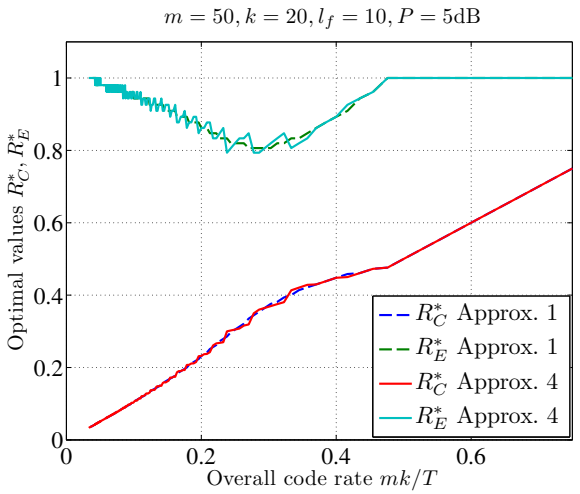



Fig. 4. Result of optimization problem (17) as a function of overall code rate for an example system with $m = 10$, $k = 20$, $l_f = 10$ and $P = 5$ dB.

## References


[1] K. Vakilinia, A. R. Williamson, S. V. S. Raganathan, D. Divsalar, and R. D. Wesel. Feedback systems using non-binary LDPC codes with a limited number of transmissions. In *IEEE Information Theory Workshop*, pages 167 – 171, Hobart, Tasmania, Australia, November 2014.

[2] S. V. S. Ranganathan, T. Mu, and R. D. Wesel. Optimality and rate-compatibility for erasure-coded packet transmissions when fading channel diversity increases with packet length. In arxiv.org/abs/1602.00761, 2016.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[4] Y. Polyanskiy, H. V. Poor, and S. Verdú. Feedback in the non-asymptotic regime. *IEEE Trans. Inf. Theory*, 57(8):4903–4925, Aug. 2011.

[5] A. R. Williamson, T.-Y. Chen, and R. D. Wesel. Variable-length convolutional coding for short blocklengths with decision feedback. *IEEE Trans. Commun.*, 63(7):2389 – 2403, July 2015.

[6] A. R. Williamson, T.-Y. Chen, and R. D. Wesel. Firing the genie: Two-phase short-blocklength convolutional coding with feedback. In *IEEE Inf. Theory and Applicat. Workshop*, pages 1 – 6, San Diego, CA, February 2013.

[7] K. Vakilinia, T.-Y. Chen, S. V. S. Raganathan, A. R. Williamson, D. Divsalar, and R. D. Wesel. Short-blocklength non-binary LDPC codes with feedback-dependent incremental transmissions. In *IEEE Int. Symp. Inf. Theory*, pages 426 – 430, Honolulu, Hawaii, June 2014.

[8] A. R. Williamson, T.-Y. Chen, and R. D. Wesel. Reliability-based error detection for feedback communication with low latency. In *IEEE Int. Symp. Inf. Theory*, pages 2552 – 2556, Istanbul, Turkey, July 2013.

[9] M. G. Luby, M. Mitzenmacher, M. A. Shokrallahi, and D. A. Spielman. Efficient erasure correcting codes. *IEEE Trans. on Info. Th.*, 47(2):569–584, Feb. 2001.

[10] M. Luby. LT codes. In *The 43rd Annual IEEE Symposium on the Foundations of Computer Science*, pages 271–280, Atlanta, Georgia, Nov. 2002.

[11] A. Shokrollahi. Raptor codes. *IEEE Trans. on Info. Th.*, 52(6):2551–2567, June 2006.

[12] T.A. Courtade and R.D. Wesel. Optimal allocation of redundancy between packet-level erasure coding and physical-layer channel coding in fading channels. *IEEE Trans. on Comm.*, 59(8):2101–2109, Aug. 2011.

[13] A. Goldsmith. *Wireless Communications*. Cambridge University Press, New York, 2005.

[14] E. Biglieri. *Coding for Wireless Channels*. Springer, 2005.

[15] A. Yilmaz. Calculating outage probability of block fading channels based on moment generating functions. *IEEE Trans. Commun.*, 59(11):2945–2950, November 2011.

# A Beta-Beta Achievability Bound with Applications

Wei Yang[1], Austin Collins[2], Giuseppe Durisi[3],
Yury Polyanskiy[2], and H. Vincent Poor[1]
[1]Princeton University, Princeton, NJ, 08544 USA
[2]Massachusetts Institute of Technology, Cambridge, MA, 02139 USA
[3]Chalmers University of Technology, 41296 Gothenburg, Sweden

We consider an abstract channel that consists of an input set $\mathcal{A}$, an output set $\mathcal{B}$, and a random transformation $P_{Y|X} : \mathcal{A} \to \mathcal{B}$. An $(M, \epsilon)$ code for the channel $(\mathcal{A}, P_{Y|X}, \mathcal{B})$ comprises a message set $\mathcal{M} \triangleq \{1, \ldots, M\}$, an encoder $f : \mathcal{M} \to \mathcal{A}$ , and a decoder $g : \mathcal{B} \to \mathcal{M} \cup \{e\}$, where $e$ denotes an error event, that satisfies the *average* error probability constraint

$$\frac{1}{M}\sum_{j=1}^{M}\Big(1 - P_{Y\,|\,X}\big(g^{-1}(j)\,|\,f(j)\big)\Big) \le \epsilon. \tag{1}$$

Here, $g^{-1}(j) \triangleq \{y \in \mathcal{Y} : g(y) = j\}$. In Theorem 1 below, we provided a novel lower bound (i.e., achievability bound) on the largest number of codewords for which an $(M, \epsilon)$ code exists.

*Theorem 1 ($\beta\beta$ achievability bound):* For every $0 < \epsilon < 1$ and every input distribution $P_X$, there exists an $(M, \epsilon)$ code for the channel $(\mathcal{A}, P_{Y|X}, \mathcal{B})$ satisfying

$$\frac{M}{2} \ge \sup_{0<\tau<\epsilon}\sup_{Q_Y}\frac{\beta_\tau(P_Y, Q_Y)}{\beta_{1-\epsilon+\tau}(P_{XY}, P_X Q_Y)}. \tag{2}$$

Here, $P_Y \triangleq P_{Y|X} \circ P_X$, and $\beta_\alpha(P, Q)$ is defined as

$$\beta_\alpha(P, Q) \triangleq \min \int P_{Z\,|\,W}(1\,|\,w)Q(dw) \tag{3}$$

where the minimum is over all conditional probability distributions $P_{Z\,|\,W} : \mathcal{W} \to \{0, 1\}$ satisfying

$$\int P_{Z\,|\,W}(1\,|\,w)P(dw) \ge \alpha \tag{4}$$

and $\mathcal{W}$ denotes the support of $P$ and $Q$.

The proof of Theorem 1, which can be found in [1], relies on Shannon's random coding technique and on a suboptimal decoder that is based on the Neyman-Pearson test [2] between $P_{XY}$ and $P_X Q_Y$. Hypothesis testing is used twice in the proof: to relate the decoding error probability to $\beta_{1-\epsilon+\tau}(P_{XY}, P_X Q_Y)$, and to perform a change of measure from $P_Y$ to $Q_Y$.

The bound (2) is the dual of a converse bound recently established by Polyanskiy and Verdú [3, Th. 15]. Furthermore, both (2) and [3, Th. 15] can be viewed as a finite-blocklength analog of the following identity for mutual information (also known as the *golden formula*) [4, Eq. (8.7)], which is exceedingly useful for computing or bounding capacity [5]–[8]:

$$I(X;Y) = D(P_X P_{Y|X} \| P_X Q_Y) - D(P_Y \| Q_Y). \tag{5}$$

The connections between (2) and existing achievability bounds in the literature are discussed in [1].

The bound (2) is useful in situations where $P_Y$ is not a product distribution (although the underlying channel law $P_{Y|X}$ is stationary memoryless), for example due to cost constraints, or structural constraints on the channel input, such as orthogonality or constant composition. In such cases, traditional achievability bounds such as Feinstein's bound [9] and the dependence testing bound [10, Th. 18], which are explicit in $dP_{Y|X}/dP_Y$, become difficult to evaluate. In contrast, the $\beta\beta$ bound (2) requires the evaluation of $dP_{Y|X}/dQ_Y$, which factorizes for product $Q_Y$. This allows for an analytical computation of the bound (2). Furthermore, the term $\beta_\tau(P_Y, Q_Y)$—which captures the cost of the change of measure from $P_Y$ to $Q_Y$—can be evaluated or bounded even when a closed-form expression for $P_Y$ is not available. Applications of the bound (2), which illustrate these properties, are provided in [1].

The work of H. V. Poor and W. Yang was supported in part by the US National Science Foundation under Grants CCF-1420575 and ECCS-1343210. The work of G. Durisi was partly supported by the Swedish Research Council, under grant 3222452.

## References

[1] W. Yang, A. Collins, G. Durisi, Y. Polyanskiy, and H. V. Poor, "A beta-beta achievability bound with applications," *in preparation*.
[2] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Trans. Royal Soc. A*, vol. 231, pp. 289–337, 1933.
[3] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 5–21, Jan. 2014.
[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[5] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.
[6] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.
[7] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.
[8] R. E. Blahut, "Computation of channel capacity and rate-distortion function," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.
[9] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inform. Theory*, vol. 4, no. 4, pp. 2–22, 1954.
[10] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

# Pairwise Secret Key Agreement based on Location-derived Common Randomness

Somayeh Salimi, Panos Papadimitratos
Networked Systems Security Group, School of Electrical Engineering, KTH, Stockholm, Sweden
somayen@kth.se, papadim@.kth.se

***Abstract*—A source model of key sharing between three users is considered in which each pair of them wishes to agree on a secret key hidden from the remaining user. There are rate-limited public channels for communications between the users. We give an inner bound on the secret key capacity region in this framework. Moreover, we investigate a practical setup in which localization information of the users as the correlated observations are exploited to share pairwise keys between the users. The inner and outer bounds of the key capacity region are analyzed in this setup for the case of i.i.d. Gaussian observations.**

## I. Introduction

Secret key sharing at the physical layer is a promising approach for deriving shared secret keys. Ahlswede and Csiszar [1] and Maurer [2] introduced source and channel models of key sharing between two legitimate users in the presence of an eavesdropper using source and channel common randomness along with an unlimited public channel. Various extensions considered a limited public channel [3], sharing of one secret key in a network of users [4], and more than one secret key in different scenarios [5]– [11].

*Pairwise key sharing* first introduced in [11], is a specific problem in this area, requiring that each pair of users shares a secret key concealed from the remaining user(s). In a basic setup including three users with access to correlated source observations and communication over an unlimited public channel, inner and outer bounds on the secret key capacity region were derived. In this paper, we extend the pairwise key sharing framework in [11] to the rate-limited public channel for communications. The public channel is full duplex and each of the users can simultaneously send/receive information over/from the public channel. Based on the correlated observations, users communicate over the rate-limited public channel. Then, each user generates the respective keys as functions of its source observations and the information received over the rate-limited public channel. We derive an inner bound on the key capacity region in this framework; the explicit outer bound given in [11] holds here for the rate-limited public channel case.

We consider location-derived common randomness here because it is a promising, towards practical applications, approach. This is so because a multitude of emerging wireless systems are location-aware and devices can and need to perform distance measurements over RF communication, notably for security reasons, for example [12], [13].

Location-derived common randomness was considered in [14] in a different setup, with a key established between a mobile node and a wireless infrastructure. In a setup closer to the one considered here, [15] considered two users that move according to a discrete time stochastic mobility model and measure their respective distance, after exchanging messages, in the presence of an eavesdropper. In this paper, leveraging the latter approach, we generalize location-derived key sharing to the "pairwise secret key"setting, notably with three users. We present inner bounds of the pairwise key capacity region for both unlimited and limited public channels. Furthermore, the explicit outer bound in [11] is analyzed in this i.i.d. Gaussian setup. Some numerical results are given for the Gaussian setup as well.

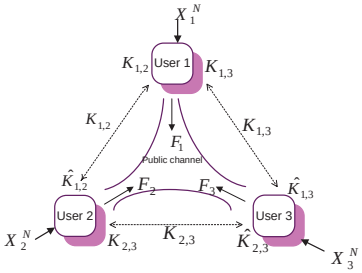



Fig. 1: Pairwise secret key sharing in the source model

The proposed scheme can be extended to the case of more than three users as the future work in which collusion of curious users needs to be investigated. Here we consider simply users curious about the keys their peers derive. But they do not otherwise deviate from the specification and disrupt the protocol.

The rest of the paper is organized as follows: in Section II, the preliminaries of the key sharing setup are given. An inner bound of the pairwise key capacity region with rate-limited public channel is given in Sections III. Deriving pairwise keys from localization information along with the respective inner and outer bounds are presented in Section IV. Numerical results and concluding remarks are given in Sections V and VI, respectively. Proofs of the results are presented in Appendices.

## II. Preliminaries

Users 1, 2 and 3, respectively, have access to $n$ i.i.d. observations $X_1, X_2$ and $X_3$ according to Fig. 1. The observations are correlated according to distribution $P_{X_1X_2X_3}$. The random variable $X_i$ takes values from the finite set $\mathcal{X}_i$ for $i = 1, 2, 3$. Furthermore, there exists a noiseless public channel of limited capacity for communication between the three users where user $i$ is subject to rate constraint $R_i$ for its transmission. Each pair of the three users intends to share a secret key concealed from the remaining user. $K_{i,j}$ denotes the shared key between users $i$ and $j$, hidden from user $m$, for $i, j, m \in \{1, 2, 3\}$, $i < j$, $m \neq i, j$.

We represent the formal definition of the described secret key sharing setup.

User $i$ sends stochastic function $F_i = f_i(X_i^n)$ over the rate-limited public channel for $i = 1, 2, 3$ subject to

$$\frac{1}{n} H(F_i) \leq R_i \tag{1}$$

Upon receiving the information over the public channel, key generation is performed at the users. Key generation function $g_i$ is used by user $i$ for $i = 1, 2, 3$ as:

$$g_1 : \mathcal{F}_2 \times \mathcal{F}_3 \times \mathcal{X}_1^n \to \mathcal{K}_{1,2} \times \mathcal{K}_{1,3} \tag{2}$$
$$g_2 : \mathcal{F}_1 \times \mathcal{F}_3 \times \mathcal{X}_2^n \to \mathcal{K}_{1,2} \times \mathcal{K}_{2,3} \tag{3}$$
$$g_3 : \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{X}_3^n \to \mathcal{K}_{1,3} \times \mathcal{K}_{2,3}. \tag{4}$$

Thus, user 1 calculates $K_{1,2}$ and $K_{1,3}$ to share with users 2 and 3, respectively. Similarly, user 2 calculates $\hat{K}_{1,2}$ and $K_{2,3}$ to share with users 1 and 3 and user 3 calculates $\hat{K}_{1,3}$ and $\hat{K}_{2,3}$ to share with users 1 and 2.

*Definition 1:* In the pairwise secret key sharing over public channels of limited rates $(R_1, R_2, R_3)$ at the respective users 1, 2, 3, the rate triple $(R_{12}, R_{13}, R_{23})$ is an achievable key rate pair if for every $\varepsilon > 0$ and sufficiently large $n$, we have:

$$\forall i < j \in \{1,2,3\} \quad \frac{1}{n} H(K_{i,j}) \geq R_{ij} - \epsilon \tag{5}$$
$$\forall i < j \in \{1,2,3\} \quad \Pr\{K_{i,j} \neq \hat{K}_{i,j}\} < \varepsilon \tag{6}$$
$$\forall i < j, m \in \{1,2,3\}, m \notin \{i,j\} \quad I(K_{i,j}; F_i, F_j, X_m^n) < \varepsilon \tag{7}$$
$$\forall i \in \{1,2,3\} \quad \frac{1}{n} H(F_i) \leq R_i. \tag{8}$$

Equation (5) means that rate $R_{ij}$ is the rate of the secret key between users $i$ and $j$. Equation (6) means that each user can correctly estimate the respective keys. Equation (7) means that each user effectively has no information about the remaining users' secret key. Equation (8) denotes that the key sharing is subject to the constraint of the public channel.

*Definition 2:* The region containing the entire achievable secret key rate triples $(R_{12}, R_{13}, R_{23})$ is the secret key capacity region.

## III. Main Result

In the following, an inner bound on the pairwise key capacity region of the source model with rate-limited public channel is given. First, we define:

$$\begin{aligned}
&\mathbf{r_{12}} = [I(S_{12}; X_2 \,|S_{23}S_{32}) - I(S_{12}; X_3, S_{13} \,|S_{23}, S_{32})]^+,\\
&\mathbf{r_{21}} = [I(S_{21}; X_1 \,|S_{13}S_{31}) - I(S_{21}; X_3, S_{23} \,|S_{13}, S_{31})]^+,\\
&\mathbf{r_{13}} = [I(S_{13}; X_3 \,|S_{23}S_{32}) - I(S_{13}; X_2, S_{12} \,|S_{23}, S_{32})]^+,\\
&\mathbf{r_{31}} = [I(S_{31}; X_1 \,|S_{12}S_{21}) - I(S_{31}; X_2, S_{32} \,|S_{12}, S_{21})]^+,\\
&\mathbf{r_{23}} = [I(S_{23}; X_3 \,|S_{13}S_{31}) - I(S_{23}; X_1, S_{21} \,|S_{13}, S_{31})]^+,\\
&\mathbf{r_{32}} = [I(S_{32}; X_2 \,|S_{12}S_{21}) - I(S_{32}; X_1, S_{31} \,|S_{12}, S_{21})]^+,\\
&\mathbf{I_{12}} = I(S_{12}; S_{21} \,|X_3, S_{13}, S_{23}),\\
&\mathbf{I_{13}} = I(S_{13}; S_{31} \,|X_2, S_{12}, S_{32}),\\
&\mathbf{I_{23}} = I(S_{23}; S_{32} \,|X_1, S_{21}, S_{31}), \mathbf{I_1} = I(S_{21}; S_{31} \,|X_1),\\
&\mathbf{I_2} = I(S_{12}; S_{32} \,|X_2), \mathbf{I_3} = I(S_{13}; S_{23} \,|X_3).
\end{aligned}$$

*Theorem 1:* In the described setup, all rates in the closure of the convex hull of the set of all key rate triples $(R_{12}, R_{13}, R_{23})$ that satisfy the following region, are achievable:

$$\begin{aligned}
&R_{12} > 0, R_{13} > 0, R_{23} > 0,\\
&R_{12} \leq \mathbf{r_{12}} + \mathbf{r_{21}} - \mathbf{I_{12}},\\
&R_{13} \leq \mathbf{r_{13}} + \mathbf{r_{31}} - \mathbf{I_{13}},\\
&R_{23} \leq \mathbf{r_{23}} + \mathbf{r_{32}} - \mathbf{I_{23}},\\
&R_{12} + R_{13} \leq \mathbf{r_{12}} + \mathbf{r_{21}} + \mathbf{r_{13}} + \mathbf{r_{31}} - \mathbf{I_{12}} - \mathbf{I_{13}} - \mathbf{I_1},\\
&R_{12} + R_{23} \leq \mathbf{r_{12}} + \mathbf{r_{21}} + \mathbf{r_{23}} + \mathbf{r_{32}} - \mathbf{I_{12}} - \mathbf{I_{23}} - \mathbf{I_2},\\
&R_{13} + R_{23} \leq \mathbf{r_{13}} + \mathbf{r_{31}} + \mathbf{r_{23}} + \mathbf{r_{32}} - \mathbf{I_{13}} - \mathbf{I_{23}} - \mathbf{I_3},\\
&R_{12} + R_{13} + R_{23} \leq \mathbf{r_{12}} + \mathbf{r_{21}} + \mathbf{r_{13}} + \mathbf{r_{31}} + \mathbf{r_{23}} + \mathbf{r_{32}} -\\
&\qquad \mathbf{I_{12}} - \mathbf{I_{13}} - \mathbf{I_{23}} - \mathbf{I_1} - \mathbf{I_2} - \mathbf{I_3}
\end{aligned} \tag{9}$$

for random variables taking values in sufficiently large finite sets and according to the distribution:

$$\begin{aligned}
&p(s_{12}, s_{13}, s_{21}, s_{23}, s_{31}, s_{32}, x_1, x_2, x_3) = p(x_1, x_2, x_3).\\
&p(s_{12}|x_1)p(s_{13}|x_1)p(s_{21}|x_2)p(s_{23}|x_2)p(s_{31}|x_3)p(s_{32}|x_3)
\end{aligned}$$

and subject to the constraints:

$$I(S_{12};X_1|X_2,S_{32}) + I(S_{13};X_1|X_3,S_{23}) \leq R_1, \tag{10}$$
$$I(S_{21};X_2|X_1,S_{31}) + I(S_{23};X_2|X_3,S_{13}) \leq R_2, \tag{11}$$
$$I(S_{31};X_3|X_1,S_{21}) + I(S_{32};X_3|X_2,S_{12}) \leq R_3, \tag{12}$$
$$\begin{aligned}
&I(S_{12};X_1|X_2,S_{32}) + I(S_{21};X_2|X_1,S_{31}) + I(S_{13},S_{23};X_1,X_2|X_3)\\
&\leq R_1 + R_2,
\end{aligned} \tag{13}$$
$$\begin{aligned}
&I(S_{13};X_1|X_3,S_{23}) + I(S_{31};X_3|X_1,S_{21}) + I(S_{12},S_{32};X_1,X_3|X_2)\\
&\leq R_1 + R_3,
\end{aligned} \tag{14}$$
$$\begin{aligned}
&I(S_{23};X_2|X_3,S_{13}) + I(S_{32};X_3|X_2,S_{12}) + I(S_{21},S_{31};X_2,X_3|X_1)\\
&\leq R_2 + R_3.
\end{aligned} \tag{15}$$
$$\begin{aligned}
&I(S_{21},S_{31};X_2,X_3|X_1) + I(S_{12},S_{32};X_1,X_3|X_2) + I(S_{13},S_{23};X_1,X_2|X_3)\\
&\leq R_1 + R_2 + R_3.
\end{aligned} \tag{16}$$

*Proof:* The proof of Theorem 1 is given in Appendix A in the extended version [18]. ■

The rate region in Theorem 1 is achieved by double random binning as well as Wyner-Ziv coding [16] and rate splitting. In the achievability scheme, the rate of the key between users $i$ and $j$ consists of two parts. A part is rate of the key generated by user $i$ to share with user $j$ ($\mathbf{r}_{ij}$) and the other part is the rate of the key generated by user $j$ to share with user $i$ ($\mathbf{r}_{ji}$). The auxiliary random variable $S_{ij}$ stands for the former key while $S_{ji}$ is associated with the latter key. The total rate of the key between users $i$ and $j$ is the sum of $\mathbf{r}_{ij}$ and $\mathbf{r}_{ji}$ in which term $\mathbf{I}_{ij}$ is subtracted to avoid revealing any information about one of the key to the remaining user (as the eavesdropper) in the case that the other key is disclosed. The limitation of the public channel at the users is reflected in (10)-(16).

*Remark 1:* The region in Theorem 1 reduces to key rate regions in [7] by considering subset of keys and assuming unlimited public channel. It also reduces to the key rate region in [11] by removing public channel limitations.

We do not present a new outer bound on the key capacity region. The explicit outer bound in [11] with unlimited public channel holds in this new setup.

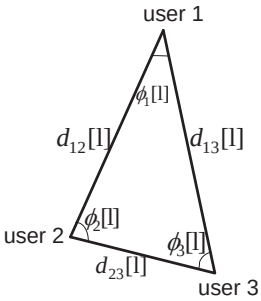



Fig. 2: Using location information for Pairwise secret key sharing

## IV. A Real-World Example of the Pairwise Key Sharing

In this section, we consider pairwise key sharing between three users who move in two-dimensional space according to a discrete time stochastic mobility model. The idea of using localization information to share a secret key between two users in the presence of an eavesdropper was first introduced in [15]. Here, we extend this idea to the pairwise key sharing between three users. The users are mobile in continuous space according to a discrete time stochastic mobility model, independent of each other. Each pair of the three mobile users exploits the distance between themselves as a source of common randomness to share a key while the remaining user tries to make an estimate of that distance as precise as possible. We borrow some notations from [15]. We assume the considered time is divided into $n$ discrete time slots where time slot $l$ includes the time interval $[lT,(l+1)T]$. The users' locations are assumed constant during a time slot. As shown in Fig. 2, at time slot $l$, the distance between users $i$ and $j$ is $d_{ij}[l]=|x_i[l]-x_j[l]|$ in which $x_i[l]\in\mathbb{R}^2$ is the random variable which denotes user $i$ location at time slot $l$. In the same figure, $\phi_i[l]$ shows the angle of the triangle at user $i$ at time slot $l$. Each pair first exchanges beacon signals (e.g., using propagation delay) to make correlated observations and then, they communicate over the (limited) public channel to share a key hidden from the remaining user. This is performed in two phases as follow.

**Localization phase:** User $i$ broadcast some beacons (as a short signal bearing localization information on the initiating node) at the beginning of time slot $l$ and users $j$ and $m$ obtain noisy observations of $d_{ji}[l]$ and $d_{mi}[l]$, respectively, for $i\in\{1,2,3\}, j\neq m\in\{1,2,3\}-i$. We assume the users are equipped to directional antenna and hence, user $i$ obtain $\hat{\phi}_i[l]$ as the noisy version of the angle between the remaining two users. The same as in [15], we assume the sent information by the users is corrupted by Gaussian noises. We have:

$$\tilde{d}_{ij}[l]=d_{ij}[l]+N_{ij}[l] \tag{17}$$

$$\tilde{\phi}_i[l]=\phi_i[l]+N_i[l] \tag{18}$$

where $N_{ij}[l]$ and $N_i[l]$ are zero-mean Gaussian noises with variances $\sigma_{ij}^2$ and $\sigma_i^2$, respectively. All the noises are independent of each other. In the rest of the paper, we consider the case of i.i.d. locations and additive noises. Thus, we drop index $l$ in equations (17)-(18). If the number of broadcast beacons by each user is $J\geq 1$, then $\sigma_{ij}^2$ and $\sigma_i^2$ are divided by $J$ [15]. We assume that users are perfectly clock synchronized (it is shown in [15] that clock mismatch does not affect the theoretical bounds of secret key rates).

**Key generation by public channel communications:** At the beginning of this phase, user $i$ has access to its observations

$$\mathbf{o}_i=\{\tilde{\mathbf{d}}_{ij}=\{\tilde{d}_{ij}[l]\}_{l=1}^n,\tilde{\mathbf{d}}_{im}=\{\tilde{d}_{im}[l]\}_{l=1}^n,\tilde{\boldsymbol{\phi}}_i=\{\tilde{\phi}_i[l]\}_{l=1}^n\} \tag{19}$$

The users communicate over a (rate-limited) public channel to share secret keys in the pairwise manner. Users $i$ and $j$ exploit the reciprocity of the distance between themselves to share a key based on their noisy observations $\tilde{\mathbf{d}}_{ij}$ and $\tilde{\mathbf{d}}_{ji}$, respectively:

$$\tilde{d}_{ij}=d_{ij}+N_{ij} \tag{20}$$

$$\tilde{d}_{ji}=d_{ji}+N_{ji}, \tag{21}$$

where $d_{ij}=d_{ji}$ is the real distance and $N_{ij}\sim\mathcal{N}(0,\sigma_{ij}^2/J)$, $N_{ji}\sim\mathcal{N}(0,\sigma_{ji}^2/J)$ assuming each user broadcasted $J$ beacons at the localization phase. On the other hand, the remaining user $m$ tries to estimate $d_{ij}$ to obtain information about the key between users $i$ and $j$ as much as possible with access to $(\tilde{d}_{mi},\tilde{d}_{mj},\tilde{\phi}_m)$.

Due to simplicity, we assume $\sigma_{ij}=\sigma_{ji}$ between each pair $i$ and $j$. In continue, we consider unlimited and rate-limited public channels separately.

### A. unlimited public channel

Since the observation between pair $i$ and $j$ is symmetric (because of $\sigma_{ij}=\sigma_{ji}$) and the public channels at both sides are unlimited, we choose one-way communication between each pair. Without loss of generality, it is assumed that user 1 communicates to user 2, user 2 communicates to user 3 and user 3 communicates to user 1. According to the directions of communications between users, we choose $S_{12}=\tilde{d}_{12},S_{23}=\tilde{d}_{23},S_{31}=\tilde{d}_{31},S_{21}=S_{32}=S_{13}=\phi$ in Theorem 1. Then the rate region in Theorem 1 is reduced to:

$$R_{12}>0,R_{13}>0,R_{23}>0, \tag{22}$$

$$R_{12}\leq I(\tilde{d}_{12};\tilde{d}_{21})-I(\tilde{d}_{12};\tilde{d}_{31},\tilde{d}_{32},\tilde{\phi}_3) \tag{23}$$

$$R_{13}\leq I(\tilde{d}_{31};\tilde{d}_{13})-I(\tilde{d}_{31};\tilde{d}_{21},\tilde{d}_{23},\tilde{\phi}_2) \tag{24}$$

$$R_{23}\leq I(\tilde{d}_{23};\tilde{d}_{32})-I(\tilde{d}_{23};\tilde{d}_{12},\tilde{d}_{13},\tilde{\phi}_1) \tag{25}$$

Each potential eavesdropper combines its available observations to estimate the distance between the other two users to enlarge the subtracted mutual information terms in (23)-(25). Thus, user $m$ as a potential eavesdropper of the key between users $i$ and $j$ makes estimate of $d_{ij}$ as:

$$\hat{d}_{ij}=\sqrt{\tilde{d}_{mi}^2+\tilde{d}_{mj}^2-2\tilde{d}_{mi}\tilde{d}_{mj}\cos(\tilde{\phi}_m)} \tag{26}$$

where the parameters inside the square root are defined as (17) and (18). For $J\gg 1$, $\sigma_{ij}^2/J\ll d_{ij}^2$ and $\sigma_i^2/J\approx 0$, $\forall i\neq j\in\{1,2,3\}$ with high probability and (26) can be approximated as [15]:

$$\hat{d}_{ij}=d_{ij}+\mathcal{N}(0,\frac{\hat{\sigma}_{ij}^2}{J}) \tag{27}$$

Substituting (27) as the estimate of $d_{ij}$ in (23)-(25) results in the following rate region (it can be shown that this is the best that each potential eavesdropper can do):

*Theorem 2:* Using unlimited public channel in the pairwise key sharing from the localization information, all rates in the closure of the convex hull of the set of all key rate triples $(R_{12}, R_{13}, R_{23})$ that satisfy the following region, are achievable:

$$R_{12} > 0, R_{13} > 0, R_{23} > 0, \tag{28}$$

$$R_{12} \leq \frac{1}{2}\mathbb{E}([\log(1 + \frac{d_{12}^4 J^2(\hat{\sigma}_{12}^2 - \sigma_{12}^2)}{(d_{12}^2 J + \hat{\sigma}_{12}^2)(2d_{12}^2 J\sigma_{12}^2 + \sigma_{12}^4)})]^+) \tag{29}$$

$$R_{13} \leq \frac{1}{2}\mathbb{E}([\log(1 + \frac{d_{13}^4 J^2(\hat{\sigma}_{13}^2 - \sigma_{13}^2)}{(d_{13}^2 J + \hat{\sigma}_{13}^2)(2d_{13}^2 J\sigma_{13}^2 + \sigma_{13}^4)})]^+)) \tag{30}$$

$$R_{23} \leq \frac{1}{2}\mathbb{E}([\log(1 + \frac{d_{23}^4 J^2(\hat{\sigma}_{23}^2 - \sigma_{23}^2)}{(d_{23}^2 J + \hat{\sigma}_{23}^2)(2d_{23}^2 J\sigma_{23}^2 + \sigma_{23}^4)})]^+) \tag{31}$$

in which $\mathbb{E}$ is the expectation with respect to $(d_{12}, d_{13}, d_{23})$ and

$$\hat{\sigma}_{ij}^2 \triangleq \sigma_{im}^2 + \sigma_{jm}^2 + \text{Const}_{d_{12},d_{13},d_{23}}(\frac{\sigma_m^2}{4d_{ij}^2} - \frac{\sigma_{im}^2}{4d_{ij}^2 d_{im}^2} - \frac{\sigma_{jm}^2}{4d_{ij}^2 d_{jm}^2}) \tag{32}$$

for $\text{Const}_{d_{12},d_{13},d_{23}} = (d_{12} + d_{13} + d_{23})(d_{12} + d_{13} - d_{23})(d_{13} + d_{23} - d_{12})(d_{12} + d_{23} - d_{13})$.

*Proof:* The proof is given in Appendix B in [18]. ■

In the following, we give an outer bound on the key capacity region in the described setup for unlimited public channel based on the explicit outer bound in [11].

*Corollary 1:* Using unlimited public channel in the pairwise key agreement from localization information, the following is an outer bound on the pairwise key capacity region:

$$R_{12} > 0, R_{13} > 0, R_{23} > 0, R_{12} \leq \frac{1}{2}\log(1 + \frac{\mathbb{E}(\hat{\sigma}_{12}^2)}{\sigma_{12}^2}) \tag{33}$$

$$R_{13} \leq \frac{1}{2}\log(1 + \frac{\mathbb{E}(\hat{\sigma}_{13}^2)}{\sigma_{13}^2}) \tag{34}$$

$$R_{23} \leq \frac{1}{2}\log(1 + \frac{\mathbb{E}(\hat{\sigma}_{23}^2)}{\sigma_{23}^2}) \tag{35}$$

in which $\mathbb{E}$ is expected value with respect to $(d_{12}, d_{13}, d_{23})$ and $\hat{\sigma}_{ij}^2$ is defined as (32).

*Proof:* The proof is given in Appendix C [18]. ■

### B. rate-limited public channel

In this case, the information sent by the users over the public channel should be subject to the respective rate constraints. In particular, a noisy version of the observation at each user can be considered for the key generation. To apply this constraint, we set:

$$S_{ij} = \tilde{d}_{ij} + D_{ij} \tag{36}$$

in Theorem 1 where $D_{ij} \sim \mathcal{N}(0, \sigma_{ij}'^2)$. The noises $D_{ij}$ are independent of each other and of all the observations. In fact $S_{ij}$ is a noisy version of $\tilde{d}_{ij}$ where its related information can be sent by user $i$ through the public channel with rate constraint $R_i$. It should be noted that in the case of rate-limited public channel, we can not assume one-way communication between each pair and we need to consider the general two-way communications to derive the largest rate region. By considering all the auxiliary random variables of Theorem 1 as (36) and applying the rate constraints in (10)-(16) in Theorem 1, we deduce:

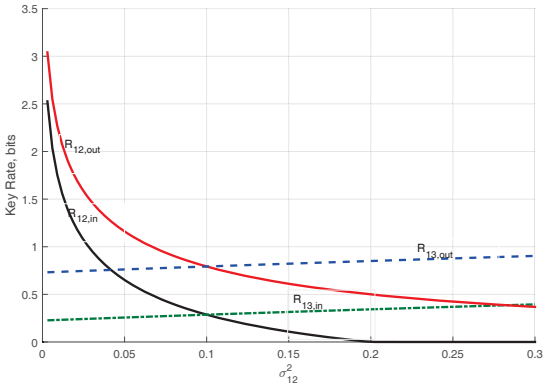



Fig. 3: inner and outer bounds on $R_{12}$ and $R_{13}$

*Theorem 3:* Using public channels with rates $(R_1, R_2, R_3)$, respectively, at users 1,2,3 in the pairwise key sharing from localization information, the pairwise key rate region on the top of the next page is achievable which is subject to the constraints:

$$\frac{1}{2}\mathbb{E}(\log(1 + \frac{(2d_{12}^2 J + \sigma_{12}^2)\sigma_{12}^2}{(d_{12}^2 J + \sigma_{12}^2)\sigma_{12}'^2}) + \log(1 + \frac{(2d_{13}^2 J + \sigma_{13}^2)\sigma_{13}^2}{(d_{13}^2 J + \sigma_{13}^2)\sigma_{13}'^2})) \leq R_1$$

$$\frac{1}{2}\mathbb{E}(\log(1 + \frac{(2d_{12}^2 J + \sigma_{12}^2)\sigma_{12}^2}{(d_{12}^2 J + \sigma_{12}^2)\sigma_{21}'^2}) + \log(1 + \frac{(2d_{23}^2 J + \sigma_{23}^2)\sigma_{23}^2}{(d_{23}^2 J + \sigma_{23}^2)\sigma_{23}'^2})) \leq R_2$$

$$\frac{1}{2}\mathbb{E}(\log(1 + \frac{(2d_{13}^2 J + \sigma_{13}^2)\sigma_{13}^2}{(d_{13}^2 J + \sigma_{13}^2)\sigma_{31}'^2}) + \log(1 + \frac{(2d_{23}^2 J + \sigma_{23}^2)\sigma_{23}^2}{(d_{23}^2 J + \sigma_{23}^2)\sigma_{32}'^2})) \leq R_3 \tag{37}$$

*Proof:* The proof is given in Appendix B in [18]. ■

## V. Numerical Results

In this section, numerical evaluation of the results in Sections IV-A and IV-B is given. We assume that at each time slot, all users' locations are characterized by i.i.d. circularly symmetric zero mean, unit variance Gaussian random variables. First we consider unlimited public channel case. We set $\sigma_{13}^2 = \sigma_{23}^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.1$ and plot the key rates as functions of $\sigma_{12}^2$. Because of symmetry, the bounds on the rates $R_{13}$ and $R_{23}$ are the same and hence, we analyse one of them. In Fig. 3, the inner and outer bounds on key rates $R_{12}$ and $R_{13}$ are shown as functions of $\sigma_{12}^2$. Clearly the bounds on $R_{12}$ decrease as $\sigma_{12}^2$ increases, while the bounds on $R_{13}$ increase with the growth of $\sigma_{12}^2$. However, for small values of $\sigma_{12}^2$, the bounds on $R_{12}$ are more affected compared to the bounds on $R_{13}$.

Then, we analyse the key rate region in the rate-limited public channel case. We set $R_1 = .5$, $R_2 = .2$, $R_3 = .8$ and $\sigma_{12}^2 = \sigma_{13}^2 = \sigma_{23}^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.1$. In order to clarify the rate region, we project the 3-D region into three 2-D regions. As we discussed in Section IV-B, in the case of rate-limited pubic channel, we have two-way communication between each pair. Each user splits its available public channel rate to share keys with the other users while the public channel rates of the other users affect this splitting. As shown in Fig. 4–6, the rate regions are not necessarily rectangular in contrast to the case of unlimited public channel. Obviously, the achievable rates are significantly smaller than the corresponding values in Fig. 3 where unlimited public channel is assumed (respective rates at Fig. 3 for $\sigma_{12}^2 = 0.1$).

$$R_{12} > 0, R_{13} > 0, R_{23} > 0, \tag{38}$$

$$R_{12} \leq \frac{1}{2}\mathbb{E}([\log(1+\frac{d_{12}^4 J^2(\hat{\sigma}_{12}^2-\sigma_{12}^2)}{(d_{12}^2 J+\hat{\sigma}_{12}^2)(d_{12}^2 J(2\sigma_{12}^2+\sigma_{12}'^2)+(\sigma_{12}^2+\sigma_{12}'^2)\sigma_{12}^2)})]^+ + [\log(1+\frac{d_{12}^4 J^2(\hat{\sigma}_{12}^2\sigma_{12}'^2-\sigma_{12}^2(\sigma_{12}^2+\sigma_{12}'^2))}{(d_{12}^2 J(\hat{\sigma}_{12}^2+\sigma_{12}^2+\sigma_{12}'^2)+\hat{\sigma}_{12}^2(\sigma_{12}^2+\sigma_{12}'^2))(d_{12}^2 J(2\sigma_{12}^2+\sigma_{21}'^2)+(\sigma_{12}^2+\sigma_{21}'^2)\sigma_{12}^2)})]^+)$$

$$R_{13} \leq \frac{1}{2}\mathbb{E}([\log(1+\frac{d_{13}^4 J^2(\hat{\sigma}_{13}^2-\sigma_{13}^2)}{(d_{13}^2 J+\hat{\sigma}_{13}^2)(d_{13}^2 J(2\sigma_{13}^2+\sigma_{13}'^2)+(\sigma_{13}^2+\sigma_{13}'^2)\sigma_{13}^2)})]^+ + [\log(1+\frac{d_{13}^4 J^2(\hat{\sigma}_{13}^2\sigma_{13}'^2-\sigma_{13}^2(\sigma_{13}^2+\sigma_{13}'^2))}{(d_{13}^2 J(\hat{\sigma}_{13}^2+\sigma_{13}^2+\sigma_{13}'^2)+\hat{\sigma}_{13}^2(\sigma_{13}^2+\sigma_{13}'^2))(d_{13}^2 J(2\sigma_{13}^2+\sigma_{32}'^2)+(\sigma_{13}^2+\sigma_{31}'^2)\sigma_{13}^2)})]^+)$$

$$R_{23} \leq \frac{1}{2}\mathbb{E}([\log(1+\frac{d_{23}^4 J^2(\hat{\sigma}_{23}^2-\sigma_{23}^2)}{(d_{23}^2 J+\hat{\sigma}_{23}^2)(d_{23}^2 J(2\sigma_{23}^2+\sigma_{23}'^2)+(\sigma_{23}^2+\sigma_{23}'^2)\sigma_{23}^2)})]^+ [\log(1+\frac{d_{23}^4 J^2(\hat{\sigma}_{23}^2\sigma_{23}'^2-\sigma_{23}^2(\sigma_{23}^2+\sigma_{23}'^2))}{(d_{23}^2 J(\hat{\sigma}_{23}^2+\sigma_{23}^2+\sigma_{23}'^2)+\hat{\sigma}_{23}^2(\sigma_{23}^2+\sigma_{23}'^2))(d_{23}^2 J(2\sigma_{23}^2+\sigma_{32}'^2)+(\sigma_{23}^2+\sigma_{32}'^2)\sigma_{23}^2)})]^+) \tag{39}$$

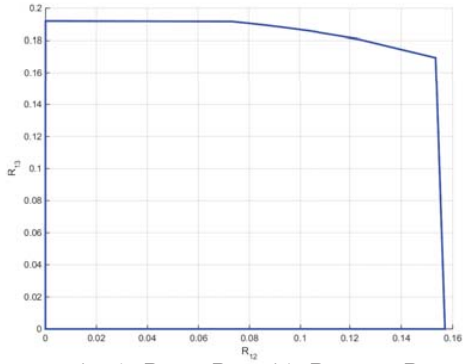



Fig. 4: $R_{12} - R_{13}$ with $R_1 = .5$, $R_2 = .2$,$R_3 = .8$

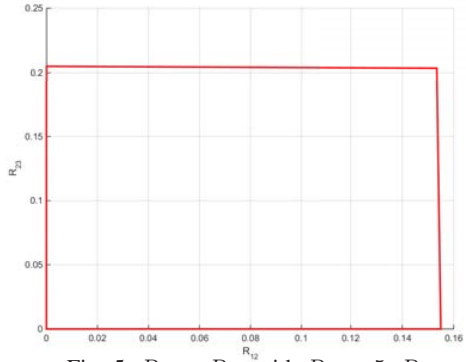



Fig. 5: $R_{12} - R_{23}$ with $R_1 = .5$, $R_2 = .2$,$R_3 = .8$

## VI. Conclusion

The source model of pairwise secret key sharing was investigated with rate-limited pubic channel between three users. An inner bound on the key capacity region was derived for the general case of discrete memoryless source observations. We considered a setup in which the users exploited the distance between themselves as correlated observations to generate keys. Inner and outer bounds on the key capacity region were analyzed for the case of i.i.d. Gaussian observations. As a future work, we analyze the problem of pairwise key sharing between arbitrary number of users who access to limited public channel.

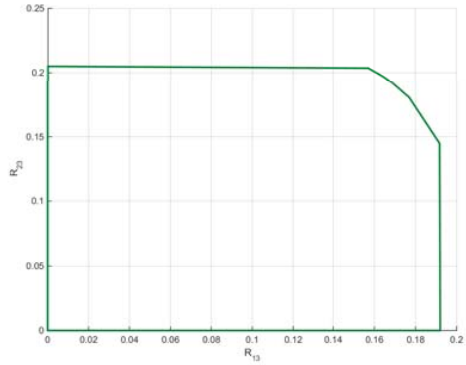



Fig. 6: $R_{13} - R_{23}$ with $R_1 = .5$, $R_2 = .2$,$R_3 = .8$

## References


[1] R. Ahlswede and I. Csiszar, "Common randomness in information theory and cryptography, part I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1121–1132, Jul. 1993.

[2] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 733–742, May 1993.

[3] I. Csiszar, P. Narayan, "Common randomness and secret key generation with a helper," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp.344-366, Mar 2000.

[4] I. Csiszr and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.

[5] C. Ye, P. Narayan, "The secret key-private key capacity region for three terminals," *IEEE Int .Symp. Inf. Theory*, Adelaide, Australia, pp. 2142–2146, Sep. 2005.

[6] S. Nitinawarat, C. Ye, A. Barg, P. Narayan, A. Reznik, "Secret Key Generation for a Pairwise Independent Network Model," *IEEE Int. Symp. Inf. Theory (ISIT)*, Toronto, Canada, pp. 1015–1019, Jul. 2008.

[7] S. Salimi, M. Salmasizadeh, M. R. Aref, "Rate Regions of Secret Key Sharing in a New Source Model," *IET Communications*, Vol. 5, Issue 4, pp. 443–455, March 2011.

[8] S. Salimi, M. Salmasizadeh, M. R. Aref, J. Dj Golić, "Key Agreement over Multiple Access Channel," *IEEE Trans. on Information Forensics and Security*, vol. 6, Issue 3, pp. 775-790, Sep. 2011.

[9] S. Salimi, M. Salmasizadeh, M. R. Aref, "Key Agreement over Multiple Access Channel Using Feedback Channel," *IEEE Int. Symp. Inf. Theory (ISIT)*, Saint Petersburg, Russia, pp. 1936-1940, Aug. 2011.

[10] S. Salimi, M. Skoglund, J. Dj Golić, M. Salmasizadeh, M. R. Aref, "Key Agreement over a Generalized Multiple Access Channel Using Noiseless and Noisy Feedback," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1765–1778, Sep. 2013.

[11] S. Salimi, M. Skoglund, M. Salmasizadeh, M. R. Aref, "Pairwise Secret Key Agreement Using the Source Common Randomness," *Int. Sym. on Wireless Communication Systems (ISWCS)*, pp. 751–755 , Paris, France, Aug. 2012.

[12] P. Papadimitratos, M. Poturalski, P. Schaller, P. Lafourcade, D. Basin, S. Čapkun, J.-P. Hubaux, "Secure Neighborhood Discovery: A Fundamental Element for Mobile Ad-Hoc Networking," *IEEE Communications Magazine*, vol. 46, no. 2, pp. 132–139, Feb. 2008.

[13] M. Fiore, C. Casetti, C.-F. Chiasserini, P. Papadimitratos, "Discovery and Verification of Neighbor Positions in Mobile Ad Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 2, pp. 289 – 303, Feb. 2013.

[14] C. Neuberg, P. Papadimitratos, C. Fragouli, R. Urbanke, "A Mobile World of Security - The Model," *IEEE Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, Mar. 2011.

[15] O. Gungor, F. Chen, C. E. Koksal, "Secret Key Generation From Mobility," *GLOBECOM Workshop on Physical Layer Security*, pp. 874–878 , Texas, US, Dec. 2011.

[16] S. N. Diggavi, V. A. Vaishampayan "On multiple description source coding with decoder side information," *IEEE Information Theory Workshop (ITW)*, San Antonio, Texas, pp. 1-6, Oct. 2004.

[17] G.B. Dantzig, and B.C. Eaves, "Fourier-Motzkin Elimination and its Dual," *Journal of Combinatorial Theory*, Ser. A, 14:288-297, 1973.

[18] S. Salimi, P. Papadimitratos, "Pairwise Secret Key Agreement based on Location-derived Common Randomness,"Extended version, Available at http://www.arxiv.org

# Strong Secrecy for Cooperative Broadcast Channels

Ziv Goldfeld
Ben Gurion University
gziv@post.bgu.ac.il

Gerhard Kramer
Technische Universität München
gerhard.kramer@tum.de

Haim H. Permuter
Ben Gurion University
haimp@bgu.ac.il

Paul Cuff
Princeton University
cuff@princeton.edu

***Abstract*—The broadcast channel (BC) with one confidential message and where the decoders cooperate via a one-sided link is considered. A messages triple with one common and two private messages is transmitted. The private message to the cooperative user is kept secret from the cooperation-aided user. An inner bound on the strong-secrecy-capacity region of the BC is derived. The inner bound is achieved by a channel-resolvability-based Marton code construction that *double-bins* the codebook of the secret message. Both the resolvability and the BC codes use the *likelihood encoder* to choose the transmitted codeword. The protocol uses the cooperation link to convey information on a portion of the non-confidential message and the common message. The inner bound is shown to be tight for the semi-deterministic and physically degraded cases.**

***Index Terms*—Broadcast channel, resolvabiliy, cooperation, physical-layer security, secrecy.**

## I. Introduction

We study broadcast channels (BCs) with one-sided decoder cooperation and one confidential message (Fig. 1). Cooperation is modeled as *conferencing*, i.e., information exchange via a rate-limited link that extends from one receiver (referred to as the *cooperative receiver*) to the other (the *cooperation-aided receiver*). The cooperative receiver possesses confidential information that should be kept secret from the other user.

By extending the coding schemes of Wyner [1] and Csiszár-Körner [2], multiuser settings with secrecy were extensively treated in the literature (e.g., cf. [3], [4] and references therein). These extensions use the so-called *weak-secrecy* metric, i.e., a vanishing information *rate* leakage to the eavesdropper. Observe that although the rate leakage vanishes with the block-length, the eavesdropper can decipher an increasing number of bits from the confidential message. This drawback was highlighted in [5], which instead advocated a secrecy measure referred to as *strong-secrecy*. We consider strong-secrecy by relying on work by Csiszár [6] and Hayashi [7] to relate the coding mechanism for secrecy to *channel-resolvability* rather than channel-capacity (see also [8]).

We first consider a state-dependent channel over which an encoder with non-causal access to the channel state sequence transmits a codeword and aims to make the conditional probability mass function (PMF) of the output given the state resemble a conditional product PMF. The underlying codebook coordinates the transmitted codeword with the state sequence by means of multicoding, i.e., by associating with every message a bin that contains enough codewords to ensure joint encoding (similar to a Gelfand-Pinsker codebook). Most encoders use joint typicality tests to determine the transmitted codeword. We instead adopt the *likelihood encoder* recently proposed in [9].

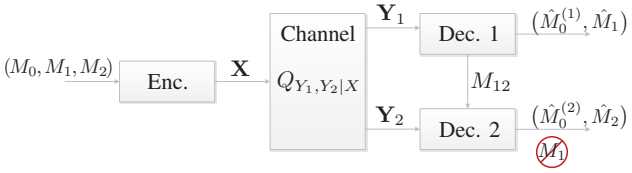



Fig. 1: Cooperative BCs with one confidential message.

Our code ensures that the relation between its codewords correspond to the relation between the channel states and the input in the corresponding resolvability problem. A double-binning of the confidential message codebook allows joint encoding (outer bin layer) and preserves confidentiality (inner bin layer). The sizes of the inner bins are determined by conditions on the rates in our resolvability lemma. To match the conditions of the lemma, we use the likelihood encoder as the multicoding mechanism. Our protocol uses the cooperation link to convey information on a public message that is assembled from portions of the non-confidential message and the common message. The inner bound is shown to be tight for semi-deterministic (SD) and physically-degraded (PD) BCs. As a special case, our results captures the strong-secrecy-capacity region of the SD-BC (without cooperation) where the message to the deterministic user is confidential - an unsolved problem that has merit on its own.

We focus on the cooperative scenario to shed light on the interaction between user cooperation and secure communication. Without secrecy constraints, the public message comprises parts of both private messages [10]. This difference is fundamental when coding for secrecy because a cooperation protocol that shares information about the confidential message violates the secrecy constraint. Since the protocol relies on the cooperative user decoding the public message before sharing it, this difference results in an additional loss in the rate of the confidential message (on top of the loss due to secrecy). The restricted cooperation protocol encapsulates the tension between secrecy and cooperation.

To the best of our knowledge, we present here the first resolvability-based Marton code. This is also a first demonstration of the likelihood encoder's usefulness in the context of secrecy for channel coding problems. From a broader perspective, our resolvability lemma is a tool for upgrading weak-secrecy to strong-secrecy in settings with Marton coding. The reader is referred to [11] for discussion and examples that

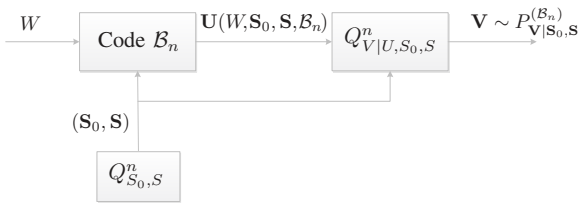



Fig. 2: Coding problem for approximating $P^{(\mathcal{B}_n)}_{\mathbf{V}|\mathbf{S}_0,\mathbf{S}_0} \approx Q^n_{V|S_0,S}$.

are not presented here due to space limitations.

This paper is organized as follows. Sections II and III provide preliminaries and state a central lemma, respectively. In section IV we introduce the cooperative BC and state our inner bound and capacity results. Proofs are given in Section V.

## II. Notations and Preliminaries

We use notation from [11, Section II]. The total variational (TV) distance between two PMFs $P$ and $Q$ is

$$||P-Q|| = \frac{1}{2}\sum_{x\in\mathcal{X}} \big|P(x)-Q(x)\big| \tag{1}$$

and the corresponding relative entropy is

$$D(P||Q) = \sum_{x\in\mathrm{supp}(P)} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \tag{2}$$

**Remark 1** *Pinsker's inequality shows that relative entropy is larger than TV distance. The reverse relation is not generally true, but there is a "reverse" Pinsker inequality for long sequences of independently and identically distributed (i.i.d.) random variables. That is, if $P \ll Q$ (i.e., $P$ is absolutely continuous with respect to $Q$), and $Q$ is an i.i.d. discrete distribution of variables, then*[1]

$$D(P||Q) \in \mathcal{O}\left(\left[n+\log\frac{1}{||P-Q||}\right]||P-Q||\right), \tag{3}$$

*as $||P-Q||$ goes to zero and $n$ goes to infinity (see [12, Equation (29)]). In particular, (3) implies that an exponential decay of the TV distance produces an exponential decay of the informational divergence with the same exponent.*

## III. Conditional Relative Entropy Approximation

Consider a state-dependant discrete memoryless channel (DMC) over which an encoder with non-causal access to the i.i.d. state sequence transmits a codeword (Fig. 2). Each channel state is a pair $(S_0,S)$ of random variables drawn according to $Q_{S_0,S}$. The encoder superimposes its codebook on $S_0$ and then uses the *likelihood encoder* with respect to $S$ to choose the channel input sequence. The conditional PMF of the channel output, given the states, should approximate a conditional product distribution in terms of unnormalized relative entropy.

[1] $f(n)\in\mathcal{O}\big(g(n)\big)$ means that $f(n)\le k\cdot g(n)$, for some $k$ independent of $n$ and sufficiently large $n$.

As shown in Section V-B, we construct a channel-resolvability-based Marton code for the cooperative BC in which the relations between the codewords correspond to those between the channel states and its input in the resolvability setup. The Marton code combines superposition coding and binning, hence the different roles the the state sequences $\mathbf{S}_0$ and $\mathbf{S}$ play in the subsequent resolvability codebook. Lemma 1 is then invoked to achieve strong-secrecy.

### A. Problem Definition

The random variable $W$ is uniformly distributed over $\mathcal{W}=[1:2^{n\tilde{R}}]$ and is independent of $(\mathbf{S}_0,\mathbf{S})\sim Q^n_{S_0,S}$. For any fixed $Q_{U|S_0,S}$, consider the following coding scheme.

**Codebook Construction:** For every $\mathbf{s}_0\in\mathcal{S}_0^n$ generate a codebook $\mathcal{B}_n(\mathbf{s_0})$ that comprises $2^{n\tilde{R}}$ bins, each associated with a different message $w\in\mathcal{W}$ and contains $2^{nR'}$ $u$-codewords that are drawn according to $Q^n_{U|S_0=\mathbf{s}_0}\triangleq\prod_{i=1}^n Q_{U|S_0}(\cdot|s_{0,i})$. Let $\mathcal{B}_n=\big\{\mathcal{B}_n(\mathbf{s}_0)\big\}_{\mathbf{s}_0\in\mathcal{S}_0^n}$ denote this collection of codebooks and denote the codewords in the bin associated with $w\in\mathcal{W}$ by $\big\{\mathbf{u}(\mathbf{s}_0,w,i,\mathcal{B}_n)\big\}_{i\in\mathcal{I}}$, where $\mathcal{I}=[1:2^{nR'}]$.

**Encoding and Induced PMF:** The encoding uses the *likelihood encoder* described by conditional PMF

$$f^{(\mathrm{LE})}(i|w,\mathbf{s}_0,\mathbf{s},\mathcal{B}_n)=\frac{Q^n_{S|U,S_0}\big(\mathbf{s}\big|\mathbf{u}(\mathbf{s}_0,w,i,\mathcal{B}_n),\mathbf{s}_0\big)}{\sum\limits_{i'\in\mathcal{I}}Q^n_{S|U,S_0}\big(\mathbf{s}\big|\mathbf{u}(\mathbf{s}_0,w,i',\mathcal{B}_n),\mathbf{s}_0\big)}. \tag{4}$$

Upon observing $(w,\mathbf{s}_0,\mathbf{s})$, an index $i\in\mathcal{I}$ is drawn according to (4). The codeword $\mathbf{u}(\mathbf{s}_0,w,i,\mathcal{B}_n)$ is passed through the DMC $Q^n_{V|U,S_0,S}$. The distribution induced by the resolvability codebook $\mathcal{B}_n$ is

$$\begin{aligned}P^{(\mathcal{B}_n)}(\mathbf{s}_0,\mathbf{s},w,i,\mathbf{u},\mathbf{v})=&Q^n_{S_0,S}(\mathbf{s}_0,\mathbf{s})2^{-n\tilde{R}}f^{(\mathrm{LE})}(i|w,\mathbf{s}_0,\mathbf{s},\mathcal{B}_n)\\&\times\mathbb{1}_{\{\mathbf{u}(\mathbf{s}_0,w,i,\mathcal{B}_n)=\mathbf{u}\}}Q^n_{V|U,S_0,S}(\mathbf{v}|\mathbf{u},\mathbf{s}_0,\mathbf{s}).\end{aligned} \tag{5}$$

Furthermore, we use $\mathbb{B}_n$ to denote a random codebook that adheres to the above construction .

**Lemma 1 (Sufficient Conditions for Approximation)** *For any $Q_{S_0,S}$, $Q_{U|S_0,S}$ and $Q_{V|U,S_0,S}$, if $(\tilde{R},R')\in\mathbb{R}_+^2$ satisfies*

$$R' > I(U;S|S_0) \tag{6a}$$

$$R'+\tilde{R} > I(U;S,V|S_0), \tag{6b}$$

*then*

$$\mathbb{E}_{\mathbb{B}_n} D\Big(P^{(\mathbb{B}_n)}_{\mathbf{V}|\mathbf{S}_0,\mathbf{S}}\Big|\Big|Q^n_{V|S_0,S}\Big|Q^n_{S_0,S}\Big)\xrightarrow[n\to\infty]{}0. \tag{7}$$

The proof of Lemma 1 shows that the TV distance decays exponentially fast with the blocklength $n$. By Remark 1 this implies an exponential decay of the desired relative entropy. See Section V-A for details.

Another useful property is that the chosen $u$-codeword is jointly letter-typical with $(\mathbf{S}_0,\mathbf{S})$ with high probability.

**Lemma 2 (Typical with High Probability)** *If $(\tilde{R},R')\in\mathbb{R}_+^2$ satisfies* (6)*, then for any $w\in\mathcal{W}$ and $\epsilon>0$, we have*

$$\mathbb{E}_{\mathbb{B}_n}\mathbb{P}\Big(\big(\mathbf{S}_0,\mathbf{S},\mathbf{U}(\mathbf{S}_0,w,I,\mathbb{B}_n)\big)\notin\mathcal{T}_\epsilon^{(n)}(Q_{S_0,S,U})\Big|\mathbb{B}_n\Big)\xrightarrow[n\to\infty]{}0,$$

*where $I$ is a random variable that represents the index chosen by the likelihood encoder $f^{(\mathrm{LE})}$.*

The proof of Lemma 2 relies on [9, Property 1]: for any $\epsilon > 0$ and $f : \mathcal{X} \to \mathbb{R}$ bounded by $b > 0$, if $\big|\big|\Pi - \Lambda\big|\big| < \epsilon$ then $\Big|\mathbb{E}_\Pi f(X) - \mathbb{E}_\Lambda f(X)\Big| < \epsilon b$. The proposition follows by taking $f(\mathbf{S}_0, \mathbf{S}, \mathbf{U}) \triangleq \mathbb{1}_{\left\{(\mathbf{S}_0,\mathbf{S},\mathbf{U}) \notin \mathcal{T}_\epsilon^{(n)}(Q_{S_0,S,U})\right\}}$ and using (7).

## IV. Cooperative Broadcast Channels with One Confidential Message

### A. Problem Definition

The discrete memoryless broadcast channel (DMBC) with cooperation and one confidential message is illustrated in Fig. 1. The channel has one sender and two receivers. The sender chooses a triple $(m_0, m_1, m_2)$ of indices uniformly and independently from the set $[1 : 2^{nR_0}] \times [1 : 2^{nR_1}] \times [1 : 2^{nR_2}]$ and maps them to a sequence $\mathbf{x} \in \mathcal{X}^n$, which is the channel input. The sequence $\mathbf{x}$ is transmitted over a BC with transition probability $Q_{Y_1,Y_2|X}$. If $Q_{Y_1,Y_2|X}$ factors as $\mathbb{1}_{\{Y_1 = f(X)\}} Q_{Y_2|X}$ or $Q_{Y_1|X} Q_{Y_2|Y_1}$ then we call the BC SD or PD, respectively. The output sequence $\mathbf{y}_j \in \mathcal{Y}_j^n$, where $j = 1, 2$, is received by decoder $j$. Decoder $j$ produces a pair of estimates $\big(\hat{m}_0^{(j)}, \hat{m}_j\big)$ of $(m_0, m_j)$. Furthermore, the message $m_1$ is to be kept secret from Decoder 2. There is a one-sided noiseless cooperation link of rate $R_{12}$ from Decoder 1 to Decoder 2. By conveying a message $m_{12} \in [1 : 2^{nR_{12}}]$ over this link, Decoder 1 can share with Decoder 2 information about $\mathbf{y}_1$, $\big(\hat{m}_0^{(1)}, \hat{m}_1\big)$, or both.

**Definition 1 (Code)** *An $(n, R_{12}, R_0, R_1, R_2)$ code $\mathcal{C}_n$ for the BC with cooperation and one confidential message has: (i) Four message sets $\mathcal{M}_{12} = [1 : 2^{nR_{12}}]$ and $\mathcal{M}_j = [1 : 2^{nR_j}]$, for $j = 0, 1, 2$; (ii) A stochastic encoder described by a stochastic matrix $f^{(\mathcal{E})}_{\mathbf{X}|M_0,M_1,M_2}$ on $\mathcal{X}^n$; (iii) A decoder cooperation function $f_{12} : \mathcal{Y}_1^n \to \mathcal{M}_{12}$; (iv) Two decoding functions $\phi_1 : \mathcal{Y}_1^n \to \mathcal{M}_0 \times \mathcal{M}_1$ and $\phi_2 : \mathcal{Y}_2^n \times \mathcal{M}_{12} \to \mathcal{M}_0 \times \mathcal{M}_2$.*

**Definition 2 (Error Probability)** *The average error probability for an $(n, R_{12}, R_0, R_1, R_2)$ code $\mathcal{C}_n$ is*

$$P_e(\mathcal{C}_n) = \mathbb{P}_{\mathcal{C}_n}\Big((\hat{M}_0^{(1)}, \hat{M}_0^{(2)}, \hat{M}_1, \hat{M}_2) \neq (M_0, M_0, M_1, M_2)\Big),$$

*where $\mathbb{P}_{\mathcal{C}_n}(\cdot)$ means that the probability is calculated with respect to the joint PMF induced by $\mathcal{C}_n$. Furthermore, $(\hat{M}_0^{(1)}, \hat{M}_1) = \phi_1(\mathbf{Y}_1)$ and $(\hat{M}_0^{(2)}, \hat{M}_2) = \phi_2\big(\mathbf{Y}_2, f_{12}(\mathbf{Y}_1)\big)$.*

The *information leakage* at receiver 2 is measured by $\mathbb{L}(\mathcal{C}_n) = I_{\mathcal{C}_n}(M_1; M_{12}, \mathbf{Y}_2)$, which is also calculated with respect to PMF induced by $\mathcal{C}_n$.

**Definition 3 (Achievability)** *A rate tuple $(R_{12}, R_0, R_1, R_2) \in \mathbb{R}_+^4$ is achievable if for any $\epsilon > 0$ there is an $(n, R_{12}, R_0, R_1, R_2)$ code $\mathcal{C}_n$ with $P_e(\mathcal{C}_n) \leq \epsilon$ and $\mathbb{L}(\mathcal{C}_n) \leq \epsilon$, for any $n$ sufficiently large.*

The *strong-secrecy-capacity region* $\mathcal{C}_\mathrm{S}$ is the closure of the set of the achievable rates.

### B. Strong-Secrecy-Capacity Bounds and Results

We state an inner bound on the strong-secrecy-capacity region $\mathcal{C}_\mathrm{S}$ of a cooperative BC with one confidential message.

**Theorem 3 (Inner Bound)** *Let $\mathcal{R}_I$ be the closure of the union of rate tuples $(R_{12}, R_0, R_1, R_2) \in \mathbb{R}_+^4$ satisfying:*

$$\begin{aligned} R_1 &\leq I(U_1; Y_1|U_0) - I(U_1; U_2, Y_2|U_0) \\ R_0 + R_1 &\leq I(U_0, U_1; Y_1) - I(U_1; U_2, Y_2|U_0) \\ R_0 + R_2 &\leq I(U_0, U_2; Y_2) + R_{12} \\ \sum_{j=0,1,2} R_j &\leq I(U_0, U_1; Y_1) + I(U_2; Y_2|U_0) - I(U_1; U_2, Y_2|U_0), \end{aligned} \tag{8}$$

*where the union is over all PMFs $Q_{U_0,U_1,U_2,X} Q_{Y_1,Y_2|X}$. Then the inclusion $\mathcal{R}_I \subseteq \mathcal{C}_S$ holds.*

The proof of Theorem 3 relies on a channel-resolvability-based Marton code and is given in Section V-B. The inner bound in Theorem 3 is tight for SD and PD BCs.

**Theorem 4 (Secrecy-Capacity for SD-BC)** *The strong-secrecy-capacity region $\mathcal{C}_S^{(SD)}$ of a cooperative SD-BC with one confidential message is the closure of the union of rate tuples $(R_{12}, R_0, R_1, R_2) \in \mathbb{R}_+^4$ satisfying:*

$$\begin{aligned} R_1 &\leq H(Y_1|W, V, Y_2) \\ R_0 + R_1 &\leq H(Y_1|W, V, Y_2) + I(W; Y_1) \\ R_0 + R_2 &\leq I(W, V; Y_2) + R_{12} \\ \sum_{j=0,1,2} R_j &\leq H(Y_1|W, V, Y_2) + I(V; Y_2|W) + I(W; Y_1), \end{aligned} \tag{9}$$

*where the union is over all $Q_{W,V,Y_1,X} Q_{Y_2|X}$ with $Y_1 = f(X)$.*

The direct part of Theorem 4 follows from Theorem 3 by setting $U_0 = W$, $U_1 = Y_1$ and $U_2 = V$.

**Theorem 5 (Secrecy-Capacity for PD-BC)** *The strong-secrecy-capacity region $\mathcal{C}_S^{(PD)}$ of a cooperative PD-BC with on confidential message is the closure of the union of rate tuples $(R_{12}, R_0, R_1, R_2) \in \mathbb{R}_+^4$ satisfying:*

$$\begin{aligned} R_1 &\leq I(X; Y_1|W) - I(X; Y_2|W) \\ R_0 + R_2 &\leq I(W; Y_2) + R_{12} \\ \sum_{j=0,1,2} R_j &\leq I(X; Y_1) - I(X; Y_2|W), \end{aligned} \tag{10}$$

*where the union is over all $Q_{W,X} Q_{Y_1|X} Q_{Y_2|Y_1}$.*

The achievability of $\mathcal{C}_\mathrm{S}^{(\mathrm{PD})}$ follows by setting $U_0 = W$, $U_1 = X$ and $U_2 = 0$ in Theorem 3.

**Remark 2 (Converse)** *The converse proofs for Theorems 4 and 5 are omitted due to space limitations (see [11] for details). We remark that we used two distinct converse proofs. In the converse of Theorem 4, the fourth bound in* (9) *does not involve $R_{12}$ since the auxiliary random variable $W_i$ contains $M_{12}$. With respect to this choice of $W_i$, showing that $W - X - (Y_1, Y_2)$ forms a Markov chain relies heavily on the SD property of the channel. For the PD-BC, however, such*

*an auxiliary is not feasible as it violates the Markov relation $W - X - Y_1 - Y_2$ induced by the channel. To circumvent this, in the converse of Theorem 5 we define $W_i$ without $M_{12}$ and use the structure of the channel to keep $R_{12}$ from appearing in the third rate bound in* (10)*. Specifically, this argument relies on the relation $M_{12} = f_{12}(\mathbf{Y}_1)$ and that $Y_2$ is a degraded version of $Y_1$, implying that all three messages $(M_0, M_1, M_2)$ are reliably decodable from $\mathbf{Y}_1$ only.*

## V. Proofs

### A. Proof of Lemma 1

Note that the factorization of $P^{(\mathcal{B}_n)}$ from (5) implies that $P^{(\mathcal{B}_n)}_{\mathbf{S}_0,\mathbf{S}} = Q^n_{S_0,S}$. Therefore, to establish Lemma 1 we show that

$$\mathbb{E}_{\mathbb{B}_n} D\Big(P^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}}\Big\|Q^n_{S_0,S,V}\Big) \xrightarrow[n\to\infty]{} 0. \tag{11}$$

For every fixed codebook $\mathcal{B}_n$, $P^{(\mathcal{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}}$ is absolutely continuous with respect to $Q^n_{S_0,S,V}$. Combining this with Remark 1, a sufficient condition for (11) is that

$$\mathbb{E}_{\mathbb{B}_n}\Big\|P^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}} - Q^n_{S_0,S,V}\Big\| \xrightarrow[n\to\infty]{} 0. \tag{12}$$

To evaluate the TV distance in (12), define the *ideal* PMF of $(\mathbf{S}_0, \mathbf{S}, W, I, \mathbf{U}, \mathbf{V})$ as

$$\begin{aligned}\Gamma^{(\mathcal{B}_n)}(\mathbf{s}_0, w, i, \mathbf{u}, \mathbf{s}, \mathbf{v}) &= Q^n_{S_0}(\mathbf{s}_0)2^{-n(\tilde{R}+R')}\mathbb{1}_{\{\mathbf{u}(\mathbf{s}_0,w,i,\mathcal{B}_n)=\mathbf{u}\}}\\ &\quad\times Q^n_{S|U,S_0}(\mathbf{s}|\mathbf{u},\mathbf{s}_0)Q^n_{V|U,S_0,S}(\mathbf{v}|\mathbf{u},\mathbf{s}_0,\mathbf{s})\end{aligned}$$

with respect to the same codebook $\mathcal{B}_n$ as $P^{(\mathcal{B}_n)}$. Note, however, that $\Gamma$ describes an encoding process where the choice of the $u$-codeword from a certain bin is uniform, as opposed to $P$ that uses the likelihood encoder. Furthermore, the structure of $\Gamma$ implies that the sequence $\mathbf{s}$ is generated by feeding $\mathbf{s}_0$ and the chosen $u$-codeword into a DMC $Q^n_{S|U,S_0}$.

Using the TV distance triangle inequality, we upper bound the left-hand side of (12) by

$$\mathbb{E}_{\mathbb{B}_n}\Big[\Big\|P^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}} - \Gamma^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}}\Big\| + \Big\|\Gamma^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}} - Q^n_{S_0,S,V}\Big\|\Big] \tag{13}$$

By [12, Corollary VII.5], the second expected TV distance decays exponentially fast as $n \to \infty$ if (6b) holds.

For the first term in (13), we use the following relations between $\Gamma$ and $P$. For every fixed codebook $\mathcal{B}_n$, we have

$$\begin{aligned}\Gamma^{(\mathcal{B}_n)}_{I|W,\mathbf{S}_0,\mathbf{S}} &= f^{(\mathrm{LE})}_{I|W,\mathbf{S}_0,\mathbf{S},\mathbb{B}_n=\mathcal{B}_n} = P^{(\mathcal{B}_n)}_{I|W,\mathbf{S}_0,\mathbf{S}}\\ \Gamma^{(\mathcal{B}_n)}_{\mathbf{U}|I,W,\mathbf{S}_0,\mathbf{S}} &= \mathbb{1}_{\{\mathbf{U}=\mathbf{U}(\mathbf{S}_0,W,I,\mathcal{B}_n)\}} = P^{(\mathcal{B}_n)}_{\mathbf{U}|I,W,\mathbf{S}_0,\mathbf{S}}\\ \Gamma^{(\mathcal{B}_n)}_{\mathbf{V}|\mathbf{U},I,W,\mathbf{S}_0,\mathbf{S}} &= Q^n_{V|U,S_0,S} = P^{(\mathcal{B}_n)}_{\mathbf{V}|\mathbf{U},I,W,\mathbf{S}_0,\mathbf{S}}.\end{aligned}$$

Consequently, basic properties of the TV distance and the symmetry in constructing $\mathcal{B}_n$ give

$$\mathbb{E}_{\mathbb{B}_n}\Big\|P^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}} - \Gamma^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S},\mathbf{V}}\Big\| \le \mathbb{E}_{\mathbb{B}_n}\Big\|Q^n_{S_0,S} - \Gamma^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S}|W=1}\Big\|.$$

Invoking [12, Corollary VII.5] once more, (6a) implies that

$$\mathbb{E}_{\mathbb{B}_n}\Big\|Q^n_{S_0,S} - \Gamma^{(\mathbb{B}_n)}_{\mathbf{S}_0,\mathbf{S}|W=1}\Big\| \xrightarrow[n\to\infty]{} 0 \tag{14}$$

exponentially fast.

### B. Proof of Theorem 3

**Codebook Generation:** Split $M_2$ into two independent parts $(M_{20}, M_{22})$ with rates $R_{20}$ and $R_{22}$ that satisfy $R_2 = R_{20} + R_{22}$, and alphabets $\mathcal{M}_{20}$ and $\mathcal{M}_{22}$, respectively. $M_p \triangleq (M_0, M_{20})$ is referred to as a *public message* while $M_{22}$ denotes *private message number 2*. We also use $\mathcal{M}_p \triangleq \mathcal{M}_0 \times \mathcal{M}_{20}$ and $R_p \triangleq R_0 + R_{20}$. Let $W$ be a random variable uniformly distributed over $\mathcal{W} = [1 : 2^{n\tilde{R}}]$ and independent of $(M_0, M_1, M_2)$.

Generate a public message codebook[2] $\mathcal{C}_0$ that comprises $2^{nR_p}$ $u_0$-codewords $\mathbf{u}_0(m_p, \mathcal{C}_0)$, $m_p \in \mathcal{M}_p$, drawn according to $Q^n_{U_0}$. Randomly and uniformly partition $\mathcal{C}_0$ into $2^{nR_{12}}$ bins $\mathcal{B}(m_{12})$, where $m_{12} \in \mathcal{M}_{12}$.

For each $\mathbf{u}_0(m_p, \mathcal{C}_0)$, $m_p \in \mathcal{M}_p$, generate a codebook $\mathcal{C}_1(m_p)$ that comprises $2^{n(R_1+R'_1+\tilde{R})}$ codewords $\mathbf{u}_1$, each drawn according to $Q^n_{U_1|U_0}\big(\,\cdot\,\big|\mathbf{u}_0(m_p, \mathcal{C}_0)\big)$ independent of all the other $u_1$-codewords. Label these codewords as $\mathbf{u}_1(m_p, m_1, i, w, \mathcal{C}_1)$, where $(m_1, i, w) \in \mathcal{M}_1 \times \mathcal{I} \times \mathcal{W}$ and $\mathcal{I} \triangleq [1 : 2^{nR'_1}]$.

For each $\mathbf{u}_0(m_p, \mathcal{C}_0)$, $m_p \in \mathcal{M}_p$, also generate a codebook $\mathcal{C}_2(m_p)$ that comprises $2^{nR_{22}}$ $u_2$-codewords, each associated with a private message $m_{22} \in \mathcal{M}_{22}$. Each $u_2$-codeword is drawn according to $Q^n_{U_2|U_0}\big(\,\cdot\,\big|\mathbf{u}_0(m_p, \mathcal{C}_0)\big)$ independent of all the other $u_2$-codewords. Denote $\mathcal{C}_2(m_p) \triangleq \big\{\mathbf{u}_2(m_p, m_{22}, \mathcal{C}_2)\big\}_{m_{22}\in\mathcal{M}_{22}}$.

The channel input $\mathbf{x}$ associated with a triple $(\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2)$ is generated according to $Q^n_{X|U_0,U_1,U_2}\big(\,\cdot\,\big|\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2\big)$.

**Encoding:** To transmit a triple $(m_0, m_1, m_2)$, the encoder transforms it into the triple $(m_p, m_1, m_{22})$, and draws $W$ uniformly over $\mathcal{W}$. Then, an index $i \in \mathcal{I}$ is chosen by the likelihood encoder described in (15) at the top of the next page. The corresponding $\mathbf{x}$ is transmitted over the BC.

**Decoding and Cooperation:** Decoder 1: Searches for a unique triple $(\hat{m}_p, \hat{m}_1, \hat{w}) \in \mathcal{M}_p \times \mathcal{M}_1 \times \mathcal{W}$, for which there is an index $\hat{i} \in \mathcal{I}$ such that $\big(\mathbf{u}_0(\hat{m}_p, \mathcal{C}_0), \mathbf{u}_1(\hat{m}_p, \hat{m}_1, \hat{i}, \hat{w}, \mathcal{C}_1), \mathbf{y}_1\big)$ are in $\mathcal{T}^{(n)}_\epsilon(Q_{U_0,U_1,Y_1})$. If such unique triple is found, $(\hat{m}_0, \hat{m}_1)$ is declared as the decoded message pair.

Cooperation: Having $(\hat{m}_{20}, \hat{m}_1, \hat{i}, \hat{w})$, Decoder 1 sends the bin number of $\mathbf{u}_0(\hat{m}_p, \mathcal{C}_0)$ to Decoder 2 via the cooperation link.

Decoder 2: Upon receiving $(\hat{m}_{12}, \mathbf{y}_2)$, Decoder 2 searches for a unique pair $(\hat{\hat{m}}_p, \hat{\hat{m}}_{22}) \in \mathcal{M}_p \times \mathcal{M}_{22}$ such that $\big(\mathbf{u}_0(\hat{\hat{m}}_p, \mathcal{C}_0), \mathbf{u}_2(\hat{\hat{m}}_p, \hat{\hat{m}}_{22}, \mathcal{C}_2), \mathbf{y}_2\big)$ is in $\mathcal{T}^{(n)}_\epsilon(Q_{U_0,U_2,Y_2})$, where $\mathbf{u}_0(\hat{\hat{m}}_{20}) \in \mathcal{B}(\hat{m}_{12})$. If such a unique pair is found, then $(\hat{\hat{m}}_0, \hat{\hat{m}}_2)$ is declared as the decoded message.

The error probability analysis, which we omit due to space limitations, uses Lemma 2 to first show that the above encoding process result in $u_0$-, $u_1$-, $u_2$- and $x$-sequences that are jointly typical. Then, by standard joint-typicality decoding arguments, reliability is established provided that

$$\begin{aligned}R' &> I(U_1; U_2|U_0)\\ R' + \tilde{R} &> I(U_1; U_2, Y_2|U_0)\end{aligned}$$

[2]The subsequent notations for codebooks omit the blocklength $n$.

$$f_{\mathrm{BC}}^{(\mathrm{LE})}\big(i\big|w,\mathbf{u}_0(m_p,\mathcal{C}_0),\mathbf{u}_2(m_p,m_{22},\mathcal{C}_2),\mathcal{C}_1\big)=\frac{Q^n_{U_2|U_1,U_0}\big(\mathbf{u}_2(m_p,m_{22},\mathcal{C}_2)\big|\mathbf{u}_1(m_p,m_1,i,w,\mathcal{C}_1),\mathbf{u}_0(m_p,\mathcal{C}_0)\big)}{\sum_{i'\in\mathcal{I}}Q^n_{U_2|U_1,U_0}\big(\mathbf{u}_2(m_p,m_{22},\mathcal{C}_2)\big|\mathbf{u}_1(m_p,m_1,i',w,\mathcal{C}_1),\mathbf{u}_0(m_p,\mathcal{C}_0)\big)}. \tag{15}$$

$$\begin{aligned}
I(M_1;M_{12},\mathbf{Y}_2|\mathbb{C}) &\overset{(a)}{\leq} \mathbb{E}_{\mathbb{C}}D\Big(P^{(\mathbb{C})}_{\mathbf{Y}_2|M_p,M_1,M_{22},\mathbf{U}_0,\mathbf{U}_2}\Big\|Q^n_{Y_2|U_0,U_2}\Big|P^{(\mathbb{C})}_{M_p,M_1,M_{22},\mathbf{U}_0,\mathbf{U}_2}\Big)\\
&\overset{(b)}{=} \mathbb{E}_{\mathbb{C}}D\Big(P^{(\mathbb{C})}_{\mathbf{Y}_2|M_p=1,M_1=1,M_{22}=1,\mathbf{U}_0,\mathbf{U}_2}\Big\|Q^n_{Y_2|U_0,U_2}\Big|P^{(\mathbb{C})}_{\mathbf{U}_0,\mathbf{U}_2|M_p=1,M_1=1,M_{22}=1}\Big)\\
&\overset{(c)}{=} \sum_{\mathbf{u}_0,\mathbf{u}_2}\mathbb{E}_{\mathbb{C}_1}\left[D\Big(P^{(\mathbb{C}_1)}_{\mathbf{Y}_2|M_p=1,M_1=1,M_{22}=1,\mathbf{U}_0=\mathbf{u}_0,\mathbf{U}_2=\mathbf{u}_2}\Big\|Q^n_{Y_2|U_0=\mathbf{u}_0,U_2=\mathbf{u}_2}\Big)\mathbb{E}_{\mathbb{C}_0,\mathbb{C}_2}\Big[\mathbb{1}_{\big\{\big(\mathbf{U}_0(1,\mathbb{C}_0),\mathbf{U}_2(1,1,\mathbb{C}_2)\big)=(\mathbf{u}_0,\mathbf{u}_2)\big\}}\Big]\right]\\
&\overset{(d)}{=} \mathbb{E}_{\mathbb{C}_1}D\Big(P^{(\mathbb{C}_1)}_{\mathbf{Y}_2|M_p=1,M_1=1,M_{22}=1,\mathbf{U}_0,\mathbf{U}_2}\Big\|Q^n_{Y_2|U_0,U_2}\Big|Q^n_{U_0,U_2}\Big)
\end{aligned} \tag{17}$$

$$\begin{aligned}
R_1+R'+\tilde{R} &< I(U_1;Y_1|U_0)\\
R_0+R_{20}+R_1+R'+\tilde{R} &< I(U_0,U_1;Y_1)\\
R_{22} &< I(U_2;Y_2|U_0)\\
R_0+R_2-R_{12} &< I(U_0,U_2;Y_2).
\end{aligned} \tag{16}$$

**Security Analysis:** Let $\mathbb{C}_0$ be random variables that represents a random public message codebook. Furthermore, let $\mathcal{C}_j \triangleq \{\mathcal{C}_j(m_p)\}_{m_p\in\mathcal{M}_p}$, for $j=1,2$, be the private message codebooks 1 and 2, and $\mathbb{C}_1$ and $\mathbb{C}_2$ be the corresponding random codebooks. With some abuse of notation, we also use $\mathcal{C}\triangleq(\mathcal{C}_0,\mathcal{C}_1,\mathcal{C}_2)$ and $\mathbb{C}\triangleq(\mathbb{C}_0,\mathbb{C}_1,\mathbb{C}_2)$. Moreover, when clear from the context, we omit the functional dependencies of the $u_j$-codewords, $j=0,1,2$, on the corresponding indices and codebooks, e.g., we write $\mathbf{U}_2$ instead of $\mathbf{U}_2(M_p,M_{22},\mathcal{C}_2)$.

We start with the upper bound in (17) at the top of the page. Step (a) uses the independence of the messages $(M_p,M_1,M_{22})$, the deterministic dependence of $(M_{12},\mathbf{U}_0,\mathbf{U}_2)$ on $(M_p,M_{22})$ and the relative entropy chain rule; (b) follows by the symmetry of the code construction with respect to the messages. To justify (c), first note that for any $\mathbb{C}=\mathcal{C}$, the conditional relative entropy is an expectation of unconditional relative entropies with respect to

$$P^{(\mathcal{C})}(\mathbf{u}_0,\mathbf{u}_2|1,1,1)=\mathbb{1}_{\big\{(\mathbf{u}_0(1,\mathcal{C}_0),\mathbf{u}_2(1,1,\mathcal{C}_2)=(\mathbf{u}_0,\mathbf{u}_2)\big\}}. \tag{18}$$

Combining (18) with the law of total expectation (conditioning the inner expectation on $\mathbb{C}_1$) and noting that $\big(\mathbf{U}_0(1,\mathbb{C}_0),\mathbf{U}_2(1,1,\mathbb{C}_2)\big)$ is independent of $\mathbb{C}_1$, gives (c); finally, (d) relies on the coding PMF being $Q^n_{U_0,U_2}$.

Note that the code construction and the RHS of (17) fall within the framework of Lemma 1. Invoking the lemma, the first two rate bounds in (16) ensure that the RHS of (17) converges to 0 as $n\to\infty$, which establishes strong secrecy. By standard existence arguments and Fourier-Motzkin elimination applied to (16), the achievability of $\mathcal{R}_I$ is established.

**Remark 3** *The main differences between the coding schemes for the cooperative BC with one confidential message and the same channel without secrecy [10] are threefold. First, a randomizer $W$ is used in the secrecy-achieving scheme. Second, the cooperation message $M_{12}$ depends on $M_{20}$ rather than on the pair $(M_{10},M_{20})$ ($M_{10}$ refers to the public part of the message $M_1$). The second difference is because conveying an $M_{12}$ that holds any $M_1$ (in the form of its public part $M_{10}$) violates the secrecy requirement. Finally, a prefix channel $Q_{X|U_0,U_1,U_2}$ is used to optimize randomness and, in turn, to conceal $M_1$ from the 2nd receiver.*


## ACKNOWLEDGEMENTS

The work of Z. Goldfeld and H. H. Permuter was supported by the Israel Science Foundation (grant no. 684/11), an ERC starting grant and the Cyber Security Research Grant at Ben-Gurion University of the Negev. The work of G. Kramer was supported by an Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. The work of P. Cuff was supported by the National Science Foundation (grant CCF-1350595) and the Air Force Office of Scientific Research (grant FA9550-15-1-0180).


## REFERENCES


[1] A. D. Wyner. The wire-tap channel. *Bell Sys. Techn.*, 54(8):1355–1387, Oct. 1975.

[2] I. Csiszár and J. Körner. Broadcast channels with confidential messages. *IEEE Trans. Inf. Theory*, 24(3):339–348, May 1978.

[3] R. Liu, I. Maric, P. Spasojević, and R. D. Yates. Discrete memoryless interference and broadcast channels with confidential messages: Secrecy rate regions. *IEEE Trans. Inf. Theory*, 54(6):2493–2507, Jun. 2008.

[4] Z. Goldfeld, G. Kramer, and H. H. Permuter. Broadcast channels with privacy leakage constraints. *Submitted for publication to IEEE Trans. Inf. Theory*, 2015. Available on ArXiv at http://arxiv.org/abs/1504.06136.

[5] U. Maurer. *Communications and Cryptography: Two Sides of One Tapestry*, chapter The Strong Secret Key Rate of Discrete Random Triples, pages 271–285. Springer US, Norwell, MA, USA, 1994.

[6] I. Csiszár. Almost independence and secrecy capacity. *Prob. Inf. Trans.*, 32(1):40–47, Jan.-Mar. 1996.

[7] M. Hayashi. General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channels. *IEEE Trans. Inf. Theory*, 52(4):1562–1575, Apr. 2006.

[8] M. Bloch and N. Laneman. Strong secrecy from channel resolvability. *IEEE Trans. Inf. Theory*, 59(12):8077–8098, Dec. 2013.

[9] E. Song, P. Cuff, and V. Poor. The likelihood encoder for lossy compression. *submitted to IEEE Trans. on Inf. Theory*, Aug. 2014.

[10] Z. Goldfeld, H. H. Permuter, and G. Kramer. Duality of a source coding problem and the semi-deterministic broadcast channel with rate-limited cooperation. *Accepted for publication to IEEE Trans. Inf. Theory*, 2014. Available on ArXiv at http://arxiv.org/abs/1405.7812.

[11] Z. Goldfeld, H. H. Permuter G. Kramer, and P. Cuff. Strong-secrecy for cooperative broadcast channels. *To be submitted for publication to IEEE Trans. Inf. Theory*, 2015.

[12] P. Cuff. Distributed channel synthesis. *IEEE. Trans. Inf. Theory*, 59(11):7071–7096, Nov. 2013.

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# Context trees for privacy-preserving modeling of genetic data

Lieneke Kusters and Tanya Ignatenko
Eindhoven University of Technology
Department of Electrical Engineering
Eindhoven, The Netherlands
Email: c.j.kusters@tue.nl

***Abstract*—In this work, we use context trees for privacy-preserving modeling of genetic sequences. The resulting estimated models are applied for functional comparison of genetic sequences in a privacy preserving way. Here we define privacy as uncertainty about the genetic source sequence given its model and use equivocation to quantify it. We evaluate the performance of our approach on publicly available human genomic data. The simulation results confirm that the context trees can be effectively used to detect similar genetic sequences while guaranteeing high privacy levels. However, a trade-off between privacy and utility has to be taken into account in practical applications.**

## I. Introduction

With recent advances in genome sequencing technologies, genetic data is increasingly being used in everyday applications. Genes are studied to better understand inherited diseases, such as cancer; personalized medicine based on genetic analysis is used to determine the most effective medicine for patients; genetic material is used to identify individuals in forensic investigations. Therefore, more and more genetic sequences are being collected for research and analysis purposes. Several projects [1], [2], [3] have been initiated to identify and catalog similarities and differences between genetic sequences and relate these findings to medical conditions. The human reference genome has been reconstructed as a general reference and is a representative example of the human genome.

Disclosing one's genetic data is often beneficial for medical purposes, however, this information is also considered to be privacy sensitive. Genetic sequences contain health-related information as well as information about one's ancestry. Furthermore, genetic sequences are unique identifiers of human beings. Therefore privacy concerns are raised as this information can be misused by insurance companies, employers and forensic institutions. Remarkably, one's genome is unique inherited information, of which disclosure is irreversible, thus affecting not only its owner but also the owner's family members for many generations. As such, in modern society protection of genetic data becomes a crucial problem.

### A. Related work

Traditionally, privacy-sensitive information (e.g. in medical profiles) is protected by data anonymization techniques through removing or aggregating the information that can be used to identify the corresponding individual (e.g. a name, birth date or address). Uniqueness of genetic data makes traditional anonymization techniques insufficient, as the data itself uniquely identifies its owner. Therefore, recently, privacy-preserving approaches for genome analysis became the focus of a research community. These approaches aim at analysis of genetic data while protecting the privacy of individuals that are involved.

Erlich *et al.* provided an overview of possible privacy breaches based on the information that can be derived from genetic databases, and reviewed a number of techniques for their prevention [9]. These techniques range from access control and data anonymization techniques to cryptographic solutions. There is also a large branch of research focusing on privacy-preserving sequence comparison. Works in this direction focus on privacy-preserving edit-distance computation between genetic sequences, and typically deploy secure multi-party computation and homomorphic encryption techniques [5], [11]. This approach has however a number of problems. First, due to computational complexity, these algorithms do not scale well to large sequences, while genetic sequences can reach three billion base pairs in length. Therefore, to implement these techniques, outsourcing of the expensive computations to the cloud is used, examples of the corresponding secure protocols can be found in [6], [7]. Moreover, since these approaches are based on cryptographic techniques, their security relies on hardness of the underlying problems and attacker's computational power limitations. However, advances in computer power do not guarantee that such cryptographic techniques remain secure in the future. Therefore, given that the data is unique and inherited, information theoretic security is a desirable property for genomic data protection. Finally, edit-distance does not provide sufficient information to draw conclusions about functional similarity of genetic data to be used in genetic analysis.

### B. Our contribution

In this paper we study generative statistical modeling of genetic sequences applied to sequence comparison and the privacy-preserving properties of these models. Note that genetic sequences can be seen as a natural code that sequentially encodes amino acids and proteins. Therefore models that take into account source memory, such as e.g. hidden Markov models, have been successfully used in genetic sequence

analysis [20]. Here we focus on the context-tree models, proposed and studied by Willems *et al.* [19]. These models are a special case of Markov chains and are closely related to the concept of k-mers[1] used in Bioinformatics for statistical analysis of genetic data [4].

Application of context-tree models for genetic sequence comparison was proposed in [10]. Furthermore, it was shown in [12] that context trees can be applied to model coding and non-coding parts of genetic sequences. Following the Minimum Description Length (MDL) principle [15], in both [10] and [12] compression rate of genetic sequences given an estimated genetic sequence model was used to assess similarity of the compressed and modeled sequences.

In our setting, we assume that a genetic database is composed of genetic sequence models and some associated relevant metadata. This database can be used to search genetic sequences that are functionally similar to a query sequence. Here genetic models are the only genetic information that an attacker can deduce from the database, and therefore, we analyze privacy as the privacy of an underlying genetic source sequence given its model. Such models in general do not correspond to a single sequence but to a class of sequences. Clearly, the larger the class is, the better the privacy guaranties are. The privacy-preserving properties of the models are characterized by the uncertainty about the underlying source sequence and thus correspond to the entropy of the source model. Hence our approach provides information-theoretic security guarantees.

## II. Background and notation

Genetic sequences are described in terms of four symbols from the alphabet $\mathcal{A} = \{A, C, G, T\}$ that correspond to the DNA building blocks, called nucleobases, i.e. Adenine, Cytosine, Guanine and Thymine. Although genomes of different individuals are similar, each individual genome is unique. Differences between genetic sequences result from genetic variations that include *substitution* of one nucleotide with another one and *insertion* or *deletion* of a subsequence of nucleotides.

In the following, we denote by $x_1^N = x_1x_2x_3\dots x_N \in \mathcal{A}^N$ a genetic sequence of length $N$. Furthermore, a sequence of arbitrary length is denoted by $\boldsymbol{x}$, and a source sequence for which we construct the statistical model is denoted by $\tilde{\boldsymbol{x}}$.

## III. Sequence modeling

In this section we first introduce the concept of context trees, see e.g. [19], that we apply to model genetic sequences. Then, we describe our approach to estimate the model corresponding to a given sequence.

### A. Context tree model for genetic sequences

A context tree is a tree structure whose nodes represent the memory of the source. Given an observed source sequence, each node in the tree is associated with a specific context that represents the past of the current symbol in the observed sequence. The leafs of the tree represent the contexts of maximum memory and have associated conditional probabilities, characterizing probability of generating a symbol from the source alphabet given its observed context.

More precisely, a genetic tree source is described in terms of a context-tree model $\langle S_T, P_T\rangle$ represented by a quaternary-tree structure (resulting from the quaternary alphabet for genetic sequences), see e.g. Figure 1. The tree model $\langle S_T, P_T\rangle$ defines the leafs or contexts $\boldsymbol{s} \in S_T$ and the conditional probability distribution $P_T = \{p_T(x|\boldsymbol{s}) : x \in \mathcal{A}, \boldsymbol{s} \in S_T\}$, associated with occurrence of a symbol given its context. For an observed sequence $\boldsymbol{x}$, the context of a symbol $x_i$ is defined by its at most $D$ preceding symbols in reversed order, i.e. $x_{i-1}x_{i-2}\dots x_{i-D}$. Therefore, given a tree model, we can define a mapping $\omega_{S_T}(x_{i-D}^{i-1}) \to \boldsymbol{s} \in S_T$ that maps $D$ preceding symbols of $x_i$ to a leaf in the tree model and thus retrieve $p_T(x_i|\boldsymbol{s})$, the probability of occurrence of symbol $x_i$ after context $\boldsymbol{s}$ was observed. Then, given a tree source model, we can determine the probability of sequence $\boldsymbol{x}$ being generated by this source as follows:

$$\Pr(\boldsymbol{x}|\langle S_T, P_T\rangle) = \prod_i p_T(x_i|\omega_{S_T}(x_{i-D}^{i-1})). \tag{1}$$

In general, for a given sequence, we do not know the actual source model that has generated this sequence. Therefore, we have to find a good estimate for it, i.e. given this sequence we have to estimate the corresponding contexts in the tree and the associated conditional distribution. In the following, we present the algorithm to estimate such models.

### B. Model estimation

In order to estimate a tree model corresponding to sequence $\boldsymbol{x}$, we assume the model depth (or memory) to be $D$. Now to estimate the sequence statistics, we process the sequence $\boldsymbol{x}$ in a sequential way. For each symbol in the sequence we find its context, defined by its $D$ preceding symbols, and count the total number of occurrences $n_{\boldsymbol{s}}^a(\boldsymbol{x})$ of a symbol $a \in \mathcal{A}$ with each observed context in the sequence. Here each distinct observed context becomes a leaf in the tree. Moreover, for the first symbol we assume some initial context $\boldsymbol{s}_1$. Now, given the contexts and the counts, the conditional probability distribution $p_T(x|\boldsymbol{s})$ is estimated as follows:

$$p_T(x|\boldsymbol{s}) = \frac{n_{\boldsymbol{s}}^x(\boldsymbol{x}) + 1/2}{\sum_{a\in\mathcal{A}} n_{\boldsymbol{s}}^a(\boldsymbol{x}) + |\mathcal{A}|/2} \quad x \in \mathcal{A}, \boldsymbol{s} \in S_T. \tag{2}$$

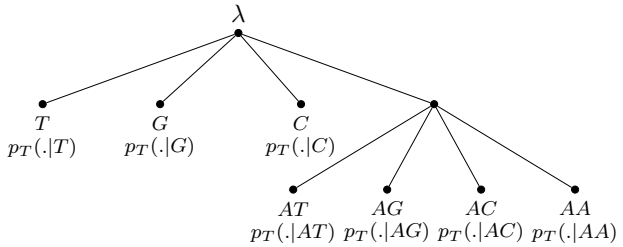



Fig. 1. An example of the context-tree model $\langle S_T, P_T\rangle$ of depth 2, with root $\lambda$ and set of leafs $S_T = \{T, G, C, AT, AG, AC, AA\}$ with corresponding conditional probabilities, given by $P_T = \{p_T(\cdot|\boldsymbol{s}) : \forall \boldsymbol{s} \in S_T\}$.

[1]Typically, k-mers refer to all subsequences of length $k$ occurring in the genetic sequence.

The source model of the sequence $\boldsymbol{x}$ is now defined by $\langle S_T, P_T\rangle_{\boldsymbol{x}}$, where $S_T$ corresponds to the distinct contexts observed in the processed sequence $\boldsymbol{x}$ and represents the leafs of the tree model, and $P_T = \{p_T(x|\boldsymbol{s}), x \in \mathcal{A}, \boldsymbol{s} \in S_T\}$ corresponds to the conditional probability distribution associated with each of these leafs.

## IV. Sequence comparison

In order to apply our context-tree models for sequence comparison, we need to define a measure to evaluate similarity of an arbitrary genetic sequence $\boldsymbol{x}$ to a source sequence $\tilde{\boldsymbol{x}}$, given only its model $\langle S_T, P_T\rangle_{\tilde{\boldsymbol{x}}}$. The MDL principle [15] states that the model that describes the data in the shortest possible way is the model that most probably generated this data. Observe that compression rate characterizes the amount of bits required to describe data in a concise way. Furthermore, the techniques presented in the previous section originally aimed at finding good source coding distributions, and are characterized by asymptotically optimal performance. Thus based on the MDL principle and the fact that a model that closely relates to an observed sequence also results in a low compression rate, we use compression rate to characterize sequence similarity.

The compression rate of a sequence $x_1^M$, given a tree model $\langle S_T, P_T\rangle$ is estimated as the logarithm of the probability of the sequence being generated by this model:

$$R(x_1^M|\langle S_T, P_T\rangle) = -\frac{1}{M}\sum_{i=1}^{M}\log_2(p_T(x_i|\omega_{S_T}(x_{i-D}^{i-1}))). \quad (3)$$

For our problem, we use parameter $p_T$ equals $1/4$, when there is no leaf in $S_T$ associated with a subsequence in $x_1^M$.

Clearly, compression rate of a sequence given its model depends on both accuracy and complexity of this model. Therefore, to take into account model complexity we will use compression rate of a sequence $\hat{\boldsymbol{x}}$ given a source model of $\tilde{\boldsymbol{x}}$ relative to the compression rate of this sequence given its own model. Thus in this work we use the normalized information gain as a similarity measure, defined as

$$\frac{\Delta R(\hat{\boldsymbol{x}})}{R_{\hat{\boldsymbol{x}}}} = \frac{R(\hat{\boldsymbol{x}}|\langle S_T, P_T\rangle_{\tilde{\boldsymbol{x}}}) - R_{\hat{\boldsymbol{x}}}}{R_{\hat{\boldsymbol{x}}}}, \quad (4)$$

where $R_{\hat{\boldsymbol{x}}} = R(\hat{\boldsymbol{x}}|\langle S_T, P_T\rangle_{\hat{\boldsymbol{x}}})$ is the compression rate of the sequence $\hat{\boldsymbol{x}}$ given its own model.

Note that while estimating the compression rate for the query sequence, the best (lowest) compression rate is achieved with the model that corresponds to the query sequence itself. Furthermore, the more similar the query sequence is to the source sequence, the more similar their models are, and thus the better the compression that is achieved using the model corresponding to the source sequence. Therefore, for statistically similar sequences the resulting information gain will be small. Also the opposite is true, and for sequences with very distinct models, the information gain achieved with the correct model will be high.

We will use information gain estimates based on the context-tree models to determine whether the genetic sequences are (functionally) similar or not. The utility of our approach is tested on the publicly available genetic data to which mutations are introduced, in Section VI.

## V. Privacy analysis

Another aspect of genetic sequence modeling that we study in this paper is privacy. Here we regard privacy as uncertainty about the source sequence $\tilde{\boldsymbol{x}}$, given its corresponding model $\langle S_T, P_T\rangle_{\tilde{\boldsymbol{x}}}$. For the attack model we assume that an attacker only has access to the sequence models. Note that the sequence models actually correspond to a class of sequences, therefore the attacker's best strategy is to reconstruct the most probable sequence that can be generated by the given sequence model. However, the context-tree models correspond to a class of sequences that are all generated by this model with equal probability. Thus the attacker cannot distinguish the original sequence $\tilde{\boldsymbol{x}}$ from the other sequences that can be generated by the model.

In order to measure the privacy level of our models, we define privacy as a function of the number of sequences corresponding to the model of the underlying source sequence. This measure of privacy is known as equivocation [16] and is given by

$$E(\tilde{\boldsymbol{x}}) = H(\tilde{\boldsymbol{x}}|\langle S_T, P_T\rangle_{\tilde{\boldsymbol{x}}}) = \log_2(K), \quad (5)$$

where $K$ is the number of sequences that are generated by the model $\langle S_T, P_T\rangle_{\tilde{\boldsymbol{x}}}$. The next step is to estimate the equivocation of our context-tree models.

### A. Type-class cardinality

The type-class $\mathcal{T}$, see [8] and [14], of a sequence $\tilde{\boldsymbol{x}}$ with respect to a tree model is defined as the set of sequences that have the same symbol counts $n_{\boldsymbol{s}}^a$ in the leafs of their tree model as the source sequence $\tilde{\boldsymbol{x}}$, i.e.:

$$\mathcal{T}_{S_T}(\tilde{\boldsymbol{x}}) = \{\hat{\boldsymbol{x}} \in \mathcal{A}^N : n_{\boldsymbol{s}}^a(\hat{\boldsymbol{x}}) = n_{\boldsymbol{s}}^a(\tilde{\boldsymbol{x}}), \forall \boldsymbol{s} \in S_T, a \in \mathcal{A}\}. \quad (6)$$

Note that our context-tree model only contains conditional probabilities and no symbol counts in its leafs. Therefore, to construct the type-class, one first needs to estimate the counts from the available conditional probabilities. This requires knowledge of the length of the original sequence used to construct the model. While this information is in general not available, we can assume that one (e.g. the attacker) can approximate or make an assumption about the sequence length $\hat{N}$, approximate the counts, and then construct the type-class. Clearly, since all the sequences from the type-class are equiprobable from an attacker point of view, the cardinality of the type-class provides us with the number of sequences $K$ that are generated by the corresponding model, and thus can be used to get the equivocation estimate.

We use Whittle's formula [18] to calculate the cardinality of a type-class, given the transition frequency matrix $F_S$ (see the next subsection for the details) and the initial $\boldsymbol{s}_1$ and final $\boldsymbol{s}_L$ contexts corresponding to the source sequence:

$$K = |\mathcal{T}_S(\tilde{\boldsymbol{x}})| = C_S\frac{\prod_{\boldsymbol{s}\in S}F_S(\boldsymbol{s},*)!}{\prod_{\boldsymbol{t},\boldsymbol{v}\in S}F_S(\boldsymbol{t},\boldsymbol{v})!}, \quad (7)$$

where $F_S(\boldsymbol{s},*) = \sum_{\boldsymbol{v}\in S} F_S(\boldsymbol{s},\boldsymbol{v})$ is the number of symbols emitted with context $\boldsymbol{s}$, and $C_S$ is the cofactor of entry $(\boldsymbol{s}_L, \boldsymbol{s}_1)$ of $I - \hat{F}_S$, with $\hat{F}_S$ being the transition frequency matrix with normalized rows.

### B. Transition frequency matrix

Given the context-tree model and the symbol counts in its leafs, an element of the transition frequency matrix $F_S(\boldsymbol{t},\boldsymbol{v})$ gives the number of times context $\boldsymbol{t}$ is being followed by context $\boldsymbol{v}$ in the source sequence, i.e.

$$F_{S_T}(\boldsymbol{t},\boldsymbol{v}) = |\{i : 1 \le i \le N, \boldsymbol{s}_i = \boldsymbol{t}, \boldsymbol{s}_{i+1} = \boldsymbol{v}\}|, \quad \boldsymbol{t},\boldsymbol{v} \in S_T, \tag{8}$$

where the context sequence $\boldsymbol{s}_1^L = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L\}$ is the concatenated set of contexts that occur successively in the source sequence.

For the context-tree models described in Section III, the next context $\boldsymbol{s}_{i+1}$ follows from the current context $\boldsymbol{s}_i$ and current ($i^{th}$) symbol $x_i$ in the sequence: $\boldsymbol{s}_{i+1} = x_i \boldsymbol{s}_1^{D-1}$, where $D$ is the depth of the tree. Therefore, the transition frequency matrix can be constructed directly from the model counts $n_{\boldsymbol{s}}^a, \boldsymbol{s} \in S_T$ as

$$F_{S_T}(\boldsymbol{t},\boldsymbol{v}) = \begin{cases} n_{\boldsymbol{t}}^a, & \boldsymbol{v} = a\boldsymbol{t}_1^{D-1}, \quad a \in \mathcal{A}, \quad \boldsymbol{t},\boldsymbol{v} \in S_T \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

In order to apply Whittle's formula and evaluate the type-class size, besides the transition frequency matrix $F_S$, we also need to know the initial and final contexts that occurred in the source sequence. We assume that the initial context $\boldsymbol{s}_1$ is predefined and that it is known. The final context $\boldsymbol{s}_L$ can then be derived from the transition frequency matrix and the initial context using the flow conservation equations:

$$F_S(*,\boldsymbol{s}) + \delta_{\boldsymbol{s},\boldsymbol{s}_1} = F_S(\boldsymbol{s},*) + \delta_{\boldsymbol{s},\boldsymbol{s}_L}, \quad \boldsymbol{s} \in S, \tag{10}$$

$$\delta_{\boldsymbol{s},\boldsymbol{t}} = \begin{cases} 1 & \text{if } \boldsymbol{s} = \boldsymbol{t} \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where $F_{S_T}(*,\boldsymbol{s}) = \sum_{\boldsymbol{t}\in S} F_S(\boldsymbol{t},\boldsymbol{s})$ is the number of transitions into context $\boldsymbol{s}$. Now, given the transition frequency matrix and initial and final contexts, the privacy-preserving properties of our context-tree models are given by equivocation (5) estimated with the help of Whittle's formula (7).

## VI. Experimental results

In order to evaluate the performance of context trees for privacy-preserving sequence modeling, we apply our method to subsequences of the human reference genome [17]. We use a set of distinct genomic subsequences, each corresponding to a different gene. Moreover, for each gene we have generated a set of similar sequences by applying a predefined number of mutations. It is well-known that on average the genomes of human individuals are $99.5\%$ similar [13]. These $0.5\%$ of genetic variations come from different types of mutations, such as single nucleotide variants, indels (insertions or deletions of a block of nucleotides of length $< 100$ ), and large-scale structural variants. Furthermore, most genetic variation occurs in a limited region of the human genome. In this work, we use a simplified mutation model, where we apply single nucleotide mutations of 1 per 1000, 1 per 100, and 1 per 10 base-pairs, as well as indels of length 10, occurring once per 1000 base-pairs, in order to simulate a set of similar sequences.

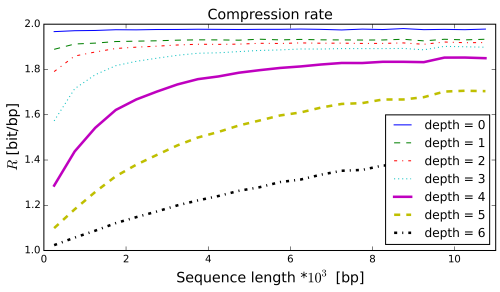



Fig. 2. Average compression rate of sequences based their own models.

### A. Utility performance

First, we evaluate the utility performance of the models for genetic sequence comparison as a function of tree-depth (characterizing model complexity) and sequence length. We start with estimating the compression rate of the source sequence, see Figure 2. Note that as the compression rate is measured in bits per base-pair (bp), a compression rate smaller than 2 corresponds to actual compression of the sequence. On average the compression performance of the context-tree models improves when larger tree-depth is used. Since compression rate indicates how well the model fits the data, we may conclude that including larger memory in the models and thus increasing its complexity helps to better describe the gene data. This can be supported by data interpretation, since genes correspond to coding regions of the genomes. However, compression rate of individual sequences does not provide information about the specificity of the model when distinguishing between different sequences.

Next, we evaluate the performance of the context-tree models in distinguishing whether the sequences are similar or not. We have evaluated the normalized information gain for sequences with various rates of mutation in comparison to the source sequence, see Figure 3. Furthermore, we have also applied the context-tree models to estimate the normalized information gain corresponding to the other genes, see

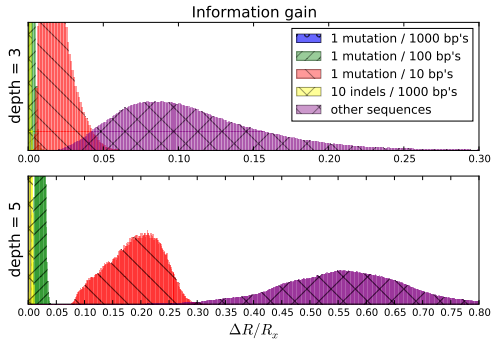



Fig. 3. Histograms of the distributions of the normalized information gain for sequences with different mutation rates.

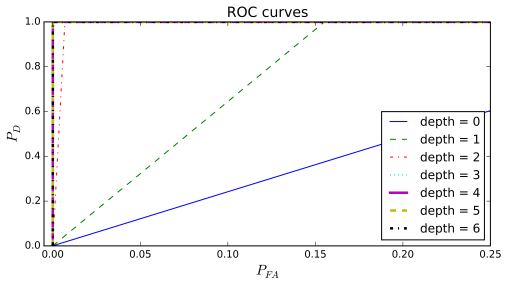



Fig. 4. Performance of context-tree models in distinguishing between sequences with similar (limited number of mutations) and distinct (other genes) functionalities.

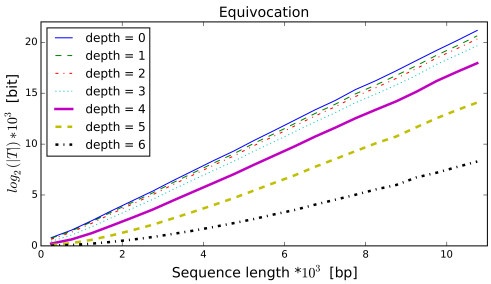



Fig. 5. The privacy and utility properties for tree models of various depths.

Figure 3. Based on our results, we conclude that the context-tree models can be applied to efficiently distinguish between similar (sequences with mutation rate smaller than 1/100) and non-similar sequences by defining an appropriate threshold. Note that models with higher complexity perform better in distinguishing between similar and distinct sequences. This is not surprising given the compression results discussed above.

Finally, we evaluate the performance of the context-tree models for distinguishing sequences with different functionalities. In this experiment we use sequences corresponding to different genes (or functionalities). Furthermore, we have simulated a set of sequences with similar functionalities by introducing the following mutations: 1 mutation per 1000, 1 mutation per 100, and 10 indels per 1000 base-pairs in comparison to the source sequences. Figure 4 shows the resulting ROC curves demonstrating the performance of models with different complexity. We conclude that a perfect distinction can be achieved with the models of depth 3 or larger, but also the models of depth 2 achieve an acceptable performance.

### B. Privacy performance

Finally, we evaluate the privacy-preserving properties of our models in terms of equivocation. We plot equivocation as a function of sequence length in Figure 5. We see that with our models we can achieve very high privacy levels. In contrast to the information gain, improved equivocation is achieved for smaller tree-depth. Therefore there exists a trade-off between privacy and utility that should be taken into account when selecting models for privacy-preserving genetic sequence comparison.

## VII. Conclusions

In this paper we have studied the use of context-tree models for privacy-preserving modeling of genetic sequences in application to sequence comparison. We have focused on functional sequence similarity that can be expressed as statistical similarity of compared sequences, and used normalized information gain as a similarity measure. Furthermore, privacy of the context-tree models is given in terms of equivocation, that characterizes uncertainty about the source sequence given its model. Based on the experimental results, we can conclude that context-tree models can be successfully applied for privacy-preserving sequence comparison resulting in both good discriminating and privacy performance. However, since there is a trade-off between privacy and utility, model complexity has to be adjusted meeting the requirements of a specific application.

## Acknowledgment

This work has been partially funded by the EC via grant agreement no. 611659 for the AU2EU project. The authors would also like to thank dr. J.F.J. Laros for the fruitful discussions on genomic data and its analysis.

## References

[1] The international hapmap project. *Nature*, (426):789–796, 2003.
[2] An integrated encyclopedia of DNA elements in the human genome. *Nature*, (489):57–74, 2012.
[3] An integrated map of genetic variation from 1,092 human genomes. *Nature*, (491):56—65, 2012.
[4] S. Anvar et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome Biol.*, 15(12):555, 2014.
[5] M. J. Atallah, F. Kerschbaum, and W. Du. Secure and private sequence comparisons. In *Proc. 2003 ACM Work. Priv. Electron. Soc.*, page 39, New York, New York, USA, 2003. ACM Press.
[6] M. J. Atallah and J. Li. Secure outsourcing of sequence comparisons. *Int. J. Inf. Secur.*, 4(4):277–287, 2005.
[7] M. Blanton et al. Secure and efficient outsourcing of sequence comparisons. In *Eur. Symp. Res. Comput. Secur.*, pages 505–522, 2012.
[8] I. Csiszâr. The method of types. *IEEE Trans. Inf. Theory*, 44(6):2505–2523, 1998.
[9] Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.*, 15(6):409–21, 2014.
[10] T. Ignatenko and M. Petković. AU2EU: Privacy-Preserving Matching of DNA Sequences. In *Proc. 8th IFIP WG 11.2 Int. Work. Inf. Secur. Theory Pract. Secur. Internet Things*, pages 180–189. Springer Berlin Heidelberg, 2014.
[11] S. Jha, L. Kruger, and V. Shmatikov. Towards practical privacy for genomic computation. In *IEEE Symp. Secur. Priv.*, pages 216–230, 2008.
[12] L. Kusters and T. Ignatenko. DNA sequence modeling based on context trees. In *Proc. 5th Jt. WIC/IEEE Symp. Inf. Theory Signal Process. Benelux*, pages 96–103, 2015.
[13] S. Levy et al. The diploid genome sequence of an individual human. *PLoS Biol.*, 5(10):2113–2144, sep 2007.
[14] Á. Martín, G. Seroussi, and M. J. Weinberger. Type classes of context trees. *IEEE Trans. Inf. Theory*, 58(7):4077–4093, jul 2012.
[15] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
[16] L. Sankar, S. R. Rajagopalan, and H. V. Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Trans. Inf. Forensics Secur.*, 8(6):838–852, 2013.
[17] T. Tatusova et al. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, 42(1):D553–9, 2014.
[18] P. Whittle. Some Distribution and Moment Formulae for the Markov Chain. *J. Roy. Statist. Soc. B*, 17(2):235–242, 1970.
[19] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. Context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory*, 41(3):653–664, May 1995.
[20] B. Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics*, 10:402–415, 2009.

# Hybrid DF-CF-DT For Buffer-Aided Relaying

Mohammad Shaqfeh*, Ammar Zafar†, Hussein Alnuweiri*, and Mohamed-Slim Alouini‡
* Texas A&M University at Qatar, Email: {mohammad.shaqfeh, hussein.alnuweiri}@qatar.tamu.edu
† University of Technology Sydney (UTS), Email: ammar.zafar@uts.edu.au
‡ King Abdullah University of Science and Technology (KAUST), Email: slim.alouini@kaust.edu.sa

***Abstract*—In this paper, we maximize the expected achievable rate of buffer-aided relaying by using a hybrid scheme that combines three transmission strategies, which are decode-and-forward (DF), compress-and-forward (CF) and direct transmission (DT). The proposed hybrid scheme is dynamically adapted based on the channel state information. This includes adjusting the data rate and compression when compress-and-forward is selected. We apply this scheme to three different models of the Gaussian block-fading relay channel, depending on whether the relay is half or full duplex and whether the source and the relay have orthogonal or non-orthogonal channel access. The integration and optimization of these three strategies provide a more generic and fundamental solution and give better achievable rates than the known schemes for buffer-aided relaying. We compare the achievable rates to the upper-bounds of the ergodic capacity for each one of the three channel models.**

## I. Introduction

As well-known, important capacity theorems were established for the physically degraded and reversely degraded discrete memoryless full-duplex relay channel in [1]. This topic has emerged as an important research area in the wireless communication field as well [2], [3]. Achievable rates and capacity upper-bound results for half-duplex relays in fixed-gain Gaussian channels were provided in the literature assuming non-orthogonal channel access of the source and relay [4], and also assuming orthogonal channel access [5], [6]. More recent results were provided in [7]. We know from these references that, similar to the full-duplex case [1], [8], the best known upper bounds on the capacity are the max-flow min-cut bounds, and that there are three different coding strategies that maximize the achievable rates, which are decode-and-forward (DF), compress-and-forward (CF) and direct transmission (DT) from the source to the destination. None of these three strategies is globally dominant over the other two, but rather each one of them can achieve higher rates than the others in specific scenarios depending on the qualities of the source-relay, source-destination and relay-destination channels. Furthermore, there are other contributions in the literature that consider fading relay channels. For example, the quasi-static (block-fading) half-duplex relay channel was studied, and it was shown that dynamic adaptation of the transmission strategies using DF and DT is needed in order to maximize the expected achievable rates [9]. However, CF was not considered and channel allocation was fixed beforehand and not subject to optimization therein. It is obvious that making channel allocation dynamic and subject to optimization would add to the degrees of freedom in the system design and enable achieving higher rates. Optimal channel allocation for Gaussian (non-fading) orthogonal and non-orthogonal relay channels was considered in a number of papers, and the obtained results for the best achievable schemes were based on DF only [5], [10], [4]. Recently, "buffer-aided relaying" was proposed and studied for the cases when there is no direct link from the source to the destination [11], [12], [13], and also when the direct link is available and utilized [14]. We are interested in the latter case in this work.

Having gone through many of the most important works in the literature that considered block-fading relay channels, we still believe that there is room for improvement since they all focus on dynamic adaptation of decode-and-forward relaying strategies and they do not consider compress-and-forward as well, although there are certain scenarios over which CF can be better than DF as we know from the case of fixed-gain channels. So, in this work, we consider a buffer-aided hybrid scheme that combines DF, CF and DT and switches among them dynamically based on the channel conditions, and we consider optimizing the resource allocation for this hybrid scheme to maximize the expected achievable rates. We believe that this is an important contribution to the literature since it is more generic than the known schemes and, hence, it can achieve higher rates when optimized properly.

Before we end this section, we want to mention that the concept of "buffer-aided relaying" was also considered for dual-hop broadcast channels and it was called "joint user-and-hop scheduling" since the buffering capabilities are actually needed to enable dynamic and flexible scheduling (i.e. channel allocation) among multiple users (destination nodes) and the relay [15]. Also, it was also applied to other channel models that involve relaying such as the bi-directional relay channel [16], [17], the shared relay channel [18] and overlay cognitive radio networks [19]. The list of references on buffer-aided relaying provided here is not exhaustive.

## II. Block-Fading Relay Channel Models

We consider a three-node network that consists of a source (S) that wants to send information to a destination (D) with the assistance of a relay (R). We assume a Gaussian block-fading model for the channels between the nodes. We also assume that all channel blocks have the same duration ($T$ in seconds) and bandwidth ($W$ in Hz) and that they are large enough to

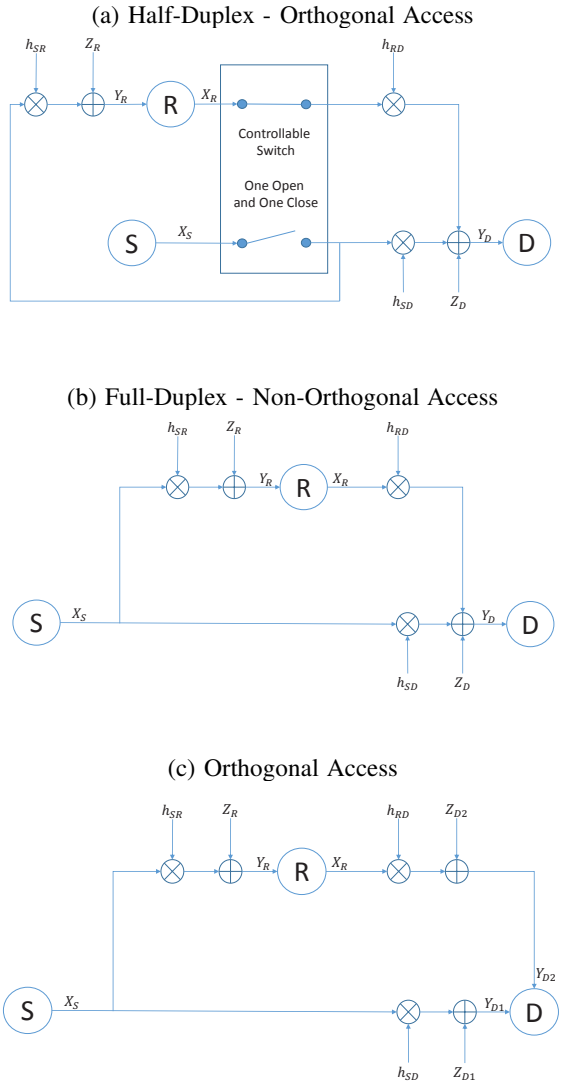



Fig. 1. Channel Models

achieve the instantaneous capacity[1]. Furthermore, we assume that the source and relay transmit using a constant (maximum) power per unit bandwidth (in Jouls/sec/Hz). We investigate three different models for the relay channel that are shown in Fig 1. We call them; (a) half-duplex – orthogonal access, (b) full-duplex – non-orthogonal access, and (c) orthogonal access. In the figure, $X_{\mathsf{S}}[k]$ and $X_{\mathsf{R}}[k]$ are the transmitted (complex field) source signal and relay signal, respectively, in channel block $k$. Similarly, $Y_{\mathsf{R}}[k]$ and $Y_{\mathsf{D}}[k]$ are the received signals at the relay and destination, respectively, and $Z_{\mathsf{R}}[k]$ and $Z_{\mathsf{D}}[k]$ are the added Gaussian noise at these two nodes, which are mutually independent and have circularly symmetric, complex Gaussian distribution with unit variance. Furthermore, $h_{\mathsf{SD}}[k]$, $h_{\mathsf{SR}}[k]$ and $h_{\mathsf{RD}}[k]$ are the channel complex coefficients, which stay constant during one channel block $k$ and change randomly afterwards. The corresponding signal-to-noise-ratio (SNR) of these channels, in a given channel block $k$, are denoted $\gamma_{\mathsf{SR}}[k]$, $\gamma_{\mathsf{SD}}[k]$ and $\gamma_{\mathsf{RD}}[k]$, respectively, where $\gamma[k] = |h[k]|^2\bar{P}$. The probability density function (PDF) of the channel gain ($|h|^2$) over each one of the three links is a continuous function. Over each link, the receiver knows the channel complex coefficient $h[k]$ perfectly, but the corresponding *transmitter*[2] *knows only the channel gain* $|h|^2$.

The controllable switch in channel model (a) makes only one of the two nodes (source or relay) transmit (subject to optimization). In channel model (c), $Y_{\mathsf{D}_1}[k]$ and $Y_{\mathsf{D}_2}[k]$ are the received signals from the source and the relay, respectively, over orthogonal channels. Both $Z_{\mathsf{D}_1}[k]$ and $Z_{\mathsf{D}_2}[k]$ are added Gaussian noise with unit variance. We assume that the two orthogonal channels have the same size ($TW$).

The instantaneous (i.e. in a given channel block $k$) channel capacities are denoted be $C_{\mathsf{SD}}[k]$, $C_{\mathsf{SR}}[k]$ and $C_{\mathsf{RD}}[k]$ for the source-destination, source-relay and relay-destination links, respectively. For channel models (a) and (c), where we have orthogonal access, the channel capacities (per unit bandwidth) follow the well-known capacity of AWGN channels $C_x[k] = \log\left(1+\gamma_x[k]\right)$, $\forall x \in \{\mathsf{SD}, \mathsf{SR}, \mathsf{RD}\}$. For channel model (b), where we have non-orthogonal access, the source-relay link will still be an AWGN channel. On the other hand, the source-destination and relay-destination links form a multiple-access channel (MAC). It can be proven[3] that for optimality, the destination should decode the relay's message first and then process the source's message. Thus, $C_{\mathsf{SD}}[k] = \log\left(1+\gamma_{\mathsf{SD}}[k]\right)$ and $C_{\mathsf{RD}}[k] = \log\left(1+\frac{\gamma_{\mathsf{RD}}[k]}{1+\gamma_{\mathsf{SD}}[k]}\right)$.

## III. Communication System Description

### A. System Requirements

We investigate a hybrid communication scheme that combines three different strategies; direct transmission (DT), decode-and-forward (DF) and compress-and-forward (CF). These schemes are adapted dynamically and optimally based on the channel conditions in order to maximize the expected achievable rate. When the source transmits a new codeword, it decides (subject to optimization) if the codeword will be used for DT, DF or CF, and it adjusts the data rate (denoted by $R_{\mathsf{DT}}[k]$, $R_{\mathsf{DF}}[k]$ and $R_{\mathsf{CF}}[k]$) of the codeword accordingly. As an optimization framework, we assume that the proposed hybrid scheme uses orthogonal time-sharing of the three transmission strategies in the same channel block. The time sharing ratios are subject to optimization. For notation, $\theta_{\mathsf{DT}}[k]$, $\theta_{\mathsf{DF}}[k]$ and $\theta_{\mathsf{CF}}[k]$ denote the time sharing ratio in a given channel block $k$ for the DT, DF and CF transmission

[1]As well-known, achieving the capacity requires infinite code length. However, with sufficiently long codewords, we can transmit at channel capacity with negligible probability of error.

[2]This assumption is stemmed from practical system design considerations. As a consequence of it, beamforming of the source and relay signals towards the destination is not feasible, and, hence, $\beta$ in formulas (5) and (7) in [4] equals zero under our assumptions.

[3]The proof is omitted here for brevity. It is available in the full-version of this paper, which is available online [20].

strategies, respectively. They refer to the source transmission phase of all of these strategies.

In the DF case, the relay fully decodes the source message and it generates and stores an amount of information, denoted by $R^*_{\mathsf{DF}}[k]$ that would be sufficient for the destination to decode the source message reliably (given that the destination utilizes both the source and relay signals to decode the source codeword). For example, the relay can store a bin index (in the sense of Slepian-Wolf coding [21]) of the source message that indicates the partition at which the source codeword lies. In the CF case, the relay encodes and stores a quantized version of the received signal using, e.g. Wyner-Ziv lossy source coding [22]. The data rate of this generated message by the relay is denoted by $R^*_{\mathsf{CF}}[k]$.

In addition to the availability of the channel state information, another important requirement to support the adaptivity of the system is having unlimited buffering capability at the relay and the destination. This is because when the source transmits a new codeword and the relay decodes or compresses it, it does not forward it directly to the destination in the same or the following channel block, but it rather stores it and it adjusts its transmission rate based on the relay-destination channel quality. Thus, the relay might send the information bits that corresponds to one codeword of the source over multiple channel blocks or combine the information bits that corresponds to more than one codeword of the source. This was properly explained in [14].

### *B. Data Rates of* CF

In CF, the data rate of the source codeword is bounded by the capacity of the single-input multiple-output (SIMO) channel assuming that the relay and destination are two antennas of the same receiver.

$$R_{\mathsf{CF}}[k] < \theta_{\mathsf{CF}}[k] \log\left(1+\gamma_{\mathsf{SR}}[k]+\gamma_{\mathsf{SD}}[k]\right) \tag{1}$$

Notice that if $R_{\mathsf{CF}}[k] \leq \theta_{\mathsf{CF}}[k] C_{\mathsf{SD}}[k]$, then the destination can decode the source message via direct transmission and the relay does not need to forward anything. For notation, we define $\gamma_{\mathsf{CF}}[k] = \exp\left(\frac{R_{\mathsf{CF}}[k]}{\theta_{\mathsf{CF}}[k]}\right) - 1$, where the data rate is measured in nats/sec/Hz.

*Theorem 1 (Rate of compressed signal at the relay):*

Given that $\gamma_{\mathsf{SD}}[k] < \gamma_{\mathsf{CF}}[k] < \gamma_{\mathsf{SR}}[k] + \gamma_{\mathsf{SD}}[k]$, the data rate of the encoded compressed signal by the relay must satisfy

$$\frac{R^*_{\mathsf{CF}}[k]}{\theta_{\mathsf{CF}}[k]} \geq \log\left(1+\frac{\left(\gamma_{\mathsf{CF}}[k]-\gamma_{\mathsf{SD}}[k]\right)\left(1+\gamma_{\mathsf{SD}}[k]+\gamma_{\mathsf{SR}}[k]\right)}{\left(\gamma_{\mathsf{SD}}[k]+\gamma_{\mathsf{SR}}[k]-\gamma_{\mathsf{CF}}[k]\right)\left(1+\gamma_{\mathsf{SD}}[k]\right)}\right) \tag{2}$$

in order to enable the destination to decode the source message reliably.

The proof is omitted here due to space constraint. It is available in [20].

### *C. Optimization Problem Formulation*

We write the main optimization problem in a generic form that is applied to the three channel models in Fig. 1. We want to maximize the average total achievable rate of the relay channel, which is the sum of the rates achieved by the three transmission strategies. The relay should transmit sufficient amount of rate to enable the destination to decode the source messages reliably.

$$\max_{\zeta[k] \ \forall k} \quad \bar{R}_{\mathsf{DT}} + \bar{R}_{\mathsf{DF}} + \bar{R}_{\mathsf{CF}} \tag{3a}$$

$$\text{subject to} \quad \bar{R}_{\mathsf{RD}} \geq \bar{R}^*_{\mathsf{DF}} + \bar{R}^*_{\mathsf{CF}}, \tag{3b}$$

$$R_{\mathsf{DT}}[k] = \theta_{\mathsf{DT}}[k] C_{\mathsf{SD}}[k], \tag{3c}$$

$$R_{\mathsf{DF}}[k] = \theta_{\mathsf{DF}}[k] C_{\mathsf{SR}}[k], \tag{3d}$$

$$R^*_{\mathsf{DF}}[k] = \theta_{\mathsf{DF}}[k]\Big(C_{\mathsf{SR}}[k] - C_{\mathsf{SD}}[k]\Big)^+, \tag{3e}$$

$$R_{\mathsf{RD}}[k] = \min\left(\theta_{\mathsf{RD}}[k] C_{\mathsf{RD}}[k], Q[k]\right) \tag{3f}$$

in addition to (1) and (2) (at equality).

In (3), $\bar{X} = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} X[k]$, $\forall X \in \{R_{\mathsf{DT}}, R_{\mathsf{DF}}, R_{\mathsf{CF}}, R_{\mathsf{RD}}, R^*_{\mathsf{DF}}, R^*_{\mathsf{CF}}\}$. Furthermore, $Q[k]$ is the normalized total amount of information stored in the relay's buffers at the start of channel block $k$, and $(x)^+ = \max(x, 0)$. $\zeta[k] = \{\theta_{\mathsf{DT}}[k], \theta_{\mathsf{DF}}[k], \theta_{\mathsf{CF}}[k], R_{\mathsf{CF}}[k], \theta_{\mathsf{RD}}[k]\}$ is the set of optimization variables for channel model (a). They are constrained by

$$\theta_{\mathsf{DT}}[k] + \theta_{\mathsf{DF}}[k] + \theta_{\mathsf{CF}}[k] + \theta_{\mathsf{RD}}[k] = 1 \tag{4}$$

Eq. (4) only applies to channel model (a). In channel models (b) and (c), $\theta_{\mathsf{RD}}[k] = 1$ over all channel blocks, and

$$\theta_{\mathsf{DT}}[k] + \theta_{\mathsf{DF}}[k] + \theta_{\mathsf{CF}}[k] = 1 \tag{5}$$

## IV. Optimal Solution

We go through the main steps to be able to obtain the solution of (3).

*Lemma 1 (Queue at edge of non-absorption):*

A necessary condition for the optimal solution of (3) is that the queue in the buffer of the relay is at the edge of non-absorption. Consequently, for $K \to \infty$, the impact of the event $Q[k] < \theta_{\mathsf{RD}}[k] C_{\mathsf{RD}}[k]$, $k = 1, \cdots, K$ is negligible. Therefore, the optimal solution will have

$$\bar{R}_{\mathsf{RD}} = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \theta_{\mathsf{RD}}[k] C_{\mathsf{RD}}[k] \tag{6}$$

and the constraint (3b) will be satisfied at equality.

The proof follows the same steps that are given in [11, Theorem 1 and Theorem 2].

By using the Lagrangian dual problem of (3), we get

$$\max_{\zeta[k] \ \forall k} \quad \bar{R}_{\mathsf{DT}} + \bar{R}_{\mathsf{DF}} + \bar{R}_{\mathsf{CF}} - \lambda\left(\bar{R}^*_{\mathsf{DF}} + \bar{R}^*_{\mathsf{CF}} - \bar{R}_{\mathsf{RD}}\right) \tag{7}$$

where $\lambda \geq 0$ is the Lagrangian multiplier. A direct consequence of Lemma 1, in particular (6), is that the achievable rates in a given channel block $k$ are only dependent on their respective optimization variables $\zeta[k]$. Therefore, (7) can be transformed into a number $K$ of independent optimization problems that are solved independently.

$$\max_{\zeta[k]} R_{\mathsf{DT}}[k] + R_{\mathsf{DF}}[k] + R_{\mathsf{CF}}[k] - \lambda\left(R^*_{\mathsf{DF}}[k] + R^*_{\mathsf{CF}}[k] - R_{\mathsf{RD}}[k]\right) \tag{8}$$

and $\lambda$ in all $K$ optimization problems should be adjusted globally such that the constraint (3b) is satisfied at equality. Therefore, the optimal value of $\lambda$ depends on the channel statistics of the three links SD, SR and RD.

Based on the new defined notations, we can show that (8) can be written as

$$\max_{\zeta[k]\setminus\{R_{\mathsf{CF}}[k]\}} \quad \theta_{\mathsf{DT}}[k]\phi_{\mathsf{DT}}[k] + \theta_{\mathsf{DF}}[k]\phi_{\mathsf{DF}}[k] + \theta_{\mathsf{CF}}[k]\phi_{\mathsf{CF}}[k] + \theta_{\mathsf{RD}}[k]\phi_{\mathsf{RD}}[k] \tag{9}$$

where

$$\phi_{\mathsf{DT}}[k] = C_{\mathsf{SD}}[k] \tag{10a}$$

$$\phi_{\mathsf{DF}}[k] = R_{\mathsf{DF}}[k]/\theta_{\mathsf{DF}}[k] - \lambda R^*_{\mathsf{DF}}[k]/\theta_{\mathsf{DF}}[k] \tag{10b}$$

$$\phi_{\mathsf{CF}}[k] = \max_{R_{\mathsf{CF}}[k]/\theta_{\mathsf{CF}}[k]} \quad (R_{\mathsf{CF}}[k]/\theta_{\mathsf{CF}}[k] - \lambda R^*_{\mathsf{CF}}[k]/\theta_{\mathsf{CF}}[k]) \tag{10c}$$

$$\phi_{\mathsf{RD}}[k] = \lambda C_{\mathsf{RD}}[k] \tag{10d}$$

Consequently, the optimization of $R_{\mathsf{CF}}[k]/\theta_{\mathsf{CF}}[k]$ is independent of the optimal value of the channel access ratios. It depends on the value of $\lambda$, which is a global variable that is not a function of the instantaneous channel capacities in a given channel block $k$. This is valid for all three channel models under consideration.

*Theorem 2 (Optimal $R_{\mathsf{CF}}$ allocation):*

Given that $0 \leq \lambda \leq 1$, then the optimal $R_{\mathsf{CF}}[k]$ allocation is given by

$$\frac{R_{\mathsf{CF}}[k]}{\theta_{\mathsf{CF}}[k]} = \max\left(\log\left((1-\lambda)(1+\gamma_{\mathsf{SD}}[k]+\gamma_{\mathsf{SR}}[k])\right), C_{\mathsf{SD}}[k]\right) \tag{11}$$

The proof can be obtained by solving the optimization problem (10c). The solution steps are omitted for brevity. They are available in [20].

A direct consequence of Theorem 2 is that in all channel blocks $k$ that have $R_{\mathsf{CF}}[k] > \theta_{\mathsf{CF}}[k]C_{\mathsf{SD}}[k]$, we will have

$$\phi_{\mathsf{CF}}[k] = \log\left(1+\gamma_{\mathsf{SR}}[k]+\gamma_{\mathsf{SD}}[k]\right) + \lambda\log\left(\frac{1+\gamma_{\mathsf{SD}}[k]}{\gamma_{\mathsf{SR}}[k]}\right) + \lambda\log(\lambda) + (1-\lambda)\log(1-\lambda) \tag{12}$$

*Theorem 3 (Selecting transmission strategy):*

Given that $\lambda < 1$, the optimal solution will have only one transmission strategy (DF, CF or DT) selected per channel block $k$, and in channel model (a), either the source or the relay transmits and not both of them. The transmission strategy is selected according to

$$\xi[k] = \arg\max_{x} \quad \phi_x[k] \tag{13}$$

where $x \in \{\mathsf{DT}, \mathsf{DF}, \mathsf{CF}, \mathsf{RD}\}$ (for channel model (a)), or $x \in \{\mathsf{DT}, \mathsf{DF}, \mathsf{CF}\}$ (for channel models (b) and (c)). Thus, we get $\theta_x[k] = 1$ if $\xi[k] = x$, and $\theta_x[k] = 0$ if $\xi[k] \neq x$.

The proof is straightforward by solving (9). Notice that we assume that the channel gains are random variables with continuous probability distribution. Therefore, $\phi$ of each transmission strategy will also be random, and the probability that two different strategies maximize (13) in a given channel block is zero. Consequently, the solution of (9) is always unique when $\lambda < 1$.

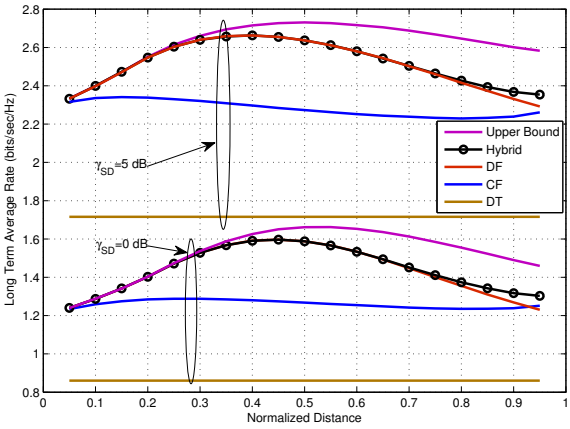



Fig. 2. Achievable rates for channel model (a).

We can also prove that we have strong duality since the time-sharing condition (refer to [23]) is satisfied in our problem. Furthermore, we can prove that, regardless of the channel statistics, the optimal value of $\lambda$ that maintains the constraint (3b) at equality will satisfy $0 \leq \lambda \leq 1$. At the special case when $\lambda = 1$, which happens when the SR link is very strong, the solution will not be unique since $\phi_{\mathsf{DT}}[k] = \phi_{\mathsf{DF}}[k]$ for all values of $k$ at which $\gamma_{\mathsf{SR}}[k] \geq \gamma_{\mathsf{SD}}[k]$. However, we can show that in this case, the optimal achievable rate will equal the capacity upper-bound. All of these proofs and extra details, including the characterization of upper bounds, can be found in [20].

## V. Numerical Results

We make our numerical results assuming that the distance between the source and the destination is $d_{\mathsf{SD}}$, and the relay is located on the straight line between the source and the destination such that the distance between the source and the relay is $d_{\mathsf{SR}}$, and the distance between the relay and the destination is $d_{\mathsf{RD}} = d_{\mathsf{SD}} - d_{\mathsf{SR}}$. The channels between the nodes are Rayleigh block-faded, and the average channel qualities are given by this formula

$$\bar{\gamma}_x = \epsilon\left(\frac{d_x}{d_{\mathsf{SD}}}\right)^{-\alpha}, \tag{14}$$

where $x \in \{\mathsf{SR}, \mathsf{RD}, \mathsf{SD}\}$, $\alpha = 3$ is the path loss exponent, and $\epsilon$ is a constant that is related to the transmission power, antenna gains and total distance. We use two cases in the simulation, $\epsilon = 10^{0.5} \approx 3.1623$, which gives $\bar{\gamma}_{\mathsf{SD}} = 5$ dB, and $\epsilon = 1$, which gives $\bar{\gamma}_{\mathsf{SD}} = 0$ dB.

In the simulations (Figs. 2,3,4), we plot the expected achievable rates versus the normalized distance of the relay to the source $\frac{d_{\mathsf{SR}}}{d_{\mathsf{SD}}}$. Also, we compare the optimal hybrid scheme to the upper-bounds and to sub-optimal schemes that use DF and DT without CF, or use CF and DT without DF.

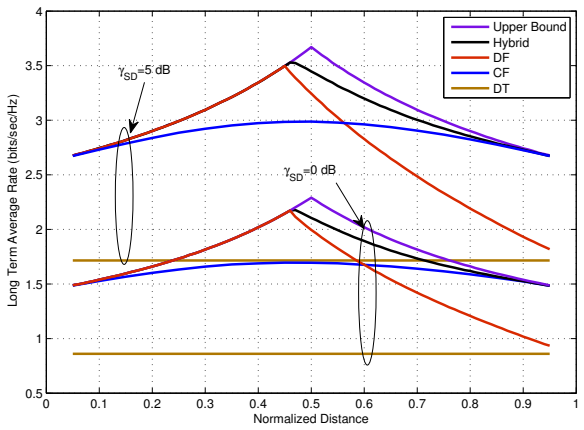



Fig. 3. Achievable rates for channel model (b).

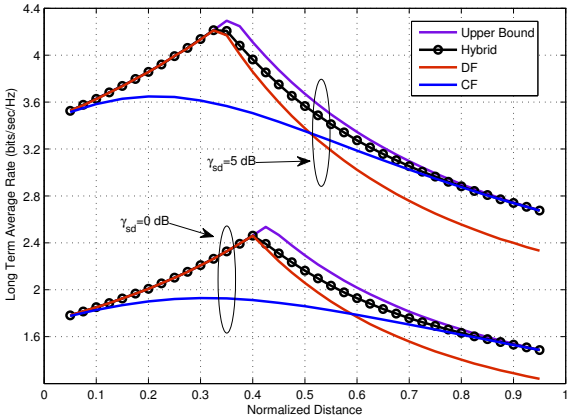



Fig. 4. Achievable rates for channel model (c).

## VI. Conclusions

We showed in this paper how to integrate compress-and-forward with decode-and-forward in buffer-aided relaying systems, and we have applied that to three different models of the relay channel. For optimality, only one transmission strategy is selected in a given channel block based on the channel conditions. The optimization of the data rate for compress-and-forward is obtained using a simple closed-form formula. The numerical results demonstrated the gains of the proposed scheme.

## Acknowledgment

This paper was made possible by NPRP grant # 5-401-2-161 from the Qatar National Research Fund (a member of Qatar Foundation). Furthermore, KAUST funded the efforts of A. Zafar partially and the efforts of M.-S. Alouini.

## References

[1] T. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sept. 1979.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[3] J. Nicholas Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[4] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, June 2005.

[5] Y. Liang and V. Veeravalli, "Gaussian orthogonal relay channels: Optimal resource allocation and capacity," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3284–3289, Sept. 2005.

[6] A. El Gamal, M. Mohseni, and S. Zahedi, "Bounds on capacity and minimum energy-per-bit for AWGN relay channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1545–1561, Apr. 2006.

[7] M. Cardone, D. Tuninetti, R. Knopp, and U. Salim, "On the Gaussian half-duplex relay channel," *IEEE Transactions on Information Theory*, vol. 57, no. 5, pp. 2542–2562, May 2014.

[8] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, Sept. 2005.

[9] D. Gunduz and E. Erkip, "Opportunistic cooperation by dynamic resource allocation," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1446–1454, Apr. 2007.

[10] Y. Liang, V. Veeravalli, and H. Vincent Poor, "Resource allocation for wireless fading relay channels: Max-min solution," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3432–3453, Oct. 2007.

[11] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 8, pp. 1530–1542, Aug. 2013.

[12] N. Zlatanov and R. Schober, "Buffer-aided relaying with adaptive link selection – fixed and mixed rate transmission," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2816–2840, May 2013.

[13] B. Xia, Y. Fan, J. Thompson, and H. Vincent Poor, "Buffering in a three-node relay network," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4492–4496, Nov. 2008.

[14] N. Zlatanov, R. Schober, and L. Lampe, "Buffer-aided relaying in a three node network," in *Proceedings IEEE International Symposium on Information Theory (ISIT)*, Cambridge, MA, July 2012, pp. 781–785.

[15] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "Exploiting multi-user diversity and multi-hop diversity in dual-hop broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3314–3325, July 2013.

[16] M. Shaqfeh, A. Zafar, H. Alnuweiri, and M. S. Alouini, "Joint opportunistic scheduling and network coding for bidirectional relay channel," in *Proceedings IEEE International Symposium on Information Theory (ISIT)*, Istanbul, Turkey, July 2013, pp. 1327–1331.

[17] H. Liu, P. Popovski, E. Carvalho, and Y. Zhao, "Sum-rate optimization in a two-way relay network with buffering," *IEEE Communications Letters*, vol. 17, no. 1, pp. 95–98, Jan. 2013.

[18] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "Resource allocation for two source-destination pairs sharing a single relay with a buffer," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1444–1457, May 2014.

[19] M. Shaqfeh, A. Zafar, H. Alnuweiri, and M.-S. Alouini, "Overlay cognitive radios with channel-aware adaptive link selection and buffer-aided relaying," *IEEE Transactions on Communications*, vol. 63, no. 8, pp. 2810–2822, Aug. 2015.

[20] M. Shaqfeh, A. Zafar, H. Alnuweiri, and M.-S. Alouini, "Maximizing expected achievable rates for block-fading buffer-aided relay channels," *available online at http://people.qatar.tamu.edu/mohammad.shaqfeh/Relay.pdf*.

[21] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, July 1973.

[22] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[23] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, July 2006.

# Extended Delivery Time Analysis for Opportunistic Secondary Packet Transmission without Work-preserving

Muneer Usman and Hong-Chuan Yang
Dept. of Electrical and Computer Engineering
University of Victoria
Victoria, BC V8P 5C2, Canada

Mohamed-Slim Alouini
Computer, Electrical and Mathematical Science
and Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)
Thuwal, Makkah Province, Saudi Arabia

***Abstract*—Cognitive radio transceiver can opportunistically access the underutilized spectrum resource of primary systems for new wireless services. With interweave implementation, secondary packet transmission may be interrupted by the primary user's transmission. To complement previous work on the resulting extended delivery time (EDT) [1], we consider secondary packet transmission with non-work-preserving strategy, i.e. interrupted packets will be re-transmitted. Both continuous sensing and periodic sensing cases are considered, for which EDT distributions are derived. Selected numerical and simulation results are presented to verify the mathematical formulation.**

***Index Terms*—Cognitive radio, opportunistic spectrum access, packet delivery time, queuing analysis.**

## I. Introduction

Radio spectrum scarcity is one of the most serious problems nowadays faced by the wireless communications industry. Cognitive radio is a promising solution to this emerging problem by exploiting temporal/spatial spectrum opportunities over the existing licensed frequency bands [2], [3]. Different implementation strategies exist for opportunistic spectrum access (OSA). In underlay cognitive radio implementation, the primary and secondary users simultaneously access the same spectrum, with a constraint on the interference that secondary user (SU) may cause to primary transmission. With interweave implementation, the SU can access a primary user (PU) channel only when the channel is not used by PUs, and must vacate the occupied channel when the PU appears. The secondary transmission of a given amount of data may involve multiple transmission attempts, resulting in extra transmission delay. When the secondary transmission is interrupted by PU activities, the secondary system can adopt either non-work-preserving strategy, where interrupted packets transmission must be repeated [4], or work-preserving strategy, where the secondary transmission can continue from the point where it was interrupted, without wasting the previous transmission [5]. In our previous work [1], we carried out a thorough statistical analysis on the extended delivery time (EDT) [4] of secondary packet transmission with work-preserving strategy, and then applied these results to the secondary queuing analysis. Typically, work-preserving packet transmission requires packets to be coded with certain rateless codes such as fountain codes, which may not be available in the secondary system.

There has been a continuing interest in the delay and throughput analysis for secondary systems [6]–[11]. There has been little previous work on delay analysis with periodic spectrum sensing. [12] discusses periodic sensing, focusing on a single secondary transmission slot. Design of periodic sensing parameters has been discussed in [13], [14]. [8] derives expressions for average delay for continuous and periodic sensing. A framework of Markov decision processes is presented by [15] to derive the optimal policy for channel access under periodic sensing assumption.

In this paper, we investigate the statistical characteristics of the EDT with non-work-preserving strategy, and apply them to evaluate the delay performance of secondary transmission considering a high SNR regime. Analysis with non-work-preserving strategy is, in general, more challenging as the transmission of a secondary packet will involve an interweaved sequence of wasted transmission slots and waiting time slots, both of which can have random time duration, followed by the final successful transmission slot. Note that with work-preserving strategy, these are no wasted transmission slots. To the best of our knowledge, the complete statistics of the EDT for non-work-preserving strategy has not been investigated in literature. In this work, we first derive the exact expressions for the distribution function of EDT assuming a fixed packet transmission time in terms of moment generating function (MGF) and probability density function (PDF). Two spectrum sensing scenarios are considered – continuous sensing and periodic sensing. The generalization to random packet transmission time, due to the effect of fading wireless channel and noise, can be addressed in a similar manner as in [1], and is omitted here due to space limitation. We then apply the results on EDT to the secondary queuing analysis. Numerical and simulation results are presented to verify the analytical approach and illustrate secondary queuing performance.

The rest of this paper is organized as follows. In section II, we introduce the system model and the problem formulation. In section III, we analyze the EDT of a single packet for both continuous sensing case and periodic sensing case. In section

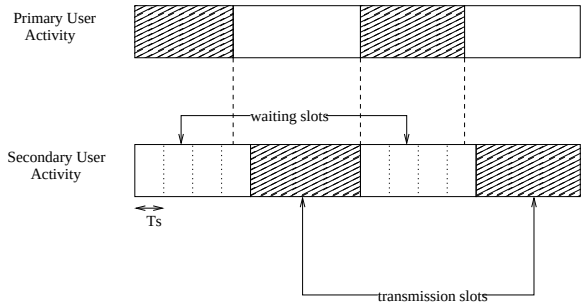



Fig. 1. Illustration of PU and SU activities and SU sensing for periodic sensing case.

IV, we present the average queuing delay of the secondary system in a general M/G/1 queuing set-up. Finally, this paper is concluded in section V.

## II. System Model and Problem Formulation

We consider the opportunistic access of a single channel of the primary system. The PU occupies that channel according to a homogeneous continuous-time Markov chain with an average on-time of $\lambda$ and an average off-time of $\mu$. The SU opportunistically accesses the channel for data transmission. Specifically, the SU can use the channel only after PU stops transmission. As soon as the PU restarts transmission, the SU instantaneously stops its transmission.

The SU monitors PU activity through spectrum sensing[1]. With continuous sensing, the SU senses the channel for availability with a very small sensing period. Thus, the SU can start its transmission soon after the channel becomes available, with a negligible time delay. With periodic sensing, the SU senses the channel with an interval of $T_s$. On each sensing instance, if the PU is sensed busy, the SU will wait for $T_s$ time period and re-sense the channel. With periodic sensing, there is a small amount of time when the PU has stopped its transmission, but the SU has not yet acquired the channel, as illustrated in Fig. 1. During transmission, the SU continuously monitors PU activity and discontinues its transmission as soon as the PU restarts. The continuous period of time during which the PU is transmitting is referred to as a waiting slot. Similarly, the continuous period of time during which the PU is off and the SU is transmitting is referred to as a transmission slot. For periodic sensing case, the waiting slot also includes the time duration when the PU has stopped transmission, but the SU has not sensed the channel yet.

In this work, we analyze the packet delivery time of secondary system, which comprises of an interleaved sequence of the wasted transmission times and the waiting times, followed by a successful transmission time. Note that a transmission slot is wasted if its duration is less than the time required to transmit the packet. The resulting EDT for a packet is mathematically given by $T_{ED} = T_w + T_{tr}$, where $T_w$ is the total of the waiting time and wasted transmission times for the SU, and $T_{tr}$ is the packet transmission time. In what follows, we first derive the distribution of the EDT $T_{ED}$ for continuous sensing and periodic sensing cases, which are then applied to the secondary queuing analysis in section IV.

[1]We assume perfect spectrum sensing here. The effect of imperfect sensing [16], [17] will be considered in future work.

## III. Extended Delivery Time Analysis

### A. Continuous Sensing

The EDT for packet transmission by the SU consists of interweaved waiting slots and wasted transmission slots, followed by the final successful transmission slot of duration $T_{tr}$. The distribution of waiting time $T_w$ depends on whether the PU was on or off at the instant of packet arrival. We denote the PDF of the waiting time of the SU for the case when PU is on at the instant of packet arrival, and for the case when PU is off at that instant, by $f^{(c)}_{T_w,p_{on}}(t)$ and $f^{(c)}_{T_w,p_{off}}(t)$, respectively. The PDF of the EDT $T_{ED}$ for the SU is then given by

$$f^{(c)}_{T_{ED}}(t) = \frac{\lambda}{\lambda+\mu} f^{(c)}_{T_w,p_{on}}(t-T_{tr}) + \frac{\mu}{\lambda+\mu} f^{(c)}_{T_w,p_{off}}(t-T_{tr}), \tag{1}$$

where $\frac{\lambda}{\lambda+\mu}$ and $\frac{\mu}{\lambda+\mu}$ are the stationery probabilities that the PU is on or off at the instant of packet arrival, respectively. The two probability density functions $f_{T_w,p_{on}}(t)$ and $f_{T_w,p_{off}}(t)$ above are calculated independently as follows.

Let $\mathcal{P}_k$ be the probability that the SU was successful in sending the packet in the $k^{th}$ transmission slot. This means that each of the first $(k-1)$ slots had a time duration of less than $T_{tr}$, while the $k^{th}$ transmission slot had a duration more than $T_{tr}$. Thus, $\mathcal{P}_k$ can be calculated, while noting that the duration of secondary transmission slots is exponentially distributed with mean $\mu$, as

$$\mathcal{P}_k = e^{-\frac{T_{tr}}{\mu}} \cdot \left(1 - e^{-\frac{T_{tr}}{\mu}}\right)^{k-1}. \tag{2}$$

For the case when PU is off at the instant of packet arrival, if a certain packet is transmitted completely in the $k^{th}$ transmission slot, then the total wait time for that packet includes $(k-1)$ secondary waiting slots and $(k-1)$ wasted transmission slots. Note that the duration of each of these $(k-1)$ waiting slots, denoted by the random variable $T_{wait}$, which is equal to PU on time, follows an exponential distribution with PDF given by $f^{(c)}_{T_{wait}}(t) = \frac{1}{\lambda} e^{\frac{-t}{\lambda}} u(t)$, while the duration of each of the previous $(k-1)$ wasted secondary transmission slots, denoted by the random variable $T_{waste}$, follows a truncated exponential distribution, with PDF given by

$$f_{T_{waste}}(t) = \frac{1}{1 - e^{-\frac{T_{tr}}{\mu}}} \frac{1}{\mu} e^{\frac{-t}{\mu}} \cdot (u(t) - u(t-T_{tr})), \tag{3}$$

where $u(t)$ is the unit step function. The MGF of $T_{w,p_{off}}$ for the continuous sensing case, $\mathcal{M}^{(c)}_{T_w,p_{off}}(s)$ can be calculated as

$$\mathcal{M}^{(c)}_{T_w,p_{off}}(s) = \sum_{k=1}^{\infty} \mathcal{P}_k \cdot \left(\mathcal{M}^{(c)}_{T_{wait}}(s)\right)^{k-1} \cdot \left(\mathcal{M}_{T_{waste}}(s)\right)^{k-1}, \tag{4}$$

where $\mathcal{M}^{(c)}_{T_{wait}}(s)$ is the MGF of $T_{wait}$ for the continuous sensing case, given by

$$\mathcal{M}^{(c)}_{T_{wait}}(s) = \frac{1}{1-\lambda s}, \tag{5}$$

and $\mathcal{M}_{T_{waste}}(s)$ is the MGF of $T_{waste}$, given by

$$\mathcal{M}_{T_{waste}}(s) = \frac{1-e^{T_{tr}(s-\frac{1}{\mu})}}{(1-\mu s)(1-e^{-\frac{T_{tr}}{\mu}})}. \tag{6}$$

After substituting Eqs. (2), (5), and (6) into Eq. (4), performing some manipulation, using the following general formula for partial fractions,

$$\frac{1}{[x(x-a)]^n} = \sum_{j=0}^{n-1}(-1)^n \binom{2n-j-2}{n-1} \frac{1}{a^{2n-j-1}} \times \left[\frac{1}{x^{j+1}} + \frac{(-1)^{j+1}}{(x-a)^{j+1}}\right], \tag{7}$$

the proof of which is given in the appendix of [18], and taking the inverse MGF, we obtain Eq. (8) as the PDF of $T_{w,p_{off}}$ for continuous sensing case, where ${}_1F_1(.,.,.)$ is the generalized Hypergeometric function [19]. Note that the impulse corresponds to the case that the packet is transmitted without waiting. Further simplification of Eq. (8) is not evident.

For the case when PU is on at the instant of packet arrival, the MGF of $T_{w,p_{on}}$ for the continuous sensing case $\mathcal{M}^{(c)}_{T_{w,p_{on}}}(s)$ can be similarly calculated as

$$\mathcal{M}^{(c)}_{T_{w,p_{on}}}(s) = \sum_{k=1}^{\infty} \mathcal{P}_k \cdot \left(\mathcal{M}^{(c)}_{T_{wait}}(s)\right)^k \cdot \left(\mathcal{M}_{T_{waste}}(s)\right)^{k-1}. \tag{9}$$

Using similar manipulations used for the PDF of $T_{w,p_{off}}$, the PDF of $T_{w,p_{on}}$ is obtained after simplification as shown in Eq. (10).

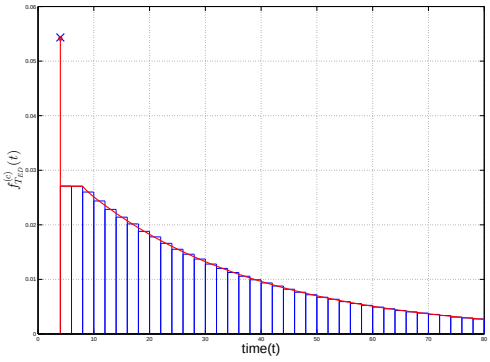



Fig. 2. Simulation verification for the analytical PDF of $T_{ED}$ with continuous sensing ($T_{tr} = 4$ ms, $\lambda = 3$ ms, and $\mu = 2$ ms).

Fig. 2 plots the analytical expression for the PDF of the EDT with continuous sensing as given in Eq. (1). The corresponding plot for the simulation results is also shown. The perfect match between analytical and simulation results verify our analytical approach.

### B. Periodic Sensing

For the periodic sensing case, the PDF of the EDT $T_{ED}$ for the SU packet transmission is given by

$$f^{(p)}_{T_{ED}}(t) = \frac{\lambda}{\lambda+\mu} f^{(p)}_{T_w,p_{on}}(t-T_{tr}) + \frac{\mu}{\lambda+\mu} f^{(p)}_{T_w,p_{off}}(t-T_{tr}), \tag{11}$$

where $f^{(p)}_{T_w,p_{on}}(t)$ and $f^{(p)}_{T_w,p_{off}}(t)$ denote the PDFs of the waiting time of the SU with periodic sensing, for the case when PU is on at the instant of packet arrival, and for the case when PU is off at that instant, respectively. We again derive the PDF of waiting time through MGF approach. The MGF of $T_{w,p_{off}}$ for the periodic sensing case, $\mathcal{M}^{(p)}_{T_w,p_{off}}(s)$, can be calculated as

$$\mathcal{M}^{(p)}_{T_w,p_{off}}(s) = \sum_{k=1}^{\infty} \mathcal{P}_k \cdot \left(\mathcal{M}^{(p)}_{T_{wait}}(s)\right)^{k-1} \cdot \left(\mathcal{M}_{T_{waste}}(s)\right)^{k-1}, \tag{12}$$

where $\mathcal{P}_k$ is given in Eq. (2), $\mathcal{M}_{T_{waste}}(s)$ is the MGF of the time duration of a wasted transmission slot $T_{waste}$, which is, noting that the PDF of $T_{waste}$ remains the same as given in Eq. (3) due to the memoryless property of exponential distribution, given in Eq. (6), and $\mathcal{M}^{(p)}_{T_{wait}}(s)$ denotes the MGF of the wait time in a single waiting slot. With periodic sensing, $T_{wait}$ consists of multiple $T_s$, and follows a geometric distribution. The MGF can be obtained as

$$\mathcal{M}^{(p)}_{T_{wait}}(s) = \sum_{n=1}^{\infty}(1-\beta)\beta^{n-1}e^{nsT_s}, \tag{13}$$

where $\beta$ denotes the probability that the primary user is on at a given sensing instant provided that it was on at the previous sensing instant $T_s$ time units earlier, given by $\beta = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\frac{1}{\lambda}+\frac{1}{\mu})T_s}$ [1]. Note that $\beta$ is a constant again due to the memoryless property of exponential distribution. Substituting Eqs. (2), (6), and (13) into Eq. (12), performing some manipulation, and taking the inverse MGF, we obtain the expression for $f^{(p)}_{T_w,p_{off}}(t)$ as shown in Eq. (14).

For the case when PU is on at the instant of packet arrival, the PDF of $T_{w,p_{on}}$ for the periodic sensing case can be similarly calculated as shown in Eq. (15). Note that the sequence of impulses corresponds to the case that the packet is transmitted in the first transmission attempt on acquiring the channel after a random number of sensing intervals/attempts.

Fig. 3 plots the cumulative distribution function (CDF) of the EDT with periodic sensing, $F^{(p)}_{T_{ED}}(t)$, obtained by numerical integration of the analytical PDF expression given by Eq. (11). The corresponding plot for the simulation results is also shown. The perfect match between analytical and simulation results verify our analytical approach.

## IV. Application to Secondary Queuing Analysis

In this section, we consider the average transmission delay for the secondary system in a queuing set-up as an application of the analytical results in previous section. In particular, the secondary traffic intensity is high and, as such, a first-in-first-out queue is introduced to hold packets until transmission.

$$f^{(c)}_{T_w,p_{off}}(t) = e^{-\frac{T_{tr}}{\mu}}\delta(t) + \frac{e^{-\frac{T_{tr}}{\mu}}}{\lambda+\mu}(1-e^{-\alpha t})u(t) - \frac{e^{-\frac{2T_{tr}}{\mu}}}{\lambda+\mu}(1-e^{-\alpha(t-T_{tr})})u(t-T_{tr}) + \sum_{i=1}^{\infty}\frac{(\lambda\mu)^i}{(\lambda+\mu)^{2i+1}}\binom{2i}{i}$$
$$\times\Big[{}_1F_1\left(-i;-2i;-\alpha(t-iT_{tr})\right)e^{-(i+1)\frac{T_{tr}}{\mu}}u(t-iT_{tr}) - {}_1F_1\left(-i;-2i;-\alpha(t-(i+1)T_{tr})\right)e^{-(i+2)\frac{T_{tr}}{\mu}}u(t-(i+1)T_{tr})$$
$$-{}_1F_1\left(-i;-2i;\alpha(t-iT_{tr})\right)e^{-\alpha t}e^{-(i+1)\frac{T_{tr}}{\mu}}u(t-iT_{tr}) + {}_1F_1\left(-i;-2i;\alpha(t-(i+1)T_{tr})\right)e^{-\alpha t}e^{-(i+2)\frac{T_{tr}}{\mu}}u(t-(i+1)T_{tr})\Big], \quad (8)$$

$$f^{(c)}_{T_w,p_{on}}(t) = \frac{e^{-\frac{T_{tr}}{\mu}}}{\lambda+\mu}\left(1+\frac{\mu}{\lambda}e^{-\alpha t}\right)u(t) + \sum_{i=1}^{\infty}\frac{(\lambda\mu)^i}{(\lambda+\mu)^{2i+1}}e^{-(i+1)\frac{T_{tr}}{\mu}}\Big[\binom{2i}{i}{}_1F_1\left(-i;-2i;-\alpha(t-iT_{tr})\right)\cdot u(t-iT_{tr})$$
$$-\binom{2i}{i}\frac{\mu}{\lambda}e^{-\alpha(t-iT_{tr})}{}_1F_1\left(-i;-2i;\alpha(t-iT_{tr})\right)\cdot u(t-iT_{tr}) - \binom{2i-1}{i}\left(1+\frac{\mu}{\lambda}\right){}_1F_1\left(1-i;1-2i;-\alpha(t-iT_{tr})\right)\cdot u(t-iT_{tr})$$
$$+\binom{2i-1}{i}\left(1+\frac{\mu}{\lambda}\right)e^{-\alpha(t-iT_{tr})}{}_1F_1\left(1-i;1-2i;\alpha(t-iT_{tr})\right)\cdot u(t-iT_{tr})\Big]. \quad (10)$$

$$f^{(p)}_{T_w,p_{off}}(t) = e^{-\frac{T_{tr}}{\mu}}\delta(t) + \sum_{n=1}^{\infty}\left[\frac{(1-\beta)\beta^{n-1}}{\mu}e^{-\frac{(t-nT_s)}{\mu}}e^{-\frac{T_{tr}}{\mu}}{}_1F_1\left(1-n;1;-\frac{1-\beta}{\beta}\frac{t-nT_s}{\mu}\right)\right.$$
$$\left.+\sum_{i=1}^{n}\left[(-1)^i e^{-(i+1)\frac{T_{tr}}{\mu}}\binom{n-1}{i-1}\frac{1}{(i-1)!}\frac{(t-iT_{tr}-nT_s)^{i-1}}{\mu^i}(1-\beta)^i\beta^{n-i}e^{\frac{-(t-nT_s-iT_{tr})}{\mu}}\times{}_2F_2\left(i+1,i-n;i,i;-\frac{1-\beta}{\beta}\frac{(t-nT_s-iT_{tr})}{\mu}\right)\right]\right]. \quad (14)$$

$$f^{(p)}_{T_w,p_{on}}(t) = e^{-\frac{T_{tr}}{\mu}}\sum_{n=2}^{\infty}(n-1)\frac{(1-\beta)^2\beta^{n-2}}{\mu}e^{-\frac{t-nT_s}{\mu}}{}_1F_1\left(2-n;2;-\frac{1-\beta}{\beta}\cdot\frac{t-nT_s}{\mu}\right) + e^{-\frac{T_{tr}}{\mu}}\sum_{n=1}^{\infty}(1-\beta)\beta^{n-1}\delta(t-nT_s)$$
$$+\sum_{n=1}^{\infty}\sum_{i=1}^{n-1}\left[(-1)^i e^{-(i+1)\frac{T_{tr}}{\mu}}\binom{n-1}{i}(1-\beta)^{i+1}\beta^{n-i-1}\times\frac{t^{i-1}e^{-\frac{t-nT_s-iT_{tr}}{\mu}}}{(i-1)!\mu^i}{}_1F_1\left(i+1-n;i;-\frac{1-\beta}{\beta}\cdot\frac{t-nT_s-iT_{tr}}{\mu}\right)\right]. \quad (15)$$

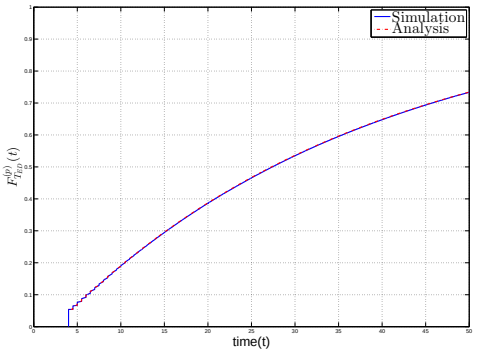



Fig. 3. Simulation verification for the analytical CDF of $T_{ED}$ with periodic sensing ($T_{tr} = 4$ ms, $\lambda = 3$ ms, $\mu = 2$ ms, and $T_s = 0.5$ ms).

We assume that equal-sized packet arrival follows a Poisson process with intensity $\frac{1}{\psi}$ [1]. For the sake of simplicity, transmission time $T_{tr}$ of all packets is assumed to be a fixed constant. As such, the secondary packet transmission can be modelled as a general M/G/1 queue, where the service time is closely related to the EDT studied in the previous section. Following the similar approach in Sec. IV-B of [1], we can analyze the queuing delay for non-work-preserving strategy. We focus on the periodic sensing case in the following while noting the analysis for the continuous sensing scenarios can be similarly solved. Note the service time of a packet depends on whether the PU is on or off when the packet is available for transmission. To facilitate queuing analysis, we now calculate the first and second moments of the service time for these two types of packets.

First moments. The service time for packets seeing PU off is denoted by $ST_{p_{off}}$. Noting that $ST_{p_{off}} = T_{w,p_{off}} + T_{tr}$, due to the memoryless property of the non-work-preserving strategy, we can calculate its first moment $E[ST_{p_{off}}]$ by following the conditional expectation approach as

$$E[ST_{p_{off}}] = e^{-\frac{T_{tr}}{\mu}}\cdot T_{tr} + (1-e^{-\frac{T_{tr}}{\mu}})\cdot E[(T_{waste}+T_{wait}+ST_{p_{off}})]. \quad (16)$$

Here the first addition term corresponds to the case that the complete packet is successfully transmitted in the first transmission slot, and the second addition term refers to the case when the complete packet is not successfully transmitted. For periodic sensing, it can be shown, from Eqs. (6) and (13), that $E[T_{wait}] = \frac{T_s}{1-\beta}$ and $E[T_{waste}] = \mu - T_{tr}\frac{e^{-\frac{T_{tr}}{\mu}}}{1-e^{-\frac{T_{tr}}{\mu}}}$. The first moment of $ST_{p_{off}}$ can be calculated from Eq. (16) as

$$E[ST_{p_{off}}] = \left(e^{\frac{T_{tr}}{\mu}}-1\right)\left(\mu+\frac{T_s}{1-\beta}\right). \quad (17)$$

Since the case with PU on at the instant of packet arrival is precisely the same as the case of PU off at the instant of packet arrival preceded by a waiting slot, we can calculate $E[ST_{p_{on}}]$, the first moment of the service time for packets seeing PU on, as

$$E[ST_{p_{on}}] = E[ST_{p_{off}}] + \frac{T_s}{1-\beta} = \left(e^{\frac{T_{tr}}{\mu}}-1\right)\cdot\mu + e^{\frac{T_{tr}}{\mu}}\cdot\frac{T_s}{1-\beta}. \quad (18)$$

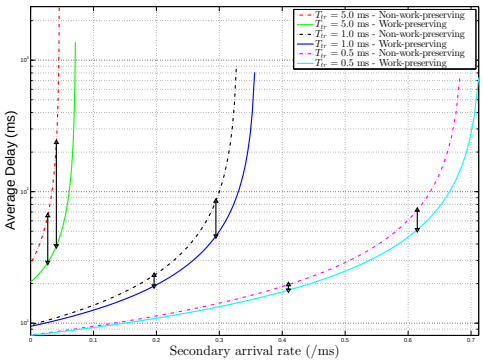



Fig. 4. Average queuing delay with periodic sensing ($T_s = 0.5$ ms, $\lambda = 10$ ms, and $\mu = 6$ ms)

Second moments. It can be shown from Eqs. (6) and (13) that $E[T_{wait}^2] = T_s^2 \frac{1+\beta}{(1-\beta)^2}$ and $E[T_{waste}^2] = 2\mu^2 + \frac{e^{-\frac{T_{tr}}{\mu}}}{1-e^{-\frac{T_{tr}}{\mu}}}(-T_{tr}^2 - 2\mu T_{tr})$. Using a similar equation as Eq. (16) for second moment, and simplifying, we obtain

$$E[ST_{p_{off}}^2] = e^{\frac{T_{tr}}{\mu}}\left[-2T_{tr}\frac{T_s}{1-\beta} - 2\mu T_{tr}\right] + \left(e^{\frac{T_{tr}}{\mu}} - 1\right)\left(e^{\frac{T_{tr}}{\mu}}\left[2\mu\frac{T_s}{1-\beta} + 2\mu^2\right] + T_s^2\frac{1+\beta}{(1-\beta)^2}\right) + \left(e^{\frac{T_{tr}}{\mu}} - 1\right)^2\left[2\frac{T_s^2}{(1-\beta)^2} + 2\mu\frac{T_s}{1-\beta}\right]. \quad (19)$$

Similarly, the second moment of the service time for packets seeing PU on can be defined as $E[ST_{p_{on}}^2] = E[(T_{wait} + ST_{p_{off}})^2]$, and calculated as

$$E[ST_{p_{on}}^2] = e^{\frac{T_{tr}}{\mu}}\left[-2T_{tr}\frac{T_s}{1-\beta} - 2\mu T_{tr} + T_s^2\frac{1+\beta}{(1-\beta)^2}\right] + 2\left(e^{\frac{T_{tr}}{\mu}} - 1\right)e^{\frac{T_{tr}}{\mu}}\left(\mu + \frac{T_s}{1-\beta}\right)^2. \quad (20)$$

Finally, these moments can be substituted into Eqs. (45)-(55) of [1] to calculate the average queuing delay.

Fig. 4 shows the average delay including the queuing delay against the rate of arrival of data packets, for various values of $T_{tr}$, both for work-preserving and non-work-preserving strategies. It can be seen that as expected, work-preserving strategy always performs better than non-work-preserving strategy. Also, the performance difference between the two strategies reduces as the packet transmission time $T_{tr}$ decreases, as shown by the vertical lines in the figure.

## V. Conclusion

This paper studied the extended delivery time of a data packet appearing at the secondary user in an interweave cognitive setup assuming non-work-preserving strategy. Exact analytical results for the probability distribution of the EDT for a fixed-size data packet were obtained for continuous sensing and periodic sensing cases. These results were then applied to analyze the expected delay of a packet at SU in a queuing setup. Simulation results were presented to verify the analytical results.

## References


[1] M. Usman, H.-C. Yang, and M.-S. Alouini, "Extended delivery time analysis for cognitive packet transmission with application to secondary queuing analysis." Accepted in IEEE Trans. Wireless Commun., available at http://arxiv.org/abs/1409.1628.

[2] A. Goldsmith, S. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. of the IEEE*, vol. 97, pp. 894–914, May 2009.

[3] Q. Zhao, S. Geirhofer, L. Tong, and B. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Trans. Signal Process.*, vol. 56, pp. 785–796, Feb 2008.

[4] F. Borgonovo, M. Cesana, and L. Fratta, "Throughput and delay bounds for cognitive transmissions," in *Advances in Ad Hoc Networking* (P. Cuenca, C. Guerrero, R. Puigjaner, and B. Serra, eds.), vol. 265 of *IFIP International Federation for Information Processing*, pp. 179–190, Springer US, 2008.

[5] C.-W. Wang and L.-C. Wang, "Analysis of reactive spectrum handoff in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, pp. 2016–2028, November 2012.

[6] F. Khan, K. Tourki, M.-S. Alouini, and K. Qaraqe, "Delay performance of a broadcast spectrum sharing network in Nakagami-m fading," *IEEE Trans. Veh. Technol.*, vol. 63, pp. 1350–1364, March 2014.

[7] Z. Liang and D. Zhao, "Quality of service performance of a cognitive radio sensor network," in *Proc. IEEE Int. Conf. Commun. (ICC), 2010*, pp. 1–5, May 2010.

[8] Z. Liang, S. Feng, D. Zhao, and X. Shen, "Delay performance analysis for supporting real-time traffic in a cognitive radio sensor network," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 325–335, January 2011.

[9] S. Wang and J. Zhang, "Opportunistic spectrum scheduling by jointly exploiting channel correlation and pu traffic memory," *Selected Areas in Communications, IEEE Journal on*, vol. 31, pp. 394–405, March 2013.

[10] L. Sibomana, H.-J. Zepernick, H. Tran, and C. Kabiri, "Packet transmission time for cognitive radio networks considering interference from primary user," in *9th Int. Wireless Commun. and Mobile Computing Conf. (IWCMC), 2013*, pp. 791–796, July 2013.

[11] H. Tran, T. Duong, and H.-J. Zepernick, "Delay performance of cognitive radio networks for point-to-point and point-to-multipoint communications," *EURASIP J. on Wireless Commun. and Networking*, vol. 2012, no. 1, pp. 1–15, 2012.

[12] F. Gaaloul, H.-C. Yang, R. Radaydeh, and M.-S. Alouini, "Switch based opportunistic spectrum access for general primary user traffic model," *IEEE Wireless Commun. Lett.*, vol. 1, pp. 424–427, October 2012.

[13] Q. Liu, X. Wang, and Y. Cui, "Robust and adaptive scheduling of sequential periodic sensing for cognitive radios," *IEEE J. Sel. Areas Commun.*, vol. 32, pp. 503–515, March 2014.

[14] A. Mariani, S. Kandeepan, and A. Giorgetti, "Periodic spectrum sensing with non-continuous primary user transmissions," *IEEE Trans. Wireless Commun.*, vol. 14, pp. 1636–1649, March 2015.

[15] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A pomdp framework," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 589–600, April 2007.

[16] E. Peh, Y.-C. Liang, Y. L. Guan, and Y. Zeng, "Optimization of cooperative sensing in cognitive radio networks: A sensing-throughput tradeoff view," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 5294–5299, Nov 2009.

[17] G. Noh, J. Lee, H. Wang, S. Kim, S. Choi, and D. Hong, "Throughput analysis and optimization of sensing-based cognitive radio systems with markovian traffic," *IEEE Trans. Veh. Technol.*, vol. 59, pp. 4163–4169, Oct 2010.

[18] M. Usman, H.-C. Yang, and M.-S. Alouini, "Extended delivery time analysis for non-work-preserving packet transmission in cognitive environment," *ArXiv e-prints*, Sept. 2014. http://arxiv.org/abs/1409.0911.

[19] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,*. Dover Publications, Incorporated, 1974.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Beamforming towards regions of interest for multi-site mobile networks

Paul Hurley*, Matthieu Simeoni*†
*IBM Zurich Research Laboratory, CH-8803 Rüschlikon, Switzerland
†École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
Email: pah@zurich.ibm.com, meo@zurich.ibm.com

***Abstract*—We show how a beamforming technique for analytical spatial filtering, called flexibeam, can be applied to mobile phone mast broadcasting so as to result in concentrations of power where most devices are.**

**To that end, flexibeam is interpreted as transmission beamforming. An analytically described radiation pattern is extended from a sphere to Euclidean space. A continuous beamforming function is then obtained by the Fourier transform of the extended radiation pattern. We then show how a Gaussian filter can be approximately achieved using beamforming.**

**The method is then expanded by means of an example of a collection of mobile phone masts covering an area of Zurich city so as to concentrate energy where devices are concentrated.**

## I. Introduction

Beamforming has been deployed in mobile phone standards starting already with 2G. The techniques have become ever more sophisticated with each iteration. 4G/LTE, for example, deploys MIMO-based beamforming.

In general, direct multi-user beamforming [1], [2] – creating beams for each individual devices from mobile phone masts – is a gargantuan task. There are simply too many users, and a steady stream of accurate channel feedback [3] would be required to account people and vehicles moving around.

Yet algorithms deployed to date are variations on a theme within the MIMO framework: steering the beam towards a single point [4], [5], [6]. In practise, one would like to be able to concentrate energy in a way commensurate with the locations of devices, namely target areas of interest not single points, all the while building in tolerance for movement and imprecise location information.

To increase received power, LTE devices can receive signal from multiple base stations. This is a complex coordination protocol in general, and one key component to its efficiency is to have base stations target areas of importance. In this paper, we apply a technique called flexibeam, which determines beamforming weights that when applied approximate an optimal radiation pattern, so as to enable base stations to jointly concentrate energy where most devices are. The analytic framework allows tractable, numerically stable determination of beamforming weights. Aiming for areas rather than points minimises the required update rate, and reduces the communication requirement.

To this end, Section II derives flexibeam from an explicit transmit beamforming perspective. We then, in Section III, illustrate its application using a Gaussian approximation to track an object in the presence of uncertainty. Afterwards, in Section III, we illustrate how energy can be concentrated around a certain area. There we take an example of devices concentrated around an area of Zurich city, and show how the target radiation pattern can be approximated by a series of Gaussian filters. We then illustrate how to determine each base station's beamforming weights so as to concentrate energy where most devices are.

## II. Flexibeam from the transmit perspective

In [7], we derived the receiving case for flexibeam. As a consequence of the reciprocity theorem [8], beam-shapes so designed can be used to receive or transmit. However, to gain insight into its operation and application, we now derive the transmission case directly.

Consider an array of $L$ omni-directional receiving antennas, with unit gains and positions $\boldsymbol{p}_1, \dots, \boldsymbol{p}_L \in \mathbb{R}^n$. Each antenna emit an identical *narrow-band* signal $s(t) \in \mathbb{C}$. Without loss of generality, let the wavelength of this signal be $\lambda = 1$. The signals originating from each antenna will sum coherently, producing a radiation pattern, also called *beam-shape* of the antenna array. To control this radiation pattern, different delays and gains are introduced at each antenna:

$$x_i(t) = \gamma_i e^{j\phi_i} s(t), \tag{1}$$

where $\gamma_i > 0$ and $\phi_i \in [0, 2\pi]$ are respectively the gain and phase delay for antenna $i$. The signal seen at a *far field* target with position $\boldsymbol{r} \in \mathbb{S}^{n-1}$ is given by [8], [9]

$$\begin{aligned} y(t, \boldsymbol{r}) &= s(t) \left( \sum_{i=1}^{L} \gamma_i e^{j\phi_i} e^{-j2\pi \langle \boldsymbol{r}, \boldsymbol{p}_i \rangle} \right), \\ &= s(t) \left( \sum_{i=1}^{L} w_i^* e^{-j2\pi \langle \boldsymbol{r}, \boldsymbol{p}_i \rangle} \right), \\ &= s(t) b^*(\boldsymbol{r}), \end{aligned} \tag{2}$$

where $b(\boldsymbol{r}) = \sum_{i=1}^{L} w_i e^{j2\pi \langle \boldsymbol{r}, \boldsymbol{p}_i \rangle}$ is the array *beam-shape*, and $w_i = \gamma_i e^{-j\phi_i} \in \mathbb{C}$ are the *beamforming weights*. We observe that beamforming is here the result of the physical summation of the signals emitted by each antenna.

For matched beamforming (cf. Fig. 1), the beamforming weights are chosen by

$$w_i = e^{-j2\pi \langle \boldsymbol{r}_0, \boldsymbol{p}_i \rangle}, \quad i = 1, \dots, L,$$

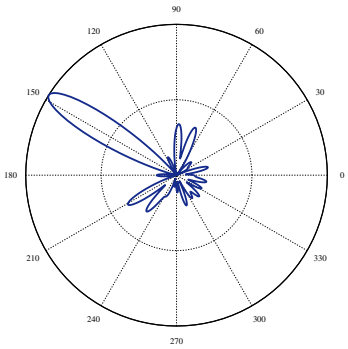



Fig. 1: Example beam-shape obtained with matched beamforming.

where $\boldsymbol{r}_0 \in \mathbb{S}^{n-1}$ is the *steering direction*. Thus, the gains and delays at each antenna are respectively $\gamma_i = 1$ and $\phi_i = 2\pi\langle \boldsymbol{r}_0, \boldsymbol{p}_i\rangle$.

Now consider a notional continuous field of antennas covering $\mathbb{R}^n$, over which we define a *broadcast function* $x(t, \boldsymbol{p}) \in \mathcal{L}^2(\mathbb{R}^n, \mathbb{C})$. This describes the signal that would be broadcasted by an antenna located at position $\boldsymbol{p}$, and extends (1) to cover all points in $\mathbb{R}^n$:

$$x(t, \boldsymbol{p}) = \gamma(\boldsymbol{p})e^{j\phi(\boldsymbol{p})}s(t) = w^*(\boldsymbol{p})s(t),$$

where $w \in \mathcal{L}^2(\mathbb{R}^n, \mathbb{C})$ is the *beamforming function*, that generalises the concept of beamforming weights, and describes the gains and delays to be applied at each position $\boldsymbol{p} \in \mathbb{R}^n$. The signals emitted by this continuous field of antennas generate constructive interference, and Eq. (2) becomes

$$\begin{aligned} y(t, \boldsymbol{r}) &= s(t)\left(\int_{\mathbb{R}^n} w^*(\boldsymbol{p})e^{-j2\pi\langle \boldsymbol{r}, \boldsymbol{p}\rangle} d\boldsymbol{p}\right), \\ &= s(t)\hat{w}^*(\boldsymbol{r}). \end{aligned} \tag{3}$$

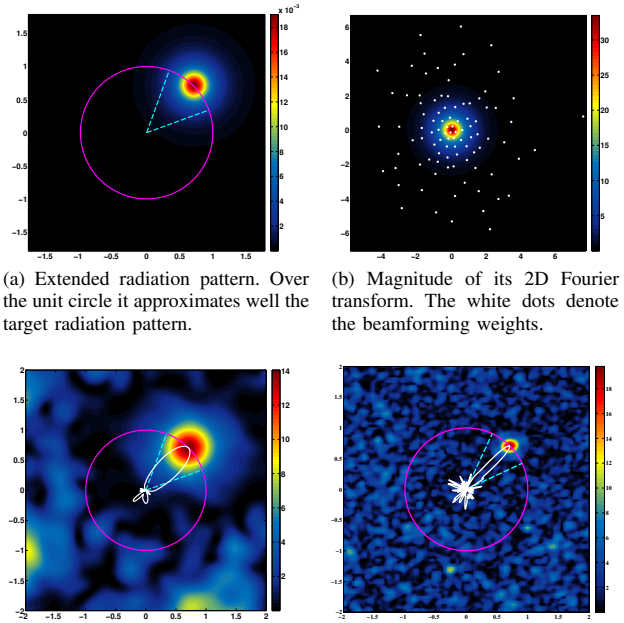
The beam-shape for the notional antenna field is then $\hat{w}(\boldsymbol{r}) \in \mathcal{L}^2(\mathbb{S}^{n-1}, \mathbb{C})$. It describes the radiation strength of the beamformed antenna field towards various directions, and as such acts as a spatial filter. The link to the beamforming function is as follows:

$$\hat{w}(\boldsymbol{r}) = \int_{\mathbb{R}^n} w(\boldsymbol{p})e^{j2\pi\langle \boldsymbol{r}, \boldsymbol{p}\rangle} d\boldsymbol{p}.$$

The beamforming function was defined thus far only over the sphere $\mathbb{S}^{n-1}$. To enable sampling at any point in the plane, and to have a realisable $n$-dimensional Fourier transform relationship, the filter needs to be extended to $\mathbb{R}^n$. Let then $\hat{\omega} : \mathbb{R}^n \to \mathbb{C}$ be a function whose $n$D Fourier transform exists, and on the hypersphere $\mathbb{S}^{n-1}$ is equal to the target radiation pattern we would like to achieve. We call $\hat{\omega}(\boldsymbol{r})$ thus designed the *extended radiation pattern*.

The actual choice of extension is application dependent, and part of the design. The beamforming function can now be computed by the Fourier transform

$$w(\boldsymbol{p}) = \int_{\mathbb{R}^n} \hat{\omega}(\boldsymbol{r})e^{-j2\pi\langle \boldsymbol{r}, \boldsymbol{p}\rangle} d\boldsymbol{r}. \tag{4}$$



(a) Extended radiation pattern. Over the unit circle it approximates well the target radiation pattern.

(b) Magnitude of its 2D Fourier transform. The white dots denote the beamforming weights.

(c) Beam-shape obtained with beamforming weights (white line), which approximates relatively well the target.

(d) Beam-shape obtained with matched beamforming.

Fig. 2: Filtering a range of directions with flexibeam for $\Theta = 40^\circ$ and 96 antennas. The beam-shape covers a much wider range of directions than matched beamforming.

which, for an arbitrary target radiation pattern, would be calculated numerically. However, the target and extended radiation pattern can be designed so that an analytical Fourier transform exists. In particular, and relevant for the example we show in the next section, the *n-dimensional symmetric Gaussian*

$$\hat{\omega}(\boldsymbol{r}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\|\boldsymbol{r}-\boldsymbol{r}_0\|^2}{2\sigma^2}}, \tag{5}$$

with mean $\boldsymbol{r}_0 \in \mathbb{S}^{n-1}$ and standard deviation $\sigma$ has Fourier transform

$$w(\boldsymbol{p}) = (2\pi)^n e^{-2\pi^2\sigma^2\|\boldsymbol{p}\|^2} e^{-j2\pi\langle \boldsymbol{p}, \boldsymbol{r}_0\rangle}. \tag{6}$$

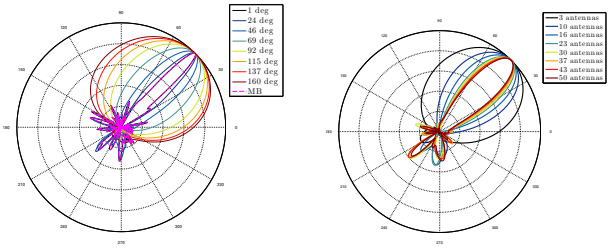
Consider now $L$ antennas with positions $\boldsymbol{p}_i$, $i = 1 \dots L$. The beamforming weight for antenna $i$ is then

$$\frac{w(\boldsymbol{p}_i)}{\sqrt{\sum_{i=1}^{L} |w(\boldsymbol{p}_i)|^2}} = \frac{w(\boldsymbol{p}_i)}{\beta}.$$

Hence, the gains $\gamma_i$ and phase delays $\phi_i$ for antenna $i$ are given by

$$\gamma_i = \frac{|w(\boldsymbol{p}_i)|}{\beta}, \quad \phi_i = \arg\left(w(\boldsymbol{p}_i)\right), \quad i = 1, \dots, L.$$

The normalisation $\beta$ prevents antennas from having too high diversity in magnitude, which would magnify their response to channel noise.




(a) Filtering a range of directions with flexibeam for various angles and 96 antennas.

(b) Filtering a range of directions with flexibeam for $\Theta = 35^\circ$ with varying number of antennas.

Fig. 3: Evolution of the flexibeam beam-shape for various angles and number of antennas.

How closely the beamforming achieves the target radiation filter over the sphere $\mathbb{S}^n$ depends strongly on the number and position of the antennas. This effectively is the ability of an FIR filter to approximate an IIR filter using a given number of taps (antennas).

## III. Tracking with flexibeam

Suppose we wish to track the direction of a target moving on the plane. We initially think it is located at $\hat{\theta}_0 = 45^\circ$, but are uncertain. If we try to target it using too narrow a beam we could miss the target altogether. We thus calculate beamforming weights using flexibeam so as to obtain a radiation pattern with a wide enough main lobe, centred around our estimate $\hat{\theta}_0 = 45^\circ$. This wider beam permits tracking for a longer period of time, which avoids having to refresh the beam too often as the target moves.

From experimental conditions, the optimal radiation pattern $\hat{w}(\theta)$ was estimated to be

$$\hat{w}(\theta) = \frac{1}{\sqrt{2\pi}\Theta} e^{-\frac{(\theta-\hat{\theta}_0)^2}{2\Theta^2}}, \tag{7}$$

where $\theta$ is an angle on the unit circle $\mathbb{S}^1$ measured in degrees, and $\Theta = 40^\circ$ is the desired width of the main lobe. For practical purposes, we propose to extend $\hat{w}(\theta)$ to $\mathbb{R}^2$ by the 2D symmetric Gaussian function (5), with $\boldsymbol{r}_0 = (1,1)/\sqrt{2} \in \mathbb{S}^1$, $\sigma = \sqrt{2 - 2\cos\Theta}$. Strictly speaking, this is only an approximate extension of Eq. (7). However, for reasonable beam widths $\Theta$, this approximation is accurate enough (see Fig. 2a), and conveniently provides us with an analytical expression for the beamforming function, shown on Fig. 2b.

The beamforming weights are determined by sampling the beamforming function at the antennas' positions (see Fig. 2b). The resultant beam-shape in Fig. 2c can be seen in general to be a good approximation. In contrast, matched beamforming would require steering towards many directions to cover the same area, and hence would be more likely to miss the moving target if the refresh rate is not high enough.

For a fixed number of antennas Fig. 3a shows that for very small $\Theta$ the beam-shape is essentially identical to the one from matched beamforming, while for larger $\Theta$, the beam-shape struggles to cover the whole range (because the 2D Gaussian extension does not approximate well enough the target radiation filter over the sphere). For fixed $\Theta$, Fig. 3b shows that the beam-shape becomes increasingly accurate as the number of antennas increases.

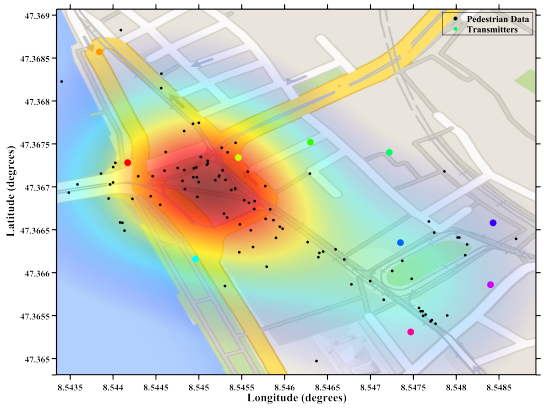



Fig. 4: Density function of pedestrians in Bellevueplatz, Zurich. The black dots are sampled positions from which the density has been inferred. The coloured dots are the transmitters.

## IV. Example using mobile base stations

We now illustrate an example for beamforming a collection of 3G/4G transmitters, in order to cover optimally Bellevueplatz, a portion of the city of Zurich, given probable client positions. Bellevueplatz has an approximate area of 0.08 km$^2$, and welcomes one of the biggest tram stations with correspondingly dense pedestrian traffic. For this experiment, we gathered positions of pedestrians in this area (black dots on Fig. 4), and inferred a continuous density function (the coloured regions). This density function is called the *preference function*. It describes where the power is most needed. The goal is then to beamform from each of the 10 transmitters (the coloured dots), so that they, based on pedestrian density, jointly cover the area well.

We assume devices are in the far-field and that the channel has a narrow bandwidth. For simplicity, we neglect signal attenuation. Each transmitter has 27 antennas arranged on three concentric circles of radii 5, 15 and 25cm respectively. Moreover, they are assumed to have an emission range of approximatively 100m. Hence, each transmitter only sees a circular cut of a 100m radius of the density function, which defines the individual transmitter preference function $f_i \in \mathcal{L}^2(\mathbb{R}^2),\ i = 1, \ldots, 10$.

The beam determination problem for each transmitter consists then of four steps:

1) Compute the target radiation pattern by taking the radial projection of the individual preference functions from

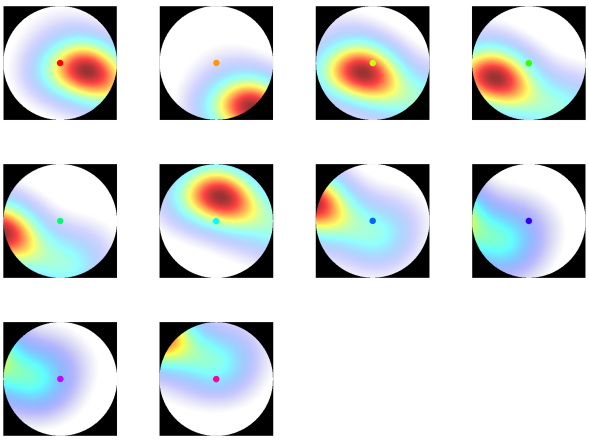
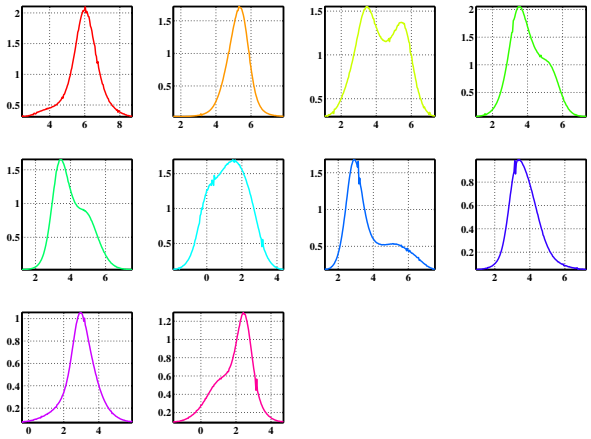
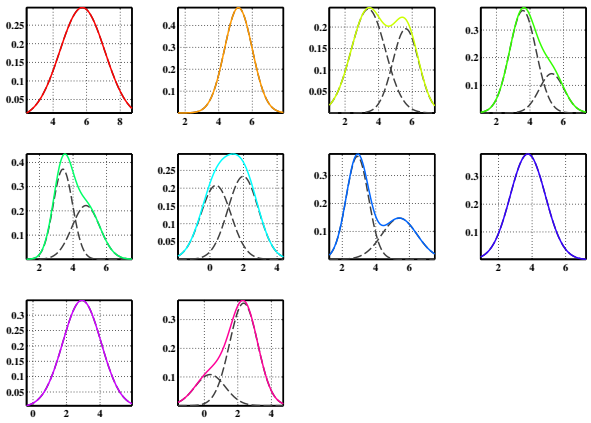


Fig. 5: Circular cuts of the density function in Fig. 4 based on the range of each transmitter.



(a) Target radiation patterns for each transmitter plotted over a segment of length $2\pi$.



(b) Approximation of the target radiation patterns by a sum of weighted Gaussian functions.

Fig. 6: Target radiation patterns for each transmitter and their approximation by a sum of weighted Gaussian functions.

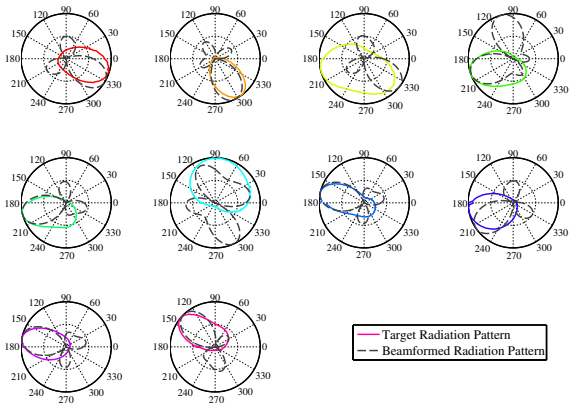



Fig. 7: Comparison between the target radiation pattern (coloured lines) and the actual achieved beam-shape (dashed grey lines) for each transmitter.

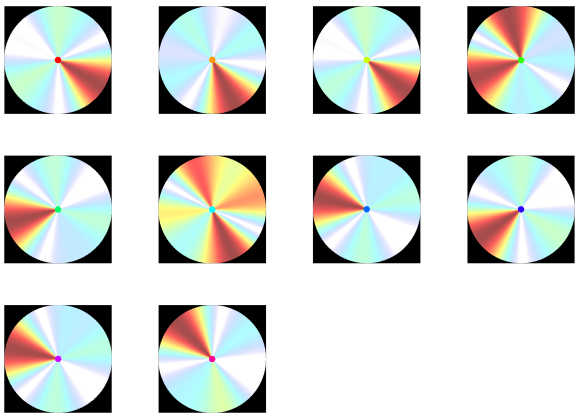


Fig. 8: The transmitters cover more areas with a high density of pedestrian (compare with Fig. 5).

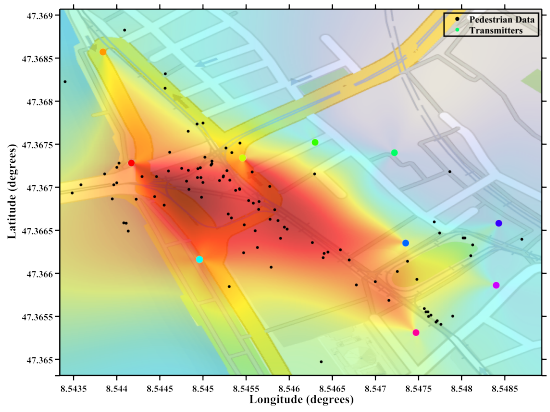



Fig. 9: The summation of the beam-shapes from each transmitter gives the joint coverage. Joint coverage achieved by all the transmitters after choosing the beamforming weights. Areas with high pedestrian density are better covered.

Fig. 5,

$$\hat{w}_i(\theta) = \int_0^{100} f_i(r\cos\theta, r\sin\theta)dr,$$

with $\theta \in [0, 2\pi]$.

2) Approximate this target filter by a sum of weighted Gaussian functions (see Fig. 6b),

$$\hat{w}_i(\theta) \simeq \sum_{k=1}^{N_i} \frac{\alpha_k^{(i)}}{\sqrt{2\pi}\Theta_k^{(i)}} e^{-\frac{\left(\theta-\mu_k^{(i)}\right)^2}{2\Theta_k^{(i)\,2}}},$$

where $N_i \in \mathbb{N}$, $\alpha_k^{(i)}, \Theta_k^{(i)} > 0$ and $\mu_k^{(i)} \in [0, 2\pi]$ are respectively the least squares estimates of the number of Gaussian components and their associated weights, standard deviations and means.

3) Extend this filter to the plane with the same technique as described in Section III and compute its Fourier transform analytically using Eq. (6). We get

$$\hat{\omega}_i(x, y) = \sum_{i=1}^{N_i} \alpha_k^{(i)} \mathbf{\Phi}\left(\frac{x - \cos(\mu_k^{(i)})}{\sigma_k^{(i)}}, \frac{y - \sin(\mu_k^{(i)})}{\sigma_k^{(i)}}\right),$$

where $(x, y) \in \mathbb{R}^2$, $\mathbf{\Phi} \in \mathcal{L}^2(\mathbb{R}^2)$ is the standard 2D Gaussian function and $\sigma_k^{(i)} = \sqrt{2 - 2\cos(\Theta_k^{(i)})}$.

4) Compute the weights to be applied to each antenna composing the transmitter by sampling the Fourier transform at the locations of the antennas.

Most of the transmitter beam-shapes, shown in Fig. 7, approximate the associated target filter well, despite unavoidable side-lobes due to the finite number of antennas. Fig. 9 shows that areas with higher pedestrian density are better covered than before, giving them better signal, as less power is dissipated in unnecessary areas.

## V. Conclusions

We took the flexibeam technique to determine beamforming weights for a target spatial filter, and explained it from the transmit beamforming perspective. We then showed on an example how it can be used by groups of mobile base stations to concentrate energy where devices are concentrated.

We argued the case for targeting regions rather than single points, and showed that this could be achieved. One interesting effect is that the MIMO optimisation tradeoff between (focused) beamforming and spatial diversity (multiple replicas of the radio signal from different directions) [10] can be circumvented.

Of course in practise real-life data transfer and the resultant communications protocol is far more complicated, and beamforming is just one component in the mix. Future work includes incorporating these together, and adding cooperation between the stations to maximise throughput and minimise latency.

## References


[1] R.-T. Juang, K.-P. Yar, K.-Y. Lin, and P. Ting, "Decentralized multiuser beamforming for cellular communication systems," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2011 IEEE 7th International Conference on*, Oct 2011, pp. 260–264.

[2] C. Jiang and L. Cimini, "Energy-efficient multiuser mimo beamforming," in *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, March 2011, pp. 1–5.

[3] J. Jin, C. Lin, Q. Wang, H. Yang, and Y. Wang, "Effect of imperfect channel estimation on multi-user beamforming in lte-advanced system," in *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, May 2010, pp. 1–5.

[4] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. Simon & Schuster, 1992.

[5] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 4 1988.

[6] Y.-S. Cheng and C.-H. Chen, "A novel 3d beamforming scheme for lte-advanced system," in *Network Operations and Management Symposium (APNOMS), 2014 16th Asia-Pacific*, Sept 2014, pp. 1–6.

[7] P. Hurley and M. Simeoni, "Flexibeam: analytic spatial filtering by beamforming," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, March 2016 (to appear).

[8] R. J. Mailloux, "Phased array antenna handbook," *Boston, MA: Artech House, 1994.*, 1994.

[9] W.-Q. Wang and H. Shao, "A flexible phased-mimo array antenna with transmit beamforming," *International Journal of Antennas and Propagation*, vol. 2012, 2012.

[10] A. Sibille, C. Oestges, and A. Zanella, *MIMO: from theory to implementation*. Academic Press, 2010.

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# EXIT Chart Analysis of the UMTS Turbo Code in VLF Channels

Alexander Hamilton
QinetiQ Portsdown Technology Park & University of Portsmouth
Maritime Systems
PO6 3RU Portsmouth, UK
Email: ajhamilton2@qinetiq.com

***Abstract*— For communications at sea and underwater, Very Low Frequency (VLF) telecommunications are typically used. These exist in a channel that is subject to highly impulsive noise, primarily due to the effects of worldwide lightning activity. There is currently a drive to increase range and depth of these communications and as such there is a need for novel energy efficient error correction schemes to be trialled and developed in these environments. This paper aims to use EXIT chart analysis to understand the performance of the Universal Mobile Telecommunications System (UMTS) Turbo code in these environments.**

## I. INTRODUCTION

VLF (3 - 30 kHz) electromagnetic waves are used for long range communication from shore infrastructure to submarines or ships due to low path attenuation, of the order of 2 - 3 dB per 1000 km [1] by utilising the ionosphere as a waveguide. VLF channels are greatly impacted by atmospheric noise, primarily due to the propagation of the impulsive noise created by lightning strikes, which can be observed around the globe [2].

Communications at VLF are assigned very little bandwidth due to electrically short antenna and frequency limitations therefore the majority of channels have a 3 dB bandwidth of approximately 120 Hz [3].

The VLF channel offers a non-trivial engineering challenge for information theory and communications as it does not conform to a traditional noise model, exhibiting a high level of kurtosis and as such can provide unusual results on communications through this channel.

The channel also exhibits characteristic properties that are hugely variant upon geographic location and diurnal conditions, due to variations in the ionosphere affecting propagation and the frequency of impulsive noise due to lightning activity.

In addition to this the effects of sea state can create large phase variations into a phase modulated communications system. These phase variations, provide a significant limitation in underwater reception, and tracking methods are currently under investigation at QinetiQ.

Current VLF systems use MSK modulation combined with a Wagner parity check [4] , the next generation will use iterative LDPC Forward Error Correction (FEC) in order to achieve as low a Bit Error Rate (BER) as possible, the details of this FEC and frame size for VLF communications is defined in various military interoperability standards. For the purpose of simulation these will be modelled as frames of 1000 bits of length.

Previous analytical methods used to provide evidence to support the decision to use a LDPC were based on simple BER analysis. However in recent years, advances by Ten Brink [5] and Hanzo et al. [6] and their contemporaries have facilitated a deeper understanding of turbo and turbo like codes by observing the message passing and the transfer of extrinsic information within the decoder itself.

## II. METHODOLOGY

In order to simulate the VLF channel it is necessary to understand the effects of impulsive noise in the VLF channel. In order to do that off air recordings at VLF were taken. Within these recordings the impulsive noise was measured at a frequency clear of any other communications, in this case 27 kHz.

These recordings then allowed a hard decision mask to be created dependent on the noise power which directly relates back to the Signal to Noise Ratio that will be observe by the submarine. This mask or error profile could then be applied to a frame, in order to represent where bits would be erased, effectively creating a Binary Erasure Channel (BEC).

An example of the VLF noise spectrum can be seen in Fig. 1

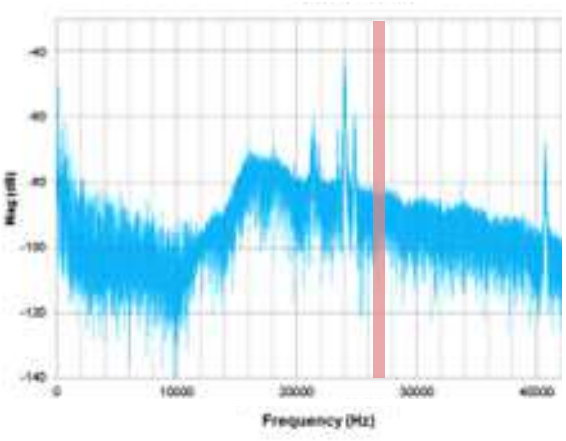



Fig 1 - VLF Spectrum

Using an off air recording taken at VLF frequencies, it is possible to create an erasure mask, representing when information is unable to pass through the VLF channel due to the presence of impulsive noise spikes. An example of this can be seen in Fig. 2 which shows the error distribution (where a 1 represents an erasure and a 0 represents a symbol being received intact)

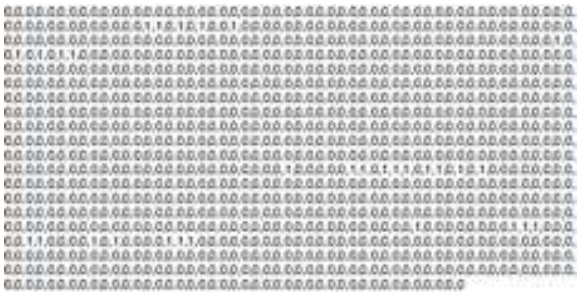


Fig 2 - VLF Erasure Mask

This mask could then be applied to the transmitted information in the channel simulation in order to understand the impact that this VLF noise will have on the received signal, and how it will affect the decoder.

For the UMTS decoder to decode the information a Gaussian distributed set of Logarithmic Likelihood Ratios (LLRs) of the information was generated to represent the transmitted encrypted information.

For the modelling the erasure mask was applied with varying levels of AWGN, this is a rudimentary technique used to simulate VLF noise - a future implementation will aim to have statistically generated 'soft' error masks which will be validated against off air recordings.

The UMTS decoder has been built according to the specification in ETSI TS 125 212.

Concurrent analysis [7] has identified that a LDPC code with a degree of 4 will require a signal to noise ratio greater than 0 dB for full convergence. in an impulsive noise channel.

All modelling assumes link encryption and that all error correction will be conducted on cipher text.

## III. RESULTS

Gaussian distributed Logarithmic Likelihood Ratios (LLRS) were generated and passed through the channel model at varying SNRs in order to generate the following EXIT charts.

As can be seen from the results the EXIT tunnel closes at -7 dB SNR (referenced to a 1 kHz Bandwidth), suggesting that at higher SNRs the error rate will exist in the error floor region of the BER curve [6] due to full convergence of the code.

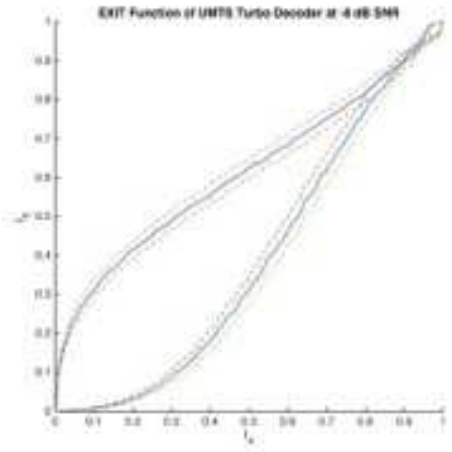


Fig 3 - EXIT chart of UMTS Turbo code in simulated VLF Atmospheric Noise at -8 dB SNR

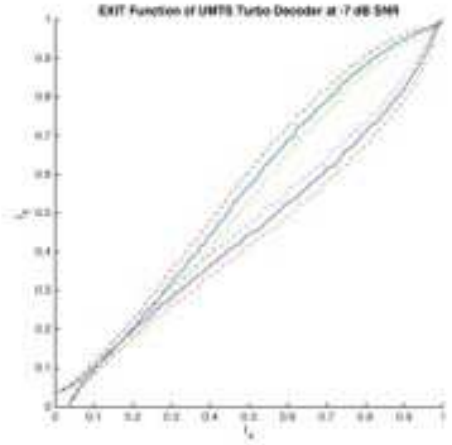


Fig 4 - EXIT chart of UMTS Turbo code in simulated VLF Atmospheric Noise at -7 dB SNR

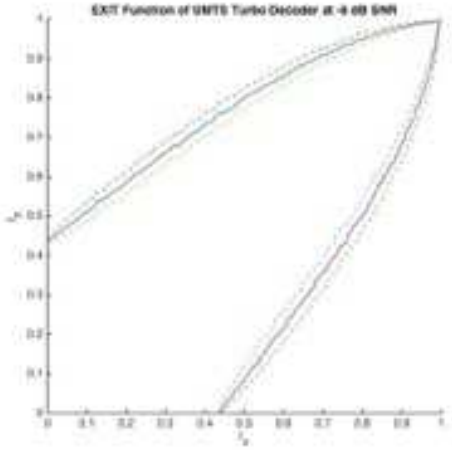


Fig 5 - EXIT chart of UMTS Turbo code in simulated VLF Atmospheric Noise at -6 dB SNR

The EXIT tunnel closure can be seen in Fig. 4 at -7 dB SNR whereas in Fig. 3 taken at -8 dB SNR the EXIT tunnel remains firmly closed preventing the passage of information through the iterative decoding process.

Conversely it can be observed that at higher SNRs such as can be seen in Fig. 5 at -6 dB SNR the EXIT tunnel is large enough to provide information transfer for all possible information trajectories, suggesting a high level of convergence for the Turbo code.

## IV. Discussion

As can be seen from the prior analysis the UMTS turbo code offers a performance advantage of 7 dB SNR over the current implementation of the LDPC error correction code which can only operate up to 0 dB SNR for full convergence.

In order to demonstrate the performance advantage that this provides the Long Wavelength Propagation Capability (LWPC) tool was used to produce of Fig. 6. The input parameters for this plot are as follows:

Transmitter Location - Cutler, Maine
Transmitter Power - 200 kW
Transmit Frequency - 24 kHz
Time and Date - July 15 2015, 0300 hrs

As can be seen the benefit that the UMTS code offers is rather significant, and can vastly increase the area of operations that VLF communications can cover.

The hard erasure mask used in this modelling will force an erasure, however this may cause the modelled UMTS code to function at pessimistic SNRs. In the future it is intended for a statically generated ‘soft’ mask to be produced which will not force an erasure due to impulsive noise, and instead permit the signal through with a reduced LLR, allowing for further analysis of a coding scheme.

## V. Conclusion

The analysis has shown that the UMTS turbo code is resilient to VLF noise to a far greater degree than an LDPC code as is currently implemented. However it may be that the LDPC code which at this moment is of degree 4 may be a sub-optimal design for the VLF channel.

For this reason it is recommended that further research must be conducted on the optimal LDPC design for resilience to noise in VLF channels, before any conclusive proof is put forward for the proposed error correction scheme for use in VLF channels.

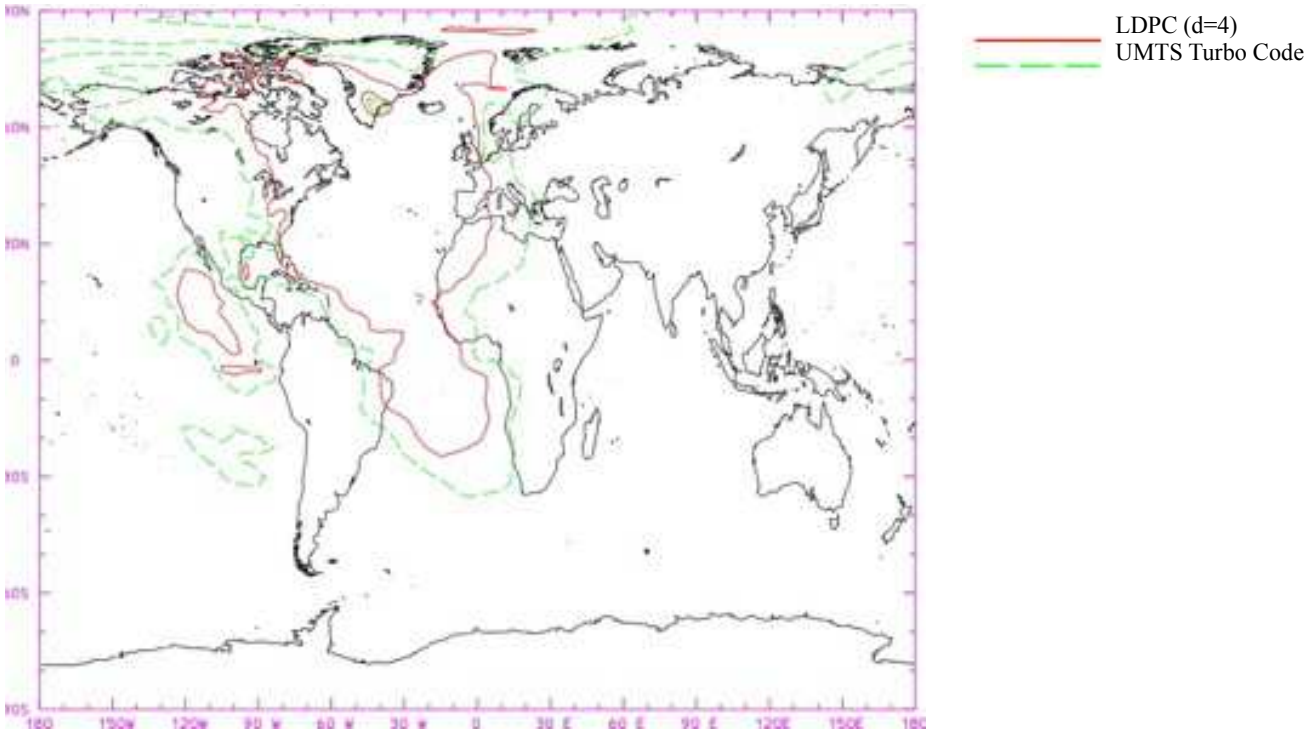



Fig 6 - VLF coverage plot showing the difference in coverage area between a LDPC (d=4) coded signal and a UMTS Turbo coded signal

## ACKNOWLEDGMENT

This research has been conducted under the VLF research program, sponsored by UK MoD. It has been supported by the author's MSc supervisor at the University of Portsmouth, Dr. Salem Al-Jareh & with support from Dr. Rob Maunder at the University of Southampton, in providing background knowledge in EXIT chart analysis.

## REFERENCES

1. MOORE, R.K [1967] "*Radio communication in the sea"* Spectrum, IEEE, 4(11), 42-15.
2. HELLIWELL, R.A. [1965] "*Whistlers and Related Ionospheric Phenomena,"* Stanford University Press, Stanford, CA, USA
3. HARRINGTON, M.D. [1993] "*A dynamic bandwidth and phase linearity measurement technique for 4-channel MSK VLF antenna systems.*" In AGARD, ELF/VLF/LF Radio Propagation and Systems Aspects 8 p (SEE N93-30727 11-32). Vol. 1.
4. PING, L., CHAN, S., & YEUNG, K.L. [1998] "*Iterative decoding of multi-dimensional concatenated single parity check codes"* Communications, IEEE International Conference on, (Vol. 1, pp. 131-155)
5. TEN BRINK, S. [2001] *"Convergence behaviour of iteratively decoded parallel concatenated codes"* Communications, IEEE Transactions on, 49(10): p. 1727-1737
6. HANZO, L., TONG, H. L. and BEE, L.Y. [2002] *"Turbo coding, turbo equalisation and space-time coding."* John Wiley & Sons,
7. HAMILTON, A.J.B. [2015] *"EXIT Chart analysis of LDPC codes in VLF Channels"* Unpublished manuscript.

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# Channel Vector Subspace Estimation from Sample Covariance of Low-Dimensional Projections

Saeid Haghighatshoar*, Giuseppe Caire*
*Communications and Information Theory Group, Technische Universität Berlin
Email: (saeid.haghighatshoar, caire)@tu-berlin.de

***Abstract*—In this paper, we propose efficient algorithms for estimating the signal subspace of mobile users in a wireless communication environment with a multi-antenna base-station with $M$ antennas. When $M$ is large, because of the high angular resolution of the receiver, any realization of the random channel vector of any given user is approximately contained in a user-specific subspace of dimension $p \ll M$. Efficient multiuser MIMO schemes can be obtained from such subspace information, which is stable in time and can be accurately estimated even in the presence of fast fading (e.g., for mm-Wave channels). We are interested in the massive MIMO regime of $M \gg 1$. In order to reduce the RF front-end complexity and overall A/D conversion rate, the $M$-antenna base-station transmitter/receiver is split into the product of a baseband linear projection (digital) and an RF reconfigurable beamforming network (analog) with only $m \ll M$ RF chains. Hence, only $m$-dimensional analog observations can be obtained for subspace estimation. We develop efficient algorithms that estimate the dominant signal subspace of the users from sampling only $m = O(2\sqrt{M})$ specific array elements according to a coprime scheme. For a given target dimension of the signal subspace $p \leq M$, our algorithms return a $p$-dimensional beamformer with a performance comparable with the best $p$-dim beamformer designed by knowing the exact covariance matrix of the received signal. We assess the performance of our proposed estimators both analytically and empirically via numerical simulations, and compare it with that of the other state-of-the-art methods in the literature.**

## 1 Introduction

Consider a multiuser MIMO channel formed by a base-station with $M$ antennas and $K$ single-antenna mobile users in a cellular network. We focus here on a flat-fading channel in which the bandwidth of the signal is less than the channel's coherence bandwidth. Following the current *massive MIMO* approach [1, 2], we assume that the uplink and the downlink are organized in Time Division Duplexing (TDD), where the base-station estimates the channel vectors of the users from orthogonal pilots that are sent by the users in the uplink in the same channel coherence time [1]. It turns out that for isotropically distributed channel vectors it is optimal to devote half the coherence time to estimate the channel, and to devote the remaining half to serve the users.

In the massive MIMO setup, the number of antennas $M$ is large, and the receiver antenna at the base-station has a high angular resolution. Consequently, in many relevant scenarios, the channel is far from isotropic. Indeed, as the propagation for a user occurs only through a small set of Angles of Arrivals (AoAs), its channel vectors in consecutive coherence blocks lie on very low-dimensional subspaces. This underlying structure can be exploited to improve the system multiplexing gain via decreasing the training overhead. For example, one approach would be to cluster the users based on the dominant subspace of their channel vectors, and apply the classical channel estimation on a per-group basis, on the low-dimensional projected channels [3]. This requires estimating the dominant signal subspace of each individual user. Although the channel vector changes in every coherence time, in many practical scenarios, the signal subspace remains stationary across many coherence blocks, thus, it can be reliably estimated.

A direct naive approach for estimating the signal subspace is to first estimate the $M \times M$ covariance matrix of the channel coefficient of each user via sampling the whole array elements, and then identify the signal subspace by applying the singular value decomposition (SVD). This requires sampling the whole array elements which requires $M$ RF chains. Since in massive MIMO setup $M \gg K$, this is inefficient and very difficult to implement. Different architectures such as Hybrid Digital Analog (HDA) have been proposed to reduce hardware complexity (notably, the A/D overall bit-rate and the number of RF modulation/demodulation chains). The main idea is to implement the $M \times K$ beamforming matrix as the product of two matrices: an $M \times m$ beamforming matrix implemented in the RF analog domain, and an $m \times K$ precoding matrix implemented in the digital baseband domain, so that only $m \ll M$ A/D converter and RF chains be used. This implies that exploiting the subspace information is possible only when it can be extracted (estimated) from $m$-dimensional sketches ($m \ll M$) of the received signal.

In this paper, we aim to design suitable subspace estimators from low-dimensional sketches of the input signal for a uniform linear array (ULA). The geometry of the array and the scattering channel is shown in Fig. 1. Array elements have a uniform distance $d = \frac{\lambda}{2\sin(\theta_m)}$, and scan the angular range $[-\theta_m, \theta_m]$ for some $\theta_m \in (0, \pi/2)$. We use a coprime sampling scheme, introduced in [4], that samples only $O(2\sqrt{M})$ specific array elements. We propose several algorithms for estimating the signal subspace and cast them as convex optimization problems that can be solved efficiently. We also analyze the performance of our estimators in terms of the dimension of the desired signal subspace $p$, array size $M$, training length $T$, and signal-to-noise ratio (SNR).

## 2 Related work

Several works in the literature are related to the problem addressed in this paper, which can be summarized in the

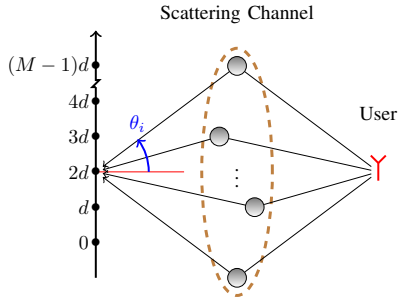



Fig. 1: Scattering channel with discrete angles of arrivals.

following four categories: Subspace tracking, Low-rank matrix recovery, Direction-of-arrival (DoA) estimation, and Multiple Measurement Vectors (MMV) problem in compressed sensing (CS). Let us consider a simple model in which the transmission between a user and the base-station occurs through $p$ scatterers (see Fig. 1). One snapshot of the received signal is given by

$$\mathbf{y} = \sum_{\ell=1}^{p} \mathbf{a}(\theta_\ell) w_\ell \, x + \mathbf{n}, \tag{1}$$

where $x$ is the transmitted (training) symbol, $w_\ell \sim \mathcal{CN}(0, \sigma_\ell^2)$ is the channel gain of the $\ell$-th multipath component, $\mathbf{n} \sim \mathcal{CN}(0, \mathbf{I}_M)$ is the additive white Gaussian noise of the receiver antenna, and where $\mathbf{a}(\theta) \in \mathbb{C}^M$ is the array response at AoA $\theta$, whose $k$-th component is given by

$$[\mathbf{a}(\theta)]_k = e^{jk\frac{2\pi d \sin(\theta)}{\lambda}} = e^{jk\pi \frac{\sin(\theta)}{\sin(\theta_m)}}. \tag{2}$$

According to the WSSUS model, the channel gains for different paths, i.e., $\{w_\ell\}_{\ell=1}^p$, are uncorrelated, and since they are (jointly) Gaussian, they are statistically independent. Without loss of generality, we suppose $x = 1$ in all training snapshots. Letting $\mathbf{A} = [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), \ldots, \mathbf{a}(\theta_p)]$, we have

$$\mathbf{y}(t) = \mathbf{A}\mathbf{w}(t) + \mathbf{n}(t), \quad t \in [T], \tag{3}$$

where $\mathbf{w}(t) = (w_1(t), w_2(t), \ldots, w_p(t))^{\mathsf{T}}$ for different $t \in [T] := \{0, 1, \ldots, T-1\}$ are statistically independent. We assume that the AoAs $\{\theta_\ell\}_{\ell=1}^p$ remain invariant over a long time $T$. From (3), the covariance of $\mathbf{y}(t)$ is given by

$$\mathbf{C}_y = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathsf{H}} + \mathbf{I}_M = \sum_{\ell=1}^{p} \sigma_\ell^2 \mathbf{a}(\theta_\ell)\mathbf{a}(\theta_\ell)^{\mathsf{H}} + \mathbf{I}_M. \tag{4}$$

Let $\mathbf{C}_y = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{H}}$ be the singular value decomposition (SVD) of $\mathbf{C}_y$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ denotes the diagonal matrix of singular values. We always assume that the singular values are sorted in a non-increasing order. If we denote by $\mathbf{U}_p$ the $M \times p$ matrix consisting of the first $p$ columns of $\mathbf{U}$, it is not difficult to see that the columns of $\mathbf{U}_p$ span the signal space. In particular, $\mathsf{span}(\mathbf{A}) = \mathsf{span}(\mathbf{U}_p)$. We need to identify this subspace from noisy low-dimensional sketches $\mathbf{x}(t) = \mathbf{B}\mathbf{y}(t)$, where $\mathbf{B}$ is the sampling matrix. This problem for the noiseless case was studied by Chi et. al. in [5] where they developed PETRELS algorithm to estimate the underlying subspace. Another algorithm named GROUSE was proposed by Balzano et. al. in [6] which uses a low-complexity stochastic gradient update over the Grassmanian manifold. Both algorithms mainly optimize the computational complexity rather than the data size, and principally suit situations in which the dimension is high (very large $M$ and $T$). We empirically compare the performance of our proposed algorithms with PETRELS for a fixed data size in Section 5.

For $p \ll M$ and for a high SNR, the covariance matrix $\mathbf{C}_y$ in (4) is nearly low-rank. Recovery of low-rank matrices from a collection of a few possibly noisy samples is of great importance in signal processing and machine learning. Recently, it has been shown that this can be done via nuclear norm minimization, which is a convex optimization and can be efficiently solved [7]:

$$\mathbf{X}^* = \arg\min_{\mathbf{M}} \ \|\mathbf{M}\|_* \text{ subject to } \mathbf{M}_\Omega = \mathbf{X}_\Omega, \tag{5}$$

where the nuclear norm $\|\mathbf{M}\|_*$ is given by the sum of the absolute value of the eigen-values of $\mathbf{M}$, and reduces to $\mathsf{Tr}(\mathbf{M})$ when $\mathbf{M}$ is positive semi-definite (PSD). In practice, we have only a collection of $T$ snapshots of sketches $\mathbf{x}(t) = \mathbf{B}\mathbf{y}(t)$, $t \in [T]$, rather than the whole covariance matrix or even the sketches thereof. Let

$$\widehat{\mathbf{C}}_y = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)\mathbf{y}(t)^{\mathsf{H}}, \quad \widehat{\mathbf{C}}_x = \mathbf{B}\widehat{\mathbf{C}}_y\mathbf{B}^{\mathsf{H}} \tag{6}$$

be the sample covariance of the full and subsampled signal. We compare the performance of our algorithms with the following extension of the nuclear norm minimization

$$\min_{\mathbf{M}} \ \mathsf{Tr}(\mathbf{M}) \text{ subject to } \mathbf{M} \in \mathbb{T}_+, \ \|\widehat{\mathbf{C}}_x - \mathbf{B}\mathbf{M}\mathbf{B}^{\mathsf{H}}\| \le \epsilon, \tag{7}$$

where $\mathbb{T}_+$ is the space of $M \times M$ PSD Toeplitz matrices, and where $\epsilon$ is an estimate of the $\ell_2$-norm of the error.

From (3), it is seen that the received signal $\mathbf{y}(t)$ is a noisy superposition of $p$ independent Gaussian sources arriving from $p$ different angles. This is the same model studied for direction-of-arrival (DoA) estimation. There are two main categories of algorithms for DoA estimation: classical algorithms such as MUSIC and ESPRIT that use subspace methods to locate the AoAs, and more recent compressed sensing based algorithms that use the angular sparsity of the signal over a prespecified grid (see [8, 9] and refs. therein). Recently, Candès and Fernandez-Granda [10] developed an off-grid *super-resolution* (SR) technique using total-variation (TV) minimization. This algorithm was extended by Tan et. al. in [11] to DoA estimation with coprime arrays when the sources are sufficiently separated. In a wireless environment the AoAs are clustered. This implies that the separation requirement for the super-resolution setup may not be met. Since in this paper we aim at estimating the subspace of the signal rather than DoAs, in section 4.2.3 we extend the super-resolution method to develop a new algorithm for estimating the signal subspace.

It is seen from (3) that, neglecting the measurement noise $\mathbf{n}(t)$, the signal $\mathbf{y}(t)$ has typically a sparse representation over the continuous dictionary $\{\mathbf{a}(\theta), \theta \in [-\theta_m, \theta_m]\}$, i.e., only $p$ atoms of the dictionary, i.e., $\{\mathbf{a}(\theta_i)\}_{i=1}^p$, are needed to represent the signal. Thus, $\mathbf{x}(t) = \mathbf{B}\mathbf{y}(t)$ can be seen as identifying a sparse vector from a collection of sketches, which coincides with the traditional CS problem. An extension of this problem involves Multiple Measurement Vectors (MMV).

The underlying assumption is that $\mathbf{y}(t)$, for different snapshots $t \in [T]$, have the same sparsity pattern or support over the underlying dictionary even though they might have different coefficients $\mathbf{w}(t)$ for each $t$. This problem has been widely studied in the literature (see [9, 12, 13] and refs. therein), where two main approaches have been proposed for estimating the common support of the signals: using a greedy algorithm or convex optimization via a regularizer promoting group sparsity; and using covariance matrix of data and subspace techniques. Once the support is identified, the standard Least-Squares method can be used to find the coefficients. Since the underlying dictionary is continuous, both classes exploit either grid-based or more recently developed off-grid techniques. We will compare the performance of our algorithm with grid-based approach in [13], and the grid-less one in [14–16].

## 3 Statement of the Problem

In (3), we introduced the channel model given by $\mathbf{y}(t) = \mathbf{A}\mathbf{w}(t)+\mathbf{n}(t)$, $t \in [T]$, where $\mathbf{A}$ contains the array response for the AoAs. Let $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ be the matrix containing the channel strengths for the AoAs $\{\theta_\ell\}_{\ell=1}^p$. We can prove the following simple result.

**Proposition** *3.1:* Let $\widehat{\mathbf{C}}_x = \frac{1}{T}\sum_{t=1}^T \mathbf{x}(t)\mathbf{x}(t)^{\mathsf{H}}$ be the sample covariance of the sketches $\mathbf{x}(t) = \mathbf{B}\mathbf{y}(t)$, $t \in [T]$. Then $\widehat{\mathbf{C}}_x$ is a sufficient statistics for estimating $\mathbf{A}$ and $\mathbf{\Sigma}$. ■

For a more practical scenario, we consider the following continuum model

$$\mathbf{y}(t) = \sqrt{\mathsf{snr}} \int_{-1}^{1} \sqrt{f(u)}\mathbf{a}(u)z(u,t)du + \mathbf{n}(t),\ t \in [T], \quad (8)$$

where $\mathsf{snr}$ is the SNR and $z(u,t)$ is a circularly symmetric Gaussian process with $\mathbb{E}\Big\{z(u,t)z(u',t')^*\Big\} = \delta(u-u')\delta_{t,t'}$. The measure $f(u)$ models the distribution of the received signal's power over $u \in [-1,1)$, where $u = \frac{\sin(\theta)}{\sin(\theta_m)}$ for $\theta \in [-\theta_m, \theta_m]$. With some abuse of notation, we denote the array vector in the $u$ domain by $\mathbf{a}(u)$ where $[\mathbf{a}(u)]_k = e^{jk\pi u}$.

Let $\mathbf{C}(f) = \mathbf{S}(f) + \mathbf{I}_M$ be the covariance matrix of the received signal, where $\mathbf{S}(f) = \mathsf{snr}\int_{-1}^{1} f(u)\mathbf{a}(u)\mathbf{a}(u)^{\mathsf{H}}du$ is the covariance of the signal of the user with power distribution $f(u)$. We define the best $p$-dim beamformer for $\mathbf{S}(f)$ as $\mathbf{V}_p = \arg\max_{\mathbf{V}\in\mathbb{H}(M,p)} \langle \mathbf{S}, \mathbf{V}\mathbf{V}^{\mathsf{H}}\rangle$, where $\mathbb{H}(M,p)$ is the space of all $M\times p$ matrices $\mathbf{U}$ with $\mathbf{U}^{\mathsf{H}}\mathbf{U} = \mathbf{I}_p$. We assess the efficiency of $\mathbf{V}_p$ for capturing the signal's power by

$$\delta_p = \frac{\langle \mathbf{S}, \mathbf{V}_p\mathbf{V}_p^{\mathsf{H}}\rangle}{\mathsf{Tr}(\mathbf{S})} = \frac{\mathsf{Tr}\{\mathbf{V}_p^{\mathsf{H}}\mathbf{S}\mathbf{V}_p\}}{\mathsf{Tr}(\mathbf{S})}, \quad (9)$$

where $\delta_p \approx 1$ implies that a significant amount of signal's power is concentrated in a $p$-dim subspace. Let $\widetilde{\mathbf{S}}$ be an estimate of $\mathbf{S}$ and let $\widetilde{\mathbf{V}}_p$ be its best $p$-dim beamformer. We can use $\widetilde{\mathbf{V}}_p$ as an estimate of the optimal beamformer $\mathbf{V}_p$. We define the following metric for the efficiency of $\widetilde{\mathbf{V}}_p$

$$\Gamma_p = \frac{\langle \mathbf{S}, \widetilde{\mathbf{V}}_p\widetilde{\mathbf{V}}_p^{\mathsf{H}}\rangle}{\langle \mathbf{S}, \mathbf{V}_p\mathbf{V}_p^{\mathsf{H}}\rangle} = 1 - \frac{\langle \mathbf{S}, \mathbf{V}_p\mathbf{V}_p^{\mathsf{H}}\rangle - \langle \mathbf{S}, \widetilde{\mathbf{V}}_p\widetilde{\mathbf{V}}_p^{\mathsf{H}}\rangle}{\langle \mathbf{S}, \mathbf{V}_p\mathbf{V}_p^{\mathsf{H}}\rangle}, \quad (10)$$

where $\langle \mathbf{S}, \mathbf{V}_p\mathbf{V}_p^{\mathsf{H}}\rangle - \langle \mathbf{S}, \widetilde{\mathbf{V}}_p\widetilde{\mathbf{V}}_p^{\mathsf{H}}\rangle \geq 0$ is the amount of power lost due to the mismatch between $\mathbf{V}_p$ and the estimate $\widetilde{\mathbf{V}}_p$. Note that $\Gamma_p \in [0,1]$, and the aim is to design an estimator with a $\Gamma_p$ as close to 1 as possible.

## 4 Sampling operator, Algorithms and Results

### *4.1 Coprime Subsampling Operator*

In this section, we introduce our coprime sampling scheme, which samples only $m \approx 2\sqrt{M}$ carefully selected array elements. Suppose $q_1, q_2$ are coprime numbers, i.e., $\gcd(q_1,q_2) = 1$, with $q_1q_2 \approx M$ and $q_1 \approx q_2 \approx \sqrt{M}$. Let $\mathcal{D}$ be the set of all nonnegative integer combinations of $q_1$ and $q_2$ less than or equal to $M-1$, i.e., $\mathcal{D} = \cup_{i=1,2}\{k \in [M], \mathrm{mod}(k,q_i) = 0\}$, where $[M] = \{0,1,\ldots,M-1\}$. Note that $|\mathcal{D}| \approx 2\sqrt{M}$. Since $q_1$ and $q_2$ are coprime, for sufficiently large $M$, we have $\mathcal{D} - \mathcal{D} \cong [M]$. Suppose the elements of $\mathcal{D}$ are sorted in an increasing order with $d_i \in \mathcal{D}$ being the $i$-th largest element in the list. Also, let $m = |\mathcal{D}|$ be the number of elements in $\mathcal{D}$ and let $\mathbf{B}$ be the $m \times M$ binary matrix with $\mathbf{B}_{i,d_i} = 1$ for $i \in \{1,2,\ldots,m\}$ and zero otherwise. We can simply check that $\mathbf{B}\mathbf{B}^{\mathsf{H}} = \mathbf{I}_m$. We also prove the following result, which shows the efficiency of the coprime matrix $\mathbf{B}$ for sampling Hermitian Toeplitz matrices.

**Proposition** *4.1:* Let $\mathbf{S}$ be an $M \times M$ Hermitian Toeplitz matrix and let $\mathbf{B}$ be the coprime sampling matrix. Then the mapping $\mathbf{S} \to \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}$ is a bijection. ■

### *4.2 Proposed Algorithms for Subspace Estimation*

*4.2.1* **Algorithm 1***: Approximate Maximum Likelihood (AML) Estimator:* Let $\mathbf{S} = \mathsf{snr}\int_{-\pi}^{\pi} f(u)\mathbf{a}(u)\mathbf{a}(u)^{\mathsf{H}}du$ be the covariance matrix of a user with power distribution $f(u)$. It is easy to check that for the coprime sampling matrix $\mathbf{B}$, we have

$$p(\mathbf{x}(1:T)|\mathbf{S}) = \frac{\exp\Big\{-T\,\mathsf{Tr}\Big(\widehat{\mathbf{C}}_x(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})^{-1}\Big)\Big\}}{\pi^{Tm}\mathsf{det}(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})^T}$$

where $\widehat{\mathbf{C}}_x = \frac{1}{T}\sum_{t=1}^T \mathbf{x}(t)\mathbf{x}(t)^{\mathsf{H}}$ is the sample covariance of the sketches $\mathbf{x}(t)$. The ML estimator for $\mathbf{S}$ can be written as $\mathbf{S}^* = \arg\min_{\mathbf{S}\in\mathbb{T}_+} L(\mathbf{S})$, where $\mathbb{T}_+$ is the space of PSD Toeplitz matrices and where $L(\mathbf{S})$ is given by

$$L(\mathbf{S}) = \log\mathsf{det}(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}) + \mathsf{Tr}\Big(\widehat{\mathbf{C}}_x(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})^{-1}\Big). \quad (11)$$

**Proposition** *4.2:* Let $L(\mathbf{S})$ be as in (11). Then, $L(\mathbf{S})$ is the sum of the concave function $L_{\mathsf{cav}}(\mathbf{S}) = \log\mathsf{det}(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})$ and the convex one $L_{\mathsf{vex}}(\mathbf{S}) = \mathsf{Tr}\Big(\widehat{\mathbf{C}}_x(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})^{-1}\Big)$. ■

As $L(\mathbf{S})$ is not convex, the ML estimation is generally intractable. However, since the signal covariance matrix $\mathbf{S}$ scales linearly with $\mathsf{snr}$, it is possible to obtain a convex (indeed, linear) approximation of $L_{\mathsf{cav}}(\mathbf{S})$, which is tight especially for low SNR.

**Proposition** *4.3:* Let $L_{\mathsf{cav}}(\mathbf{S}) = \log\mathsf{det}(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})$. Then, $L_{\mathsf{cav}}(\mathbf{S}) \leq \mathsf{Tr}(\mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})$ for all $\mathbf{S} \in \mathbb{T}_+$. Moreover, for low SNR, $L_{\mathsf{cav}}(\mathbf{S}) = \mathsf{Tr}(\mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}) + o(\mathsf{snr})$. ■

From Proposition 4.3, we define the AML cost function by

$$L_{\mathsf{app}}(\mathbf{S}) = \mathsf{Tr}(\mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}) + \mathsf{Tr}(\widehat{\mathbf{C}}_x(\mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}})^{-1}), \quad (12)$$

as the best convex upper bound for $L(\mathbf{S})$, which is tight for very low SNR. In particular, AML can be formulated as a semi-definite program (SDP) that can be solved efficiently.

**Proposition** *4.4:* Let $L_{\text{app}}(\mathbf{S})$ be as in (12). Suppose that $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{H}}$ is the SVD of $\widehat{\mathbf{C}}_x$ and set $\mathbf{\Delta} = \widehat{\mathbf{C}}_x^{1/2} = \mathbf{U}\mathbf{\Lambda}^{1/2}$. Then the AML estimate is obtained from the following SDP

$$(\mathbf{S}^*, \mathbf{W}^*) = \underset{\mathbf{S}\in\mathbb{T}_+, \mathbf{W}}{\arg\min}\, \mathsf{Tr}(\mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}) + \mathsf{Tr}(\mathbf{W}) \tag{13}$$

$$\text{subject to } \begin{bmatrix} \mathbf{I}_m + \mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}} & \mathbf{\Delta} \\ \mathbf{\Delta}^{\mathsf{H}} & \mathbf{W} \end{bmatrix} \succeq \mathbf{0}. \tag{14}$$

*4.2.2* **Algorithm 2***: MMV with Reduced Dimensionality (RMMV):* One of the main problems with grid-based and off-grid MMV optimizations is that their complexity increases fast with the sample size $T$. Here, we develop an SVD-based algorithm as in [9] to reduce the computational complexity and obtain an efficient algorithm. Let $\mathbf{D} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_G)]$ be a grid-based dictionary with $G$ points, and let $\mathbf{Y} = \mathbf{D}\mathbf{W} + \mathbf{N}$ be the whole data during the training period of length $T$, where $\mathbf{W} = [\mathbf{w}(1), \ldots, \mathbf{w}(T)]$ is a $G \times T$ matrix whose columns correspond to the random channel gains for different AoAs belonging to the grid. Recall that the common support or the position of nonzero channel gains in $\mathbf{w}(t)$, $t \in [T]$, corresponds to the AoAs. Let $\mathbf{X} = \mathbf{B}\mathbf{Y}$ be the subsampled signal. It is not difficult to see that subsampling still keeps the MMV format. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$ be the SVD of $\mathbf{X}$. We assume that $T \gg m = 2\sqrt{M}$, thus, we have $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}_m\mathbf{V}_m^{\mathsf{H}}$, where $\mathbf{V}_m$ denotes the $T \times m$ matrix consisting of the first $m$ columns of $\mathbf{V}$ and $\mathbf{\Sigma}_m$ is the $m \times m$ matrix consisting of only nonzero singular values. We define the new data $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_m = \mathbf{U}\mathbf{\Sigma}_m$. Notice that $\widetilde{\mathbf{X}}$ can be simply computed from the sample covariance matrix of the data $\widehat{\mathbf{C}}_x = \frac{1}{T}\mathbf{X}\mathbf{X}^{\mathsf{H}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)\mathbf{x}(t)^{\mathsf{H}}$, thus, it is not necessary to store the whole data $\mathbf{X}$ during the training time. Moreover, $\widehat{\mathbf{C}}_x$ can also be computed from $\widetilde{\mathbf{X}}$, thus, similar to Proposition 3.1, it is possible to show that $\widehat{\mathbf{C}}_x$ is a sufficient statistics for subspace estimation which implies that $\widetilde{\mathbf{X}}$ is also a sufficient statistics. We also have

$$\widetilde{\mathbf{X}} = \mathbf{B}\mathbf{D}\mathbf{W}\mathbf{V}_m + \mathbf{B}\mathbf{N}\mathbf{V}_m = \mathbf{B}\mathbf{D}\widetilde{\mathbf{W}} + \widetilde{\mathbf{N}}, \tag{15}$$

where $\widetilde{\mathbf{W}}_{G\times m}$ and $\widetilde{\mathbf{N}}_{m\times m}$ are the modified channel gains and array noises. It is not difficult to check that the reduced problem in (15) is still in the MMV format in the sense that the matrix $\widetilde{\mathbf{W}}$ has nonzero rows only on the grid points corresponding to the channel AoAs, but now the dimension of the problem is fixed and does not scale with $T$. The drawback is that $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{N}}$ lose their independence and Gaussianity since $\mathbf{V}_m$ depends on the channel gains and received noise.

Our second algorithm for subspace estimation, which is called *Reduced MMV* (RMMV), simply extends the off-grid atomic norm minimization for the MMV problem in [15, 16] to the low-dimensional data $\widetilde{\mathbf{X}}$. It can be cast as the following SDP

$$(\mathbf{S}^*, \mathbf{W}^*, \mathbf{Z}^*) = \underset{\mathbf{S}\in\mathbb{T}_+, \mathbf{W}\in\mathbb{C}^{m\times m}, \mathbf{Z}\in\mathbb{C}^{M\times m}}{\arg\min}\, \mathsf{Tr}(\mathbf{S}) + \mathsf{Tr}(\mathbf{W})$$

$$\text{subject to } \begin{bmatrix} \mathbf{S} & \mathbf{Z} \\ \mathbf{Z}^{\mathsf{H}} & \mathbf{W} \end{bmatrix} \succeq \mathbf{0},\ \|\widetilde{\mathbf{X}} - \mathbf{B}\mathbf{Z}\| \le \delta,$$

where $\delta$ is an estimate of the $\ell_2$-norm of $\widetilde{\mathbf{N}}$. For sufficiently large $m$, the optimal $\delta$ is give by $\delta^* = \sigma\sqrt{m^2} = m\sigma \approx 2\sigma\sqrt{M}$, where $\sigma^2$ is the noise variance in each array element, which can be estimated during the system's operation.

*4.2.3* **Algorithm 3***: Super Resolution (SR):* Consider a user with a power distribution $f(u)$ and let $\mathbf{S}(f)$ be its signal covariance matrix. Note that $\mathbf{S}$ is a Toeplitz matrix whose first column is given by $\mathbf{f} = \langle f, \mathbf{a}\rangle := \int f(u)\mathbf{a}(u)du \in \mathbb{C}^M$, where $[\langle f, \mathbf{a}\rangle]_k = \int f(u)e^{jk\pi u}du$ is the $k$-th Fourier coefficient of $f$. In this section, we assume that $f$ is merely a positive measure and not necessarily a normalized one. Since $\mathbf{S}$ is Toeplitz, from Proposition 4.1, it is seen that for the coprime sampling matrix $\mathbf{B}$ introduced in Section 4.1, all the elements of $\mathbf{S}$, and as a result the vector of Fourier coefficients $\mathbf{f}$ can be identified from $\mathbf{B}\mathbf{S}\mathbf{B}^{\mathsf{H}}$. This implies that for a sufficiently large $T$, we can estimate $\mathbf{f}$ accurately using the elements of the sample covariance matrix $\widehat{\mathbf{C}}_x = \mathbf{B}\widehat{\mathbf{C}}_y\mathbf{B}^{\mathsf{H}}$. Let $\mathcal{X}_k = \{(i, i') : i \ge i', d_i - d_{i'} = k\}$, where $\mathcal{D}$ and $d_i \in \mathcal{D}$ are as in Section 4.1. Let $c_k = |\mathcal{X}_k|$, and define the estimator $\widehat{\mathbf{f}}_k = \frac{\sum_{(i,i')\in\mathcal{X}_k}[\widehat{\mathbf{C}}_x]_{i,i'}}{c_k}$ for $\mathbf{f}_k$. We propose the following TV-minimization to recover the subspace of the signal from the estimates $\widehat{\mathbf{f}}$

$$\min \|f\|_{\text{TV}} \text{ subject to } \|\langle f, \mathbf{a}\rangle - \widehat{\mathbf{f}}\| \le \epsilon, \tag{16}$$

where $\epsilon$ is an estimate of the $\ell_2$-norm of the noise in the data. Since $f$ is a positive measure, $\|f\|_{\text{TV}}$ is given by $f\{[-1, 1)\} = \int_{-1}^{1} f(u)du = \mathbf{f}_0$, thus, we obtain the following result.

**Proposition** *4.5:* Consider the TV-minimization in (16). Then, (16) can be equivalently written as

$$\mathbf{S}^* = \underset{\mathbf{T}}{\arg\min}\mathsf{Tr}(\mathbf{T}) \text{ subject to } \mathbf{T} \in \mathbb{T}_+,$$

$$\|\mathbf{T}\,\mathbf{e}_1 - \widehat{\mathbf{f}}\| \lesssim \sqrt{\frac{M}{T}}(\sigma^2 + [\mathbf{T}]_{11}), \tag{17}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^{\mathsf{T}}$ is an $M \times 1$ vector, where $[\mathbf{T}]_{11}$ is the diagonal element of the Toeplitz matrix $\mathbf{T}$ (equivalent to $\mathbf{f}_0$), and where the $\sigma^2$ is an estimate of noise variance. ■

Algorithm (17) is a convex optimization that can be efficiently solved to recover the signal covariance matrix $\mathbf{S}$. In particular, no prior knowledge of SNR is necessary.

*4.2.4* **Algorithm 4***: Covariance Matrix Projection (CMP):* Let $\mathbf{B}$ be the $m \times M$ subsampling matrix as in Section 4.1, where $m = O(2\sqrt{M})$. Let $\widehat{\mathbf{C}}_x$ be the sample covariance of the subsampled signal. In order to recover the dominant $p$-dim subspace of the signal, we first find an estimate of the signal covariance matrix by

$$\mathbf{C}_y^* = \underset{\mathbf{R}\in\mathbb{T}_+}{\arg\min}\, \|\mathsf{LT}(\widehat{\mathbf{C}}_x) - \mathsf{LT}(\mathbf{B}\mathbf{R}\mathbf{B}^{\mathsf{H}})\|, \tag{18}$$

where $\mathsf{LT}$ keeps the lower-diagonal elements of $\widehat{\mathbf{C}}_x$. Then, an estimate of signal subspace is obtained from $\mathbf{C}_y^*$. The following theorem shows the resulting performance.

**Theorem** *4.6:* Consider the optimization problem (18). Then, for a given $p$ with $1 \le p \le M$, the CMP estimator recovers a $p$-dim subspace of the signal, and has a performance measure $\Gamma_p$ satisfying

$$\mathbb{E}(\Gamma_p) \ge \max\Big\{1 - \frac{2\sqrt{p}}{\delta_p\sqrt{T}}(1 + \frac{1}{\mathsf{snr}}),\, 0\Big\}, \tag{19}$$

$$\mathbb{V}\mathsf{ar}(\Gamma_p) \le \frac{4p}{T\delta_p^2}(1 + \frac{1}{\mathsf{snr}})^2, \tag{20}$$

where $\delta_p$ is defined as in (9), and where snr is the received SNR in one snapshot $t \in [T]$. ■

## 5 Simulations

In this section, we assess the performance of our proposed estimators via numerical simulations. We use the CVX package [17] for running all the convex optimizations. We assume that the AoAs are uniformly distributed in $\Theta = [40, 50] \cup [100, 110]$ with an angular spread of 20 degrees. We use an array of size $M = 80$, and a coprime sampling with $q_1 = 7, q_2 = 9$, where we sample only 19 out of 80 array elements that are located at $\mathcal{D} = \{0, 7, \ldots, 77\} \cup \{0, 9, \ldots, 72\}$.

Fig. 2 compares the performance of our proposed algorithms with PETRELS, nuclear norm minimization (NucNorm) in (7), grid-based (GBMMV) in [13], and grid-less MMV (GLMMV) in [14–16].

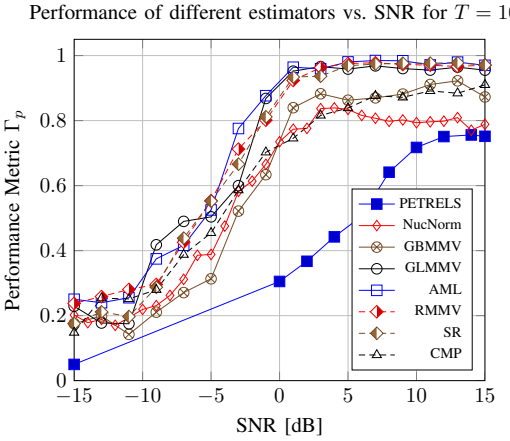



Fig. 2: Comparison of the performance of the estimators versus SNR for the training length $T = 100$. It is seen that AML, RMMV, and SR perform comparably with the GLMMV but they have lower computational complexity which does not scale with $T$. The CMP is as good as GBMMV and better than NucNorm especially for higher SNR but its complexity is much lower than GBMMV since it does not scale with $T$. PETRELS does not perform very well for the fixed data size, e.g., its performance even for $T = 800$ is worse than that of the other algorithms for $T = 100$.

Fig. 3 compares the scaling performance of our algorithms with Nuclear norm minimization for different training lengths $T$. As the performance of AML and RMMV is comparable with the GLMMV and better than GBMMV and since for large training length $T$, these algorithms are really time-consuming to run, we have not included them in this figure.

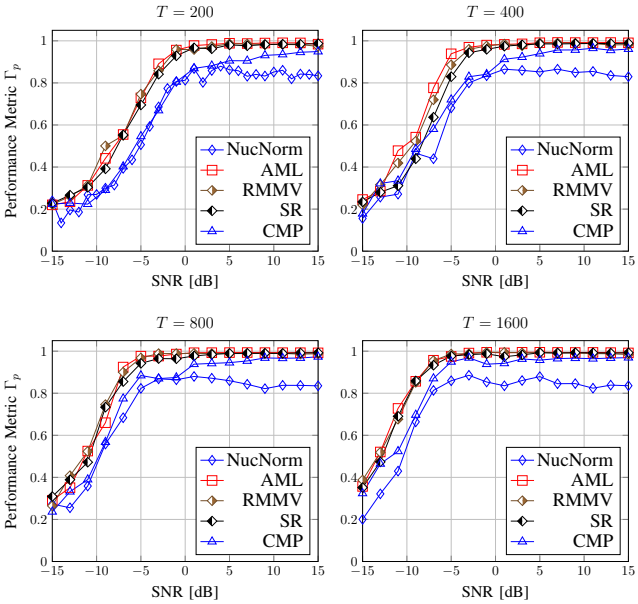



Fig. 3: Scaling of the performance of different estimators with training length $T \in \{200, 400, 800, 1600\}$

## References


[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] H. Huh, G. Caire, H. Papadopoulos, and S. Ramprashad, "Achieving massive MIMO spectral efficiency with a not-so-large number of antennas," *IEEE Trans. on Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, 2012.

[3] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexingthe large-scale array regime," *IEEE Trans. on Inform. Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[4] P. P. Vaidyanathan and P. Pal, "Sparse sensing with co-prime samplers and arrays," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 573–586, 2011.

[5] Y. Chi, Y. C. Eldar, and R. Calderbank, "Petrels: Parallel subspace estimation and tracking by recursive least squares from partial observations," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5947–5959, 2013.

[6] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 704–711.

[7] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[8] M. Herman, T. Strohmer *et al.*, "High-resolution radar via compressed sensing," *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2275–2284, 2009.

[9] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3010–3022, 2005.

[10] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[11] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *Signal Processing, IEEE Transactions on*, vol. 62, no. 21, pp. 5565–5576, 2014.

[12] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[13] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589–602, 2006.

[14] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *Information Theory, IEEE Transactions on*, vol. 59, no. 11, pp. 7465–7490, 2013.

[15] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *arXiv preprint arXiv:1408.2242*, 2014.

[16] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *arXiv preprint arXiv:1405.6585*, 2014.

[17] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.

# Factorization Approaches in Lattice-Reduction-Aided and Integer-Forcing Equalization

Robert F.H. Fischer[1], Michael Cyran[2], Sebastian Stern[1]

[1]Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany, Email: {robert.fischer,sebastian.stern}@uni-ulm.de
[2]Lehrstuhl für Informationsübertragung, Universität Erlangen-Nürnberg, Erlangen, Germany, Email: michael.cyran@fau.de

***Abstract*—In this paper, lattice-reduction-aided and integer-forcing equalization are contrasted. In both approaches, the determination of an integer matrix is essential. The different criteria for this calculation available in the literature are summarized in a unified way. A new factorization algorithm for obtaining the integer matrix is proposed. Via extensive numerical simulations the gains of the respective optimization criterion and the gain of the new algorithm over the classical Lenstra-Lenstra-Lovász algorithm are assessed. In particular, the gains achieved by dropping the constraint that the integer matrix has to be unimodular are identified.**

## I. Introduction

The design of joint receivers for signals transmitted in parallel, e.g., in multi-user uplink scenarios, is still an important topic in research. The simplest approach for handling the interference in the underlying *multiple-input/multiple-output (MIMO)* channels is to use *linear equalization* (either optimized according to the zero-forcing (ZF) or the minimum-mean squared error (MMSE) criterion). Via a (pseudo left) inverse of the channel matrix, the interference is eliminated at the cost of noise enhancement. However, as the users are perfectly decoupled, individual channel decoding can be performed and individual codes can be used. Some improvement can be gained by utilizing *decision-feedback equalization* (DFE), also known as *successive interference cancellation* (SIC) and as *Bell Laboratories space-time* (BLAST). The optimum receive strategy is *maximum-likelihood decoding*, which, however, for coded transmission has infeasible complexity.

Since more than one decade, low-complexity but well-performing approaches are of particular interest. *Lattice-reduction-aided* (LRA) techniques, e.g., [19], [16], [18], were proven to achieve the optimum diversity behavior [15]. Recently, the concept of *integer-forcing* (IF) receivers [20] was proposed. Both strategies are tightly related; the term "LRA" can be interpreted as a channel-oriented view—it puts emphasis on the mathematical tool applied to the channel matrix. In contrast, the denomination "IF" is signal-oriented—it highlights the main operation on the signals.

In this paper, a brief comparison of both approaches is given and the advantages and disadvantages of the respective procedures are enlightened. Moreover, the different criteria for selecting the integer matrix—which is central in both fields—available in the literature are summarized in a unified way. A new factorization algorithm for determining this matrix in an optimum way is presented. Using this algorithm, via extensive numerical simulations, the gains of the respective optimization criterion and the gain of the new algorithm over the classical Lenstra-Lenstra-Lovász algorithm are assessed.

The paper is organized as follows: In Sec. II the system model is introduced and in Sec. III LRA and IF strategies are contrasted and the different factorization criteria are summarized. A new factorization algorithm is presented in Sec. IV followed by numerical examples in Sec. V. Sec. VI briefly summarizes the paper.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within the framework COIN under grant FI 982/4-3.

## II. System Model

We assume a classical MIMO channel model with $K$ non-cooperating transmitters (single-antenna users) and a joint receiver with $N$ antennas. Fig. 1 shows the block diagram of the system model.

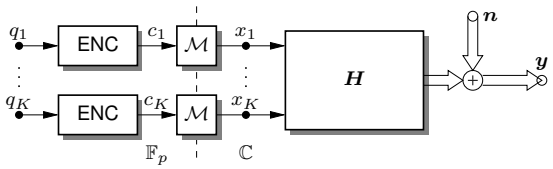



Fig. 1. System model of the MIMO communication scheme.

Each user $k$, $k = 1, \ldots, K$, wants to communicate its source symbols $q_k$ drawn from a finite field $\mathbb{F}_p$. Blocks of source symbols are encoded via some channel code; the coded symbols $c_k$ are then mapped to complex-valued transmit symbols $x_k$, drawn from some signal constellation $\mathcal{A}$. Via a suited choice of the code (including interleaving where required) and the mapping this generic model includes all types of coded modulation schemes, lattice-coding approaches, as well as uncoded transmission.

The symbols $x_k$ are then radiated over the users' antennas. Denoting the transmit vector (dimension $K$) as $\boldsymbol{x}$, the $N \times K$ channel matrix as $\boldsymbol{H}$, and the $N$-dimensional noise vector as $\boldsymbol{n}$, the receive vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \,. \tag{1}$$

The transmit symbols (per user) have variance $\sigma_x^2$ and the zero-mean Gaussian noise has variance $\sigma_n^2$ per dimension. Noteworthy, all signals and channel coefficients are complex-valued in the equivalent complex baseband domain.

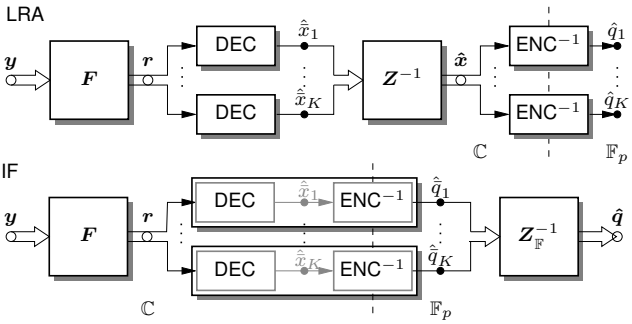



Fig. 2. System model of the receiver. Top: lattice-reduction-aided equalization; Bottom: integer-forcing receiver.

At the receiver side, the $N$ components of the receive vector $\boldsymbol{y}$ can be processed jointly in order to produce estimates of the source symbols $q_k$. To this end, some form of equalization has to take place and suited channel decoding has to be performed. In the next section, we will have a closer look at the low-complexity, well-performing LRA and IF strategies.

## III. Lattice-Reduction-Aided Equalization and Integer-Forcing Equalization

Lattice-reduction-aided (LRA) and integer-forcing (IF) receivers share the same fundamental principle. The main idea is to factorize the channel matrix as

$$\boldsymbol{H} = \boldsymbol{C}\,\boldsymbol{Z}\,. \tag{2}$$

The receive vector can then be written as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \;=\; \boldsymbol{C}\,\boldsymbol{Z}\boldsymbol{x} + \boldsymbol{n} \;=\; \boldsymbol{C}\,\bar{\boldsymbol{x}} + \boldsymbol{n}\,. \tag{3}$$

Then, not the transmit vector $\boldsymbol{x}$ itself (blocks of vectors in the coded case) is recovered but the vector $\bar{\boldsymbol{x}} \stackrel{\text{def}}{=} \boldsymbol{Z}\boldsymbol{x}$. If $\boldsymbol{Z}$ is chosen suitably, this may be done with much less noise enhancement. Taking into account that the symbols of the vector $\bar{\boldsymbol{x}}$ are correlated (due to $\boldsymbol{Z}$), the respective MMSE linear equalizer calculates to [5], [20] (inverse SNR $\zeta \stackrel{\text{def}}{=} \sigma_n^2/\sigma_x^2$)

$$\boldsymbol{F} = \left(\boldsymbol{C}^{\mathsf{H}}\boldsymbol{C} + \zeta \boldsymbol{Z}^{-\mathsf{H}}\boldsymbol{Z}^{-1}\right)^{-1}\boldsymbol{C}^{\mathsf{H}} \tag{4}$$

$$= \boldsymbol{Z}\left(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H} + \zeta \boldsymbol{I}\right)^{-1}\boldsymbol{H}^{\mathsf{H}}\,. \tag{5}$$

Hence, detection/decoding is done w.r.t. some changed basis. Having the decoding results, this basis change (the matrix $\boldsymbol{Z}$) is reversed. The LRA and IF strategies differ in the way this final step is done and how the matrix $\boldsymbol{Z}$ is chosen. The block diagrams of the respective receivers are depicted in Fig. 2.

### A. Lattice-Reduction-Aided Equalization

LRA equalization has its roots in the field of lattice reduction, i.e., the question of finding a suited basis for a given lattice; here the lattice spanned by the columns of the channel matrix $\boldsymbol{H}$. Consequently, $\boldsymbol{Z}$ is chosen as an integer unimodular matrix. In the complex case, the coefficients of $\boldsymbol{Z}$ are drawn from the *Gaussian integers* $\mathbb{G} \stackrel{\text{def}}{=} \mathbb{Z} + \mathrm{j}\mathbb{Z}$ and $|\det(\boldsymbol{Z})| = 1$, such that $\boldsymbol{Z}^{-1}$ is also an (complex) integer unimodular matrix. Using lattice-reduction algorithms, most prominently the LLL algorithm [10] or its complex-valued generalization [6], a solution can readily be found.

Decoding and resolution of the interference via $\boldsymbol{Z}^{-1}$ is done over the complex numbers; the linear combinations $\bar{\boldsymbol{x}}$ of the transmit symbols have to be estimated by the decoders. Then, an estimate of the transmit symbols is obtained via $\hat{\boldsymbol{x}} = \boldsymbol{Z}^{-1}\hat{\bar{\boldsymbol{x}}}$. Finally, via the encoder inverses, estimates $\hat{q}_k$ of the source symbols are obtained.

LRA equalization only[1] works if the signal constellation $\mathcal{A}$ is a subset[2] of $\mathbb{G}$, i.e., $\boldsymbol{x} \in \mathbb{G}^K$, such that any (complex) integer linear combination of the points is again drawn from $\mathbb{G}$. Moreover, in the coded case, the codes have to be linear, such that any (complex) integer linear combination of codewords is again a valid codeword. It is true that in the vast majority of the literature, LRA equalization is treated uncoded. This, however, is justified as equalization and decoding can simply be cascaded; coding can straightforwardly be put on top of the uncoded LRA scheme. No further specific restrictions have to be obeyed.

### B. Integer-Forcing Equalization

Recently, originating from *compute-and-forward* relaying schemes [11], an integer-forcing linear equalization scheme was proposed in [20]. The main difference, see Fig. 2, is that the integer interference is resolved over the finite field rather than over the complex numbers. To this end, linear combinations $\bar{q}_k$ of the source symbols are delivered by the decoders and the integer matrix is inverted over $\mathbb{F}_p$. Put simply, the order of encoder inverse and inverse of $\boldsymbol{Z}$ is reversed.

However, this imposes much stronger constraints on the codes and the mapping as in the LRA case. Basically, arithmetic over the complex numbers (modulo $p$) has to be isomorphic to the arithmetic of the finite field $\mathbb{F}_p$. In the simplest version this is achieved by restricting to real-valued signaling and $\mathcal{A}$ is a one-dimensional $p$-ary constellation where $p$ is a prime. Generalization to complex-valued Gaussian prime constellations [9] or other algebraic structures [4] is possible.

Since the integer interference is resolved over the finite field, the matrix $\boldsymbol{Z}$ has to be invertible over $\mathbb{F}_p$. Since $p$ is a prime this is possible as long as $\boldsymbol{Z}$ has full rank; no restriction on the determinant is required. This gives rise to a new factorization problem: not a *shortest basis problem* as in LRA has to be solved but a *shortest independent vector problem* [20].

### C. Comparison

Even though LRA and IF are tightly related, the constraints and restrictions are different. IF imposes strong constraints on the signal constellation and its cardinality and in turn on the applicable codes. In LRA only linearity in signal space is required. Contrary, here unimodularity of $\boldsymbol{Z}$ is forced.

[1] Generalization to other lattices, e.g., the Eisenstein integers [2], are possible. In each case, the signal constellation and the entries of $\boldsymbol{Z}$ have to be taken from the same lattice/algebraic structure, cf. [4].

[2] If an offset is present as in usual QAM constellations, LRA equalization still works if this offset is adequately taken into account, e.g., [17].

The presentation of the IF schemes has sparked a rethinking of the LRA approach—indeed, unimodularity is not required. If[3] $|\det(\boldsymbol{Z})| > 1$ the vector $\bar{\boldsymbol{x}} = \boldsymbol{Z}\boldsymbol{x}$, with $\boldsymbol{x} \in \mathbb{G}^K$, is not taken from $\mathbb{G}^K$ but a sublattice thereof.[4] Given the points from this sublattice, $\boldsymbol{Z}^{-1}$—which has a determinant smaller than one—will recover the original transmit vector $\boldsymbol{x}$. Hence, the LRA equalizer structure can be used with any full-rank integer matrix $\boldsymbol{Z}$, enabling the same gains as in IF but without the restrictions on the signal constellation and the codes.

IF schemes have their main justification not in central but in decentralized receivers. In a distributed antenna system, the partial equalization via $\boldsymbol{F}$ cannot be applied; the residual interference is taken as it is and the decoders produce estimates on linear combinations. In IF schemes, only symbols from $\mathbb{F}_p$ have to be communicated over the backhaul. The integer interference is resolved in some central processing unit. Conversely, using the LRA structure, complex numbers would have to be sent. In a central receiver the LRA structure is preferable.

In summary, LRA and IF have its individual advantages and constraints. However, the calculation of the integer matrix can be done in the same way for both approaches. For that we have to distinguish between the different *criteria* the optimization is based on and between different *factorization algorithms*.

### D. Factorization Criteria

We now give an overview on the different criteria the factorization task (2) is usually based on.

C-I *Based on $\boldsymbol{H}$:* In the initial publications [19], [16], lattice reduction is directly applied to the channel matrix $\boldsymbol{H}$

$$\boldsymbol{H} = \boldsymbol{C}_{\mathrm{I}}\,\boldsymbol{Z}_{\mathrm{I}}\,. \tag{6}$$

Any lattice reduction algorithm may be used, e.g., minimizing the orthogonality defect of $\boldsymbol{C}$.

C-II *Based on $\boldsymbol{H}^{-\mathsf{H}}$:* In [15], the factorization

$$\boldsymbol{H}^{-\mathsf{H}} = \boldsymbol{F}_{\mathrm{II}}^{\mathsf{H}}\,\boldsymbol{Z}_{\mathrm{II}}^{-\mathsf{H}} \tag{7}$$

has been proposed. As for square matrices $\boldsymbol{F}^{\mathsf{H}} = \boldsymbol{C}^{-\mathsf{H}} = \boldsymbol{H}^{-\mathsf{H}}\,\boldsymbol{Z}^{\mathsf{H}}$ follows from (6), $\boldsymbol{F}$ is immediately the (ZF) equalization matrix and $\boldsymbol{Z}$ is the required integer matrix. Here, lattice reduction is applied to $\boldsymbol{H}^{-\mathsf{H}}$ instead of $\boldsymbol{H}$ (for non-square channel matrices the Hermitian of the left pseudoinverse has to be used). Since the squared lengths of the columns of $\boldsymbol{F}^{\mathsf{H}}$ give the noise enhancement (in case of ZF linear equalization), this criterion directly optimizes the performance of the scheme instead of a substitute measure as above.

C-III *Based on $\boldsymbol{\mathcal{H}}$:* In [18], an MMSE version to LRA equalization has been given. The main idea is to calculate the ZF solution for the augmented[5] channel matrix; the result is exactly the MMSE solution. The factorization here reads

$$\begin{bmatrix} \boldsymbol{H} \\ \sqrt{\zeta}\boldsymbol{I} \end{bmatrix} \overset{\text{def}}{=} \boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{C}}_{\mathrm{III}}\,\boldsymbol{Z}_{\mathrm{III}} = \begin{bmatrix} \boldsymbol{C}_{\mathrm{III}} \\ \sqrt{\zeta}\boldsymbol{Z}_{\mathrm{III}}^{-1} \end{bmatrix} \boldsymbol{Z}_{\mathrm{III}}\,. \tag{8}$$

[3] As $\boldsymbol{Z} \in \mathbb{G}^{K\times K}$, $|\det(\boldsymbol{Z})| < 1$ is not possible for full-rank matrices.

[4] The individual decoding/detection of the components of $\bar{\boldsymbol{x}}$ is suboptimal, as non-valid points can be delivered. This is anyway the case as the actual boundary region of the constellation cannot be taken into account in separate decoding, cf. [17]. For sufficiently large SNR this fact is irrelevant.

[5] Augmented matrices are typeset in calligraphic font.

TABLE I
OVERVIEW ON FACTORIZATION STRATEGIES.

| based on | channel matrix $\boldsymbol{H}$ ("ZF solution") | augmented matrix $\boldsymbol{\mathcal{H}}$ ("MMSE solution") |
|---|---|---|
| $\boldsymbol{H}$ | $\boldsymbol{H} = \boldsymbol{C}\,\boldsymbol{Z}$ | $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{C}}\,\boldsymbol{Z}$ |
| $(\boldsymbol{H}^{+_{\mathrm{l}}})^{\mathsf{H}}$ | $(\boldsymbol{H}^{+_{\mathrm{l}}})^{\mathsf{H}} = \boldsymbol{F}^{\mathsf{H}}\,\boldsymbol{Z}^{-\mathsf{H}}$ | $(\boldsymbol{\mathcal{H}}^{+_{\mathrm{l}}})^{\mathsf{H}} = \boldsymbol{\mathcal{F}}^{\mathsf{H}}\,\boldsymbol{Z}^{-\mathsf{H}}$ |

Interestingly, the left pseudoinverse[6] $\boldsymbol{\mathcal{C}}^{+_{\mathrm{l}}}$ of $\boldsymbol{\mathcal{C}}$, immediately gives the augmented receive matrix, as a comparison with (4) shows [5].

C-IV *Based on $(\boldsymbol{\mathcal{H}}^{+_{\mathrm{l}}})^{\mathsf{H}}$:* In [20] a criterion for directly minimizing the noise variance after MMSE linear equalization of the part $\boldsymbol{C}$ has been given. With $\boldsymbol{Z}^{\mathsf{H}} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K]$ it reads

$$\boldsymbol{Z}_{\mathrm{IV}}^{\mathsf{H}} = \underset{\substack{\boldsymbol{Z}\in\mathbb{G}^{K\times K},\\ \operatorname{rank}(\boldsymbol{Z})=K}}{\operatorname{argmin}} \max_{m} ||\boldsymbol{L}^{\mathsf{H}}\boldsymbol{z}_m||^2\,, \tag{9}$$

where $\boldsymbol{L}\boldsymbol{L}^{\mathsf{H}} = \left(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H} + \zeta\boldsymbol{I}\right)^{-1} = \left(\boldsymbol{\mathcal{H}}^{\mathsf{H}}\boldsymbol{\mathcal{H}}\right)^{-1}$. (10)

Since $\boldsymbol{L}$ can be any "square root" of the right-hand-side matrix, as straightforward calculations show, we can set

$$\boldsymbol{L}^{\mathsf{H}} = \left(\boldsymbol{\mathcal{H}}^{+_{\mathrm{l}}}\right)^{\mathsf{H}} = \left(\boldsymbol{\mathcal{H}}^{\mathsf{H}}\right)^{+_{\mathrm{r}}} \tag{11}$$

and the respective factorization task can thus be written as

$$\left(\boldsymbol{\mathcal{H}}^{+_{\mathrm{l}}}\right)^{\mathsf{H}} = \boldsymbol{\mathcal{F}}_{\mathrm{IV}}^{\mathsf{H}}\,\boldsymbol{Z}_{\mathrm{IV}}^{-\mathsf{H}}\,. \tag{12}$$

Noteworthy, for all optimization criteria a respective factorization task[7] can be stated in which $\boldsymbol{Z}$ has to be chosen such that the squared lengths of the column of the matrix $\boldsymbol{C}_{\mathrm{I}}$, $\boldsymbol{F}_{\mathrm{II}}^{\mathsf{H}}$, $\boldsymbol{\mathcal{C}}_{\mathrm{III}}$, or $\boldsymbol{\mathcal{F}}_{\mathrm{IV}}^{\mathsf{H}}$, respectively, are as short as possible.

Table I gives an overview on the different criteria for the factorization problem.

## IV. Factorization Algorithm

The above overview has shown that regardless which optimization criterion is used, a factorization problem has to be solved in order to obtain the required integer matrix $\boldsymbol{Z}$. If we follow the original LRA approach and restrict $\boldsymbol{Z}$ to be unimodular, any *lattice reduction algorithm*, in particular the LLL algorithm [10], can be used. For the complex-valued setting at hand, the CLLL [6] may be applied.

If the unimodularity is dropped, an algorithm for solving the *shortest independent vector problem (SIVP)* has to be applied. Unfortunately, in the literature, only a few approaches are available. In [20], the optimization problem (9) or (12) is solved via a brute-force search with some restrictions to the search space. In [13], [14], low-complexity factorization approaches, all directly based on the CLLL and hence resulting in a unimodular matrix, are given. An algorithm to solve the successive minima problem has been published in [3]. For the distributed antenna setting, in [8] a suited factorization is given, taking into account that no joint feedforward equalization via $\boldsymbol{F}$ is possible.

[6] $\boldsymbol{A}^{+_{\mathrm{l}}} = (\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{H}}$ denotes the left pseudoinverse of $\boldsymbol{A}$ and $\boldsymbol{A}^{+_{\mathrm{r}}} = \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}})^{-1}$ the right pseudoinverse. $\boldsymbol{A}^{-\mathsf{H}} = (\boldsymbol{A}^{\mathsf{H}})^{-1} = (\boldsymbol{A}^{-1})^{\mathsf{H}}$.

[7] We hence denote the corresponding procedure *factorization algorithm*.

**Alg. 1** Pseudocode of the factorization algorithm.

**function** $\boldsymbol{Z} = \texttt{factorize}(\boldsymbol{H}, \zeta)$

1 $\boldsymbol{G} = \left(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H} + \zeta \boldsymbol{I}\right)^{-1/2}$ // generator matrix
2 $[\boldsymbol{G}_{\mathrm{LLL}}, \boldsymbol{Z}_{\mathrm{LLL}}] = \mathrm{LLL}(\boldsymbol{G})$ // reduced basis
3 $R^2_{\max} = \max_i ||\boldsymbol{g}_{i,\mathrm{LLL}}||^2$ // search radius
4 $\boldsymbol{U} = \texttt{getlist}(\boldsymbol{G}_{\mathrm{LLL}}, R^2_{\max})$ // get list of short vectors
5 $\{i_1, \ldots, i_K\} = \texttt{getindices}(\boldsymbol{U})$ // indices of lin. indep. vectors
6 $\boldsymbol{Z} = \boldsymbol{Z}_{\mathrm{LLL}} \cdot \boldsymbol{U}(:, [i_1, \ldots, i_K])$ // integer matrix

We now present an algorithm which is feasible for MIMO scenarios typically of interest; a pseudocode description is given in Alg. 1. Via numerical simulations we can then study the gain possible by the respective criteria and the loss when restricting $\boldsymbol{Z}$ to be unimodular. To have a compact notation, we rewrite (10), (12) as

$$\boldsymbol{G}_{\mathrm{opt}} = \boldsymbol{G}\,\boldsymbol{Z}^{\mathsf{H}}\,, \tag{13}$$

with $\boldsymbol{G} = \boldsymbol{L}^{\mathsf{H}}$, $\boldsymbol{Z}^{\mathsf{H}} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K] \in \mathbb{G}^{K\times K}$, $\mathrm{rank}(\boldsymbol{Z}) = K$, and the columns of $\boldsymbol{G}_{\mathrm{opt}}$ as short as possible. This means that given the basis $\boldsymbol{G}$ of a lattice, find $K$ linearly independent vectors (lattice points) $\boldsymbol{G}\boldsymbol{u}_i$, $\boldsymbol{u}_i \in \mathbb{G}^K$, which are as short as possible. We do this via performing the following steps:

*LLL Reduction:*

First, the LLL reduced basis $\boldsymbol{G}_{\mathrm{LLL}} = \left[\boldsymbol{g}_{1,\mathrm{LLL}}, \ldots, \boldsymbol{g}_{K,\mathrm{LLL}}\right]$ is calculated. Since the SIVP is more relaxed than the shortest basis problem, the LLL basis gives an upper bound $R^2_{\max} = \max_i ||\boldsymbol{g}_{i,\mathrm{LLL}}||^2$ on the norms of the vectors possible in the SIVP. We denote this step as $[\boldsymbol{G}_{\mathrm{LLL}}, \boldsymbol{Z}_{\mathrm{LLL}}] = \mathrm{LLL}(\boldsymbol{G})$.

*List of Lattice Points:*

Then, a (sorted) list of vectors (lattice points) with squared norms bounded by $R^2_{\max}$ is calculated. Let the list be written as matrix $\boldsymbol{U} = \left[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell\right]$, with $||\boldsymbol{G}_{\mathrm{LLL}}\boldsymbol{u}_i||^2 \leq ||\boldsymbol{G}_{\mathrm{LLL}}\boldsymbol{u}_{i+1}||^2$, $\forall i$. Since for complex lattices the volume is given by the squared magnitude of the determinant of the generator matrix [2, Eq. (87)]), the list size can be approximated by $\ell = (\pi R^2_{\max})^K/(K!|\det(\boldsymbol{G})|^2)$. We denote this step as $\boldsymbol{U} = \texttt{getlist}(\boldsymbol{G}_{\mathrm{LLL}}, R^2_{\max})$.

This step can be implemented efficiently using the idea of the list sphere decoder, Alg. AllClosestPoints in [1]. In principle, this calculation has exponential complexity, however, if the LLL basis is used and $R^2_{\max}$ is small and, thus, the number $\ell$ of points within the search sphere is small, still an efficient search is obtained.

*Select Points:*

Among the vectors in the list (matrix $\boldsymbol{U}$) the best combination of vectors, i.e., indices $i_1, \ldots, i_K$, has to be found such that $\boldsymbol{Z} = \left[\boldsymbol{u}_{i_1}, \ldots, \boldsymbol{u}_{i_K}\right]$ has full rank and $\boldsymbol{G}_{\mathrm{LLL}}\boldsymbol{u}_{i_K}$ is as small as possible.

The last step can be solved by performing Gaussian elimination on the matrix $\boldsymbol{U}$, i.e., transforming it to *row echelon form*. Since the list is sorted according to increasing (squared) norms $||\boldsymbol{G}_{\mathrm{LLL}}\boldsymbol{u}_i||^2$, the best choice is to select the vectors $\boldsymbol{u}_i$, which first define a new dimension; in row echelon form these are the vectors at the steps. We denote this step as $\{i_1, \ldots, i_K\} = \texttt{getindices}(\boldsymbol{U})$.

Please note, if some restrictions (e.g., on the determinant) of $\boldsymbol{Z}$ have to be obeyed, a search over combinations of candidates can be performed instead of the simple Gaussian elimination. This step can efficiently be implemented by the sphere decoder and offers degrees of freedom not present in other algorithms.

## V. Numerical Results

The factorization algorithm has been implemented and extensive numerical simulations have been performed. Thereby, $\boldsymbol{H}$ is an $N \times K$ i.i.d. random zero-mean unit-variance complex Gaussian matrix. The aim is to assess which gains can be attributed to which factorization criterion or algorithm. The proposed straight-forward algorithm gives the same results as the recent one in [3]. For $K$ up to 8 our strategy is faster for most of the realizations; however for a few matrices it requires significant higher complexity. A detailed complexity evaluation is beyond the scope of the present paper.

First, in Fig. 3 the cumulative distribution function of the list size (number of columns in $\boldsymbol{U}$) is plotted. Please note, all apparently linearly dependent vectors (those multiplied by $-1$, j, and $-$j) are not added to the list in `getlist`. It can be seen that for practical values of $K$ the list size is small to moderate and can easily be handled. The chosen SNR is almost the worst case; for large SNR $\boldsymbol{L}\boldsymbol{L}^{\mathsf{H}} \approx \left(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H}\right)^{-1}$. For small SNRs smaller list sizes are obtained as $\boldsymbol{L}\boldsymbol{L}^{\mathsf{H}} \approx \frac{\sigma_x^2}{\sigma_n^2}\boldsymbol{I}$ and thus $\boldsymbol{Z} = \boldsymbol{I}$ is optimum.

Second, Tab. II summarizes the distribution of the determinant of $\boldsymbol{Z}$ over $10^6$ channel realizations. As the dimension of the channel matrix increases, the number of channels where

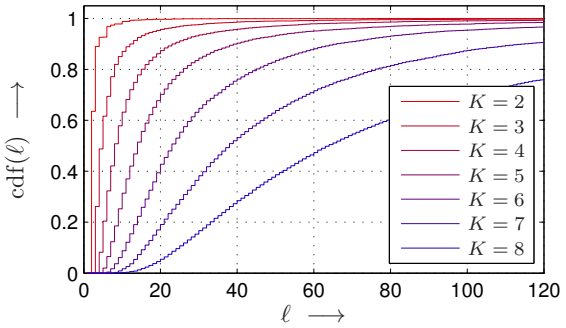



Fig. 3. Cumulative distribution function (cdf) of the list size. I.i.d. channel matrix $\boldsymbol{H}$ with $K = N$. $\sigma_x^2/\sigma_n^2 \mathrel{\widehat{=}} 20$ dB.

TABLE II
DISTRIBUTION OF $|\det(\boldsymbol{Z})|$. $\sigma_x^2/\sigma_n^2 \mathrel{\widehat{=}} 20$ dB.

| $\lvert\det(\boldsymbol{Z})\rvert =$ | 1 | $\sqrt{2}$ | 2 | $\sqrt{5}$ |
|---|---|---|---|---|
| $K = N = 2$ | 100 % | — | — | — |
| $K = N = 3$ | 99.8 % | 0.2 % | — | — |
| $K = N = 4$ | 99.0 % | 1.0 % | — | — |
| $K = N = 5$ | 97.5 % | 2.4 % | 0.005 % | — |
| $K = N = 6$ | 95.4 % | 4.5 % | 0.03 % | 0.003 % |
| $K = N = 7$ | 92.7 % | 7.1 % | 0.15 % | 0.02 % |
| $K = N = 8$ | 89.3 % | 10.2 % | 0.39 % | 0.06 % |

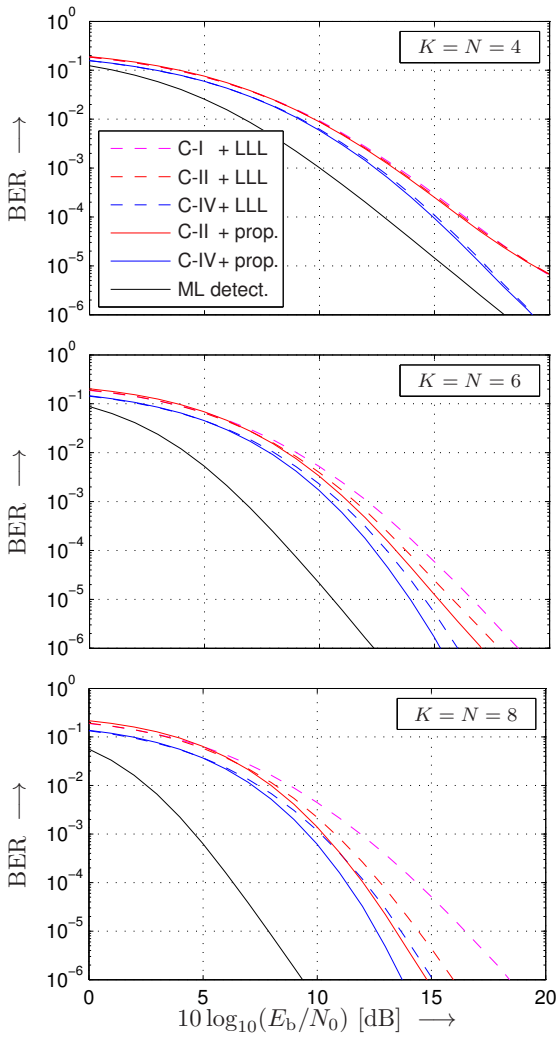



Fig. 4. Bit error rate over the SNR for uncoded transmission. 16QAM transmission. Variation of the optimization criterion (colors) and the factorization algorithm (dashed vs. solid).

$|\det(\boldsymbol{Z})| > 1$ increases. However, only for $K > 6$ non-unimodular matrices are optimum for a significant portion of channels.

Finally, bit-error-rate curves for uncoded transmission are depicted in Fig. 4. 16QAM signaling is used and the SNR is normalized to $\frac{E_b}{N_0} = \frac{\sigma_a^2}{\sigma_n^2 \log_2(16)}$. The factorization criterion and the factorization algorithm are varied; in each case the linear receiver frontend $\boldsymbol{F}$ is adjusted according to the MMSE criterion. For reference, ML detection is included.

Obviously, C-I together with the LLL algorithm has the worst performance. Using C-II gives better results (cf. [15]), best performance is obtained when applying C-IV; still the LLL is used, thus $\boldsymbol{Z}$ is unimodular. Using the proposed algorithm which relaxes the constraint on the determinant of $\boldsymbol{Z}$ some additional gain is possible. This gain, as already can be deduced from Tab. II, increases when $K$ gets larger. Compared to classical LRA equalization using C-I and the LLL, gains in the order of 5 dB are possible for $K = N = 8$ by replacing the criterion and the factorization algorithm. Thereby, however, the LRA receiver structure can be utilized as it is—the constraints on the constellation and the code design in IF can be avoided.

## VI. Summary and Conclusions

The tight relation between LRA and IF schemes has been highlighted and a new, optimum factorization algorithm has been proposed. We have restricted ourselves to linear equalization of the residual part. The extension to successive equalization and decoding (DFE/SIC, cf. [12]) is immediately possible. Moreover, the transformation to transmitter-side precoding, dual to receiver-side equalization, is also directly possible, cf. [8], [7].

## References


[1] E. Agrell, T. Eriksson, A. Vardy, K. Zeger. Closest Point Search in Lattices. *IEEE Tr. Information Theory*, pp. 2201–2214, Aug. 2002.

[2] J.H. Conway, N.J.A. Sloane. *Sphere Packings, Lattices and Groups*. Springer Verlag, New York, Berlin, 3rd edition, 1999.

[3] L. Ding, K. Kansanen, Y. Wang, J. Zhang. Exact SMP Algorithms for Integer Forcing Linear MIMO Receivers. *IEEE Tr. Wireless Communications*, pp. 6955–6966, Dec. 2015.

[4] C. Feng. D. Silva, F.R. Kschischang. An Algebraic Approach to Physical-Layer Network Coding. *IEEE Tr. Information Theory*, pp. 7576–7596, Nov. 2013.

[5] R.F.H. Fischer, C. Windpassinger, C. Stierstorfer, C. Siegl, A. Schenk, Ü. Abay. Lattice-Reduction-Aided MMSE Equalization and the Successive Estimation of Correlated Data. *AEÜ—Int. Journal of Electronics and Communications*, pp. 688–693, Aug. 2011.

[6] Y.H. Gan, C. Ling, W.H. Mow. Complex Lattice Reduction Algorithm for Low-Complexity Full-Diversity MIMO Detection. *IEEE Tr. Signal Processing*, July 2009.

[7] W. He, B. Nazer, S. Shamai. Uplink-Downlink Duality for Integer-Forcing. *IEEE Int. Symp. Information Theory*, pp. 2544–2548, 2014.

[8] S.-N. Hong, G. Caire. Compute-and-Forward Strategies for Cooperative Distributed Antenna Systems. *IEEE Tr. Information Theory*, pp. 5227–5243, Sept. 2013.

[9] K. Huber. Codes over Gaussian Integers. *IEEE Tr. Information Theory*, pp. 207–216, Jan. 1994.

[10] A.K. Lenstra, H.W. Lenstra, L. Lovász. Factoring Polynomials with Rational Coefficients, *Math. Annalen*, pp. 515–534, 1982.

[11] B. Nazer, M. Gastpar. Compute-and-Forward: Harnessing Interference Through Structured Codes. *IEEE Tr. Information Theory*, pp. 6463–6486, Oct. 2011.

[12] O. Ordentlich, U. Erez, B. Nazer. Successive Integer-Forcing and its Sum-Rate Optimality. *Allerton Conference*, pp. 282–292, 2013.

[13] A. Sakzad, J. Harshan, E. Viterbo. Integer-Forcing MIMO Linear Receivers Based on Lattice Reduction. *IEEE Tr. Wireless Communications*, pp. 4905–4915, Oct. 2013.

[14] A. Sakzad, J. Harshan, E. Viterbo. On Complex LLL Algorithm for Integer Forcing Linear Receivers. *Australian Comm. Theory Workshop*, pp. 13–17, 2013.

[15] M. Taherzadeh, A. Mobasher, A.K. Khandani. LLL Reduction Achieves the Receive Diversity in MIMO Decoding. *IEEE Tr. Information Theory*, pp. 4801–4805, Dec. 2007.

[16] C. Windpassinger, R.F.H. Fischer. Low-Complexity Near-Maximum-Likelihood Detection and Precoding for MIMO Systems using Lattice Reduction. *IEEE Information Theory Workshop*, pp. 345–348, 2003.

[17] C. Windpassinger. *Detection and Precoding for Multiple Input Multiple Output Channels*. Dissertation, Erlangen, June 2004.

[18] D. Wübben, R. Böhnke, V. Kühn, K.D. Kammeyer. Near-Maximum-Likelihood Detection of MIMO Systems using MMSE-Based Lattice Reduction. *IEEE Int. Conf. Communications*, pp. 798–802, 2004.

[19] H. Yao, G. Wornell. Lattice-Reduction-Aided Detectors for MIMO Communication Systems. *IEEE Glob. Telec. Conf.*, pp. 424–428, 2002.

[20] J. Zhan, B. Nazer, U. Erez, M. Gastpar. Integer-Forcing Linear Receivers. *IEEE Tr. Information Theory*, pp. 7661–7685, Dec. 2014.

# RECENT RESULTS ON BROADCAST NETWORKS WITH LAYERED DECODING AND SECRECY: AN OVERVIEW

*Shaofeng Zou*[⋆] *Yingbin Liang* [⋆] *Lifeng Lai*[‡] *H. Vincent Poor* [†] *Shlomo Shamai (Shitz)* [♯]

[⋆] Syracuse University [‡] Worcester Poly Insistitute [†] Princeton University [♯] Technion
Email: szou02@syr.edu, yliang06@syr.edu, llai@wpi.edu, poor@princeton.edu, sshlomo@ee.technion.ac.il

## ABSTRACT

Recent information theoretic results on a class of broadcast channels with layered decoding and/or layered secrecy are overviewed. Designs for different models are compared and applications of these results to fading wiretap channels and secret sharing are briefly discussed. An outlook, focusing on theoretical challenges concludes the overview.

## 1. INFORMATION THEORETIC MODELS

We briefly introduce four models all belonging to the class of degraded broadcast channels with layered decoding and/or layered secrecy. More details can be found in [1].

The first model is the degraded broadcast channel with layered decoding and non-layered secrecy, in which a transmitter sends $K$ messages $W_1, \ldots, W_K$ to $K$ receivers with each receiver $k$, decoding the first $k$ messages, and the eavesdropper ignorant of all messages. The capacity achieving scheme superposes multiple layers together with each layer carrying one more message than its previous layer. Furthermore, each layer applies random binning to secure not only the message in this layer but also all higher-layer messages.

The second model is the degraded broadcast channel with non-layered decoding and layered secrecy, in which a transmitter sends $K$ messages $W_1, \ldots, W_K$ to one legitimate receiver, and each eavesdropper $k$, needs to be kept ignorant of the messages $W_k, \ldots, W_K$, for $k = 1, \ldots, K$. The capacity achieving scheme encodes each codeword with multiple messages so that lower-layer messages can serve as a random source to protect higher-layer messages.

The third model is the degraded broadcast channel with layered decoding and layered secrecy, in which a transmitter sends $K$ messages $W_1, \ldots, W_K$ to $K$ receivers. Receiver $k$ is required to decode the first $k$ messages $W_1, \ldots, W_k$, and is kept ignorant of messages $W_{k+1}, \ldots, W_K$. The capacity achieving scheme is similar to that for the first model except that random binning within one layer only protects the message corresponding to the same layer.

The fourth model is the degraded broadcast channel with layered decoding and layered secrecy and with secrecy outside a bounded range. We focus on the case in which the transmitter has four messages $W_1, \ldots, W_4$ intended for the four receivers. Receiver $k$ is required to decode the messages $W_1, \ldots, W_k$. Furthermore, $W_3$ needs to be kept secure from receiver 1, and $W_4$ needs to be kept secure from receivers 1 and 2. Hence, each message is secured from a receiver with two-level worse channel quality. The capacity achieving scheme applies the joint design of superposition, embedded coding, random binning, and rate splitting and sharing.

## 2. APPLICATIONS

We discuss two applications of the broadcast models described in Section 1. The first application is to the fading wiretap channel, in which the legitimate and eavesdropping channels are corrupted by multiplicative random fading gains. In the case that the transmitter does not know the fading gains, the legitimate and eavesdropping channels can be viewed as having multiple states. A layered transmission scheme can be designed so that more layers can be decoded if the legitimate channel has better quality, and more layers can be made secure if the eavesdropper channel has lower quality. Thus, such an approach naturally yields a degraded broadcast channel with layered decoding and secrecy requirements as discussed in Section 1, and the secrecy capacity results for such models can be applied.

The second application is to the secret sharing problem with multiple secrets, which can be shown to be equivalent to the broadcast channel with secrecy requirements. Namely, the groups of participants that are required to determine secrets should be viewed as legitimate receivers and the groups of participants that are required to be ignorant of secrets should be viewed as eavesdroppers.

## 3. REFERENCES

[1] S. Zou, Y. Liang, L. Lai, H. V. Poor, and S. Shamai, "Broadcast networks with layered decoding and layered secrecy: Theory and applications," *Proceedings of the IEEE*, vol. 103, no. 10, pp. 1841–1856, Oct 2015.

The work of S. Zou and Y. Liang was supported by a National Science Foundation CAREER Award under Grant CCF-10-26565 and by the National Science Foundation under Grant CNS-11-16932. The work of L. Lai was supported by the National Science Foundation under Grant CCF-1318980 and Grant CNS-1457076. The work of H. V. Poor was supported in part by the National Science Foundation under Grant CMMI-1435778 and Grant CNS-1456793. The work of S. Shamai (Shitz) was supported in part by the Israel Science Foundation (ISF).

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Semantic Security using a Stronger Soft-Covering Lemma

Paul Cuff, Ziv Goldfeld, and Haim Permuter

## I. Soft Covering

A soft-covering theorem was introduced by Wyner [1, Theorem 6.3] and is the central analysis step for achievability proofs of information theoretic security, resolvability, and channel synthesis. It can also be used for simple achievability proofs in lossy source coding. Recently in [2] we have sharpened the claim of soft-covering by moving away from an expected value analysis. Instead, a random codebook is shown to achieve the soft-covering phenomenon with high probability. The probability of failure is doubly-exponentially small in the block-length, enabling many applications through the union bound. In particular, it can be used to achieve semantic security in wiretap channels without loss of communication rate efficiency, as we demonstrate in [3].

The soft-covering concept says that the distribution induced by selecting an $X^n$ sequence at random from a codebook of sequences and passing it through the memoryless channel $Q_{Y^n|X^n}$ will be a good approximation of $Q_{Y^n}$ in the limit of large $n$ as long as the codebook is of size greater than $2^{nR}$ where $R > I(X;Y)$. The codebook can even be constructed randomly, drawing each sequence independently from the distribution $Q_{X^n}$.

The soft-covering lemmas in the literature claim that the distance (commonly total variation or relative entropy) between the induced distribution $P_{Y^n}$ and the desired distribution $Q_{Y^n}$ vanishes in expectation over the random selection of the codebook. This phenomenon has been studied and refined numerous times in the literature (see references in [2]). Sometimes it is referred to as "resolvability" or simply as a covering lemma. But always expected distance is analyzed.

In [2] we give a stronger claim. With high probability with respect to the codebook construction, the distance will vanish exponentially quickly with the block-length $n$. The negligible probability of the random set not producing this desired result is doubly-exponentially small.

Let us define precisely the induced distribution. Let $\mathcal{C} = \{u^n(m)\}_{m=1}^M$ be the set of sequences, which will be referred to as the codebook. The size of the codebook is $M = 2^{nR}$. Then the induced distribution is:

$$P_{V^n|\mathcal{C}} = 2^{-nR} \sum_{u^n(m)\in\mathcal{C}} Q_{V^n|U^n=u^n(m)}. \tag{1}$$

**Lemma 1** ([2])**.** *For any $Q_U$, $Q_{V|U}$, and $R > I(U;V)$, where $V$ has a finite support $\mathcal{V}$, there exists a $\gamma_1 > 0$ and a $\gamma_2 > 0$ such that for $n$ large enough*

$$\mathbf{P}\left(d(P_{V^n|\mathcal{C}}, Q_{V^n}) > e^{-\gamma_1 n}\right) \le e^{-e^{\gamma_2 n}}, \tag{2}$$

*where $d(\cdot,\cdot)$ is the relative entropy.*

## II. Semantic Security

Wyner's soft-covering lemma has become a standard tool for proving that strong perfect secrecy is achieved in the wiretap channel (see e.g. [4]). Coincidentally, Wyner introduced both the idea of soft covering [1] and the wiretap channel [5] in the same year, but he didn't connect the two together.

According to the usual definition, strong perfect secrecy is achieved if the mutual information (unnormalized) between the message and the eavesdropper's channel output can be made arbitrarily small.

An even stronger notion of near-perfect secrecy is semantic security. This requires that any two messages cannot be distinguished, usually measured by total variation. This is not implied by the above strong secrecy because mutual information is an average quantity. Since there are so many messages, the mutual information can be small even if a few of the messages are perfectly distinguishable. In [6] a connection between mutual information and semantic security is established by maximizing over message distributions.

Lemma 1 allows us to show that the random codebook construction of Wyner actually achieves semantic security. This is argued by claiming that for every message, due to the randomization at the encoder, the soft-covering phenomenon causes the output distribution at the eavesdropper to be close to an i.i.d. output distribution. The union bound establishes that the probability that even a single message produces an output far from i.i.d. is vanishing quickly with $n$.

In [3] we use this technique to establish semantic security even for a wiretap channel setting with an active adversary.

## Acknowledgment

This work was supported by the National Science Foundation (grant CCF-1350595) and the Air Force Office of Scientific Research (grant FA9550-15-1-0180).

## References

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

[2] P. Cuff. "A Stronger Soft-Covering Lemma and Applications." *Proc. CNS Workshop on Physical-layer Methods for Wireless Security*, Sept. 2015.

[3] Z. Goldfeld, P. Cuff, H. Permuter. "Semantic-Security Capacity for Wiretap Channels of Type II." *CoRR:abs/1509.03619*, 2015.

[4] M. Bloch and N. Laneman, "Strong secrecy from channel resolvability," *IEEE Trans. Inf. Theory*, 59(12): 8077-8098, Dec. 2013.

[5] A. Wyner, "The wire-tap channel," *Bell Systems Technical Journal*, 54(8): 1334-87, Oct. 1975.

[6] M. Bellare, S. Tessaro, and A. Vardy, "Semantic security for the wiretap channel," *in Advances in Cryptology - CRYPTO 2012, LNCS, Springer*, 7417: 294-311, 2012.

# The Gap to Practical MIMO Wiretap Codes

Anatoly Khina
EE Dept., CalTech
Pasadena, CA
Email: khina@caltech.edu

Yuval Kochman
School of CSE, HUJI
Jerusalem, Israel
Email: yuvalko@cs.huji.ac.il

Ashish Khisti
ECE Dept., U. of Toronto
Toronto, ON M5S 3G4, Canada
Email: akhisti@comm.utoronto.ca

The wiretap channel (WTC), introduced by Wyner [1], is composed of a sender ("Alice") who wishes to convey data to a legitimate user ("Bob"), such that the eavesdropper ("Eve") cannot recover any information of this data. In the multiple-input multiple-output (MIMO) Gaussian WTC, Alice is connected to Bob and Eve bi a MIMO broadcast channel. The capacity of this channel was found in [2]–[4].

Although the capacity of WTCs is well understood, construction of practical codes is still a challange. For the scalar Gaussian case, various approaches have been suggested. The recent work of Tyagi and Vardy in [5] is particularly appealing, since it uses a black-box approach: it takes any code that is good for the ordinary (non-secrecy) AWGN channel, and turns it into a good wiretap code using a hashing procedure.

However, assuming that we have such a code for the scalar case, how do we extend it to the vector case? Do we need to construct different codes for every channel matrix? In [6] we have presented a scheme based on scalar random-binning wiretap codes, in conjunction with a linear encoder and a successive interference cancellation (SIC) decoder, which approaches the MIMO wiretap capacity. In fact, it can be described as a variant of VBLAST/GDFE schemes, used in MIMO communication without secrecy [7], [8]. Interestingly, the proof that Eve cannot extract information also hinges on the optimality of the SIC procedure, this time in a "genie-aided" setting: after Eve extracts all possible information from a stream, the content of that stream is revealed to her for the sake of trying to decode the next streams.

Given the optimal SIC scheme for the MIMO WTC, it is natural to consider an explicit code construction, where the random-binning codes are replaced by ordinary AWGN codes, combined with some structured binning procedure, e.g. the hashing of [5]. Indeed, in [9] we have pursued this idea. The key point is that, as with random-binning codes, when any good set of codes is used, a "genie-aided" Eve cannot do better than follow a SIC process. Since at any stage of a SIC decoding process, the decoder sees a multiple-access channel (MAC) where the inputs are the streams that are not decoded yet, the optimality of the scheme is intimately related to that of a scheme for the MAC WTC [10]. However, the construction of good MAC WTC codes is also not immediate.

Even without secrecy, not any collection of good AWGN codes is good for any Gaussian MAC, see e.g. [11]: if the codebooks have structure (as they should, in a practical construction), the signal resulting from one codebook may not look as noise in the process of decoding the other, as for some channel coefficients the codes may align. This compromises MAC decoding, whose optimality is needed both for the "Bob" and "Eve" parts of the secrecy proofs. This effect can be circumvented by a dithering process, which makes sure that codewords play the part of "independent noise" when decoding a different codebook. We thus define a class of MAC WTC codes that have both good individual secrecy properties, and mutual independence; such codebook sets can be obtained from any set of good AWGN codebooks by a two-stage process of hashing and then dithering. However, this still does not yield a practical code construction, as dithering, which must be performed modulo a shaping region to retain optimality, inflicts decoding complexity that may be higher than that of the original code. Thus, in order to obtain a practical construction, we need to find a "simpler" procedure to perturb any given set of codebooks, such that the resulting codes are good for the MAC WTC. It is worth noting, that such a construction will also be theoretically significant in communication without secrecy constraints, as the problem of alignment already arises in VBLAST/GDFR schemes.

## References

[1] A. D. Wyner, "The wiretap channel," *IEEE Trans. Info. Theory*, vol. 54, pp. 1355–1387, 1975.

[2] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas—part II: The MIMOME wiretap channel," *IEEE Trans. Info. Theory*, vol. 56, pp. 5515–5532, 2010.

[3] F. Oggier and B. Hassibi, "The secrecy capacity of the MIMO wiretap channel," *IEEE Trans. Info. Theory*, vol. 57, pp. 4961–4972, 2011.

[4] T. Liu and S. Shamai, "A note on the secrecy capacity of the multiple-antenna wiretap channel," *IEEE Trans. Info. Theory*, vol. 55, pp. 2547–2553, 2009.

[5] H. Tyagi and A. Vardy, "Explicit capacity-achieving coding scheme for the Gaussian wiretap channel," in *ISIT*, Honolulu, HI, USA, June/July 2014.

[6] A. Khina, Y. Kochman, and A. Khisti, "Decomposing the MIMO wiretap channel," in *ISIT*, Honolulu, HI, USA, June/July 2014.

[7] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. URSI Int. Symp. Sig., Sys., Elect. (ISSSE)*, Sep./Oct. 1998, pp. 295–300.

[8] J. M. Cioffi and G. D. Forney Jr., "Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise," in *Comm., Comp., Cont. and Sig. Proc.*, 1997, pp. 79–127.

[9] A. Khina, Y. Kochman, and A. Khisti, "From ordinary AWGN codes to optimal MIMO wiretap schemes," in *ITW*, Hobart, Tasmania, Australia, Oct./Nov. 2014.

[10] E. Tekin and A. Yener, "The Gaussian multiple access wire-tap channel," *IEEE Trans. Info. Theory*, vol. 54, pp. 5747–5755, 2008.

[11] F. Baccelli, A. El Gamal, and D. N. C. Tse, "Interference networks with point-to-point codes," *IEEE Trans. Info. Theory*, pp. 2582–2596, 2011.

# On Secure Broadcasting Over Parallel Channels with Independent Secret Keys

Rafael F. Schaefer
Information Theory and Applications Group
Technische Universität Berlin
email: rafael.schaefer@tu-berlin.de

Ashish Khisti
Dept. of Electrical and Computer Engineering
University of Toronto
email: akhisti@ece.utoronto.ca

H. Vincent Poor
Dept. of Electrical Engineering
Princeton University
email: poor@princeton.edu

***Abstract*—The broadcast channel with independent secret keys models a communication scenario in which a common message has to be securely broadcast to two legitimate receivers keeping an eavesdropper ignorant of it. The transmitter shares an independent secret key of arbitrary rate with each legitimate receiver. Depending on the channel qualities of the legitimate receivers and the eavesdropper, these secret keys must be used as one-time pads to encrypt the message, as fictitious messages for randomization in wiretap coding, or a combination of both to achieve the secrecy capacity. In this paper, communication takes place over independent parallel subchannels and the secrecy capacity is established for these cases, in which all parallel channels follow the same order of degradation.**

## I. Introduction

Security is traditionally implemented at higher layers of the communication protocol such as the application layer and usually based on cryptographic principles. Recently, *information theoretic approaches to security* have drawn considerable attention, especially for wireless communication systems, where it provides a promising complement to cryptographic approaches; see for example [1, 2] and references therein.

Information theoretic approaches realize security directly at the physical layer by exploiting the noisy properties of the underlying communication channel. This line of research was initiated by Wyner who studied the so-called wiretap channel in which a sender wants to securely transmit information over a noisy channel to a receiver keeping an eavesdropper in the dark [3]. In [4–7] this has been extended to the wiretap channel with a shared secret key.

While the basic wiretap channel with secret key is well understood, multi-user communication scenarios involving multiple secret keys have received much less attention. This is insofar surprising as the optimal use of the secret keys is no longer obvious in such scenarios. A transmitter and a receiver can use a shared secret key for encryption to securely transmit a certain message, but this might harm other receivers (of this particular message) that are not aware of this key. Accordingly, multiple secret keys can result in conflicting payoffs at different receivers making the optimal use of the secret keys a challenging and non-trivial problem.

The underlying phenomenon is captured by the broadcast channel (BC) with independent secret keys which has been studied in [8] and [9]. Here, a transmitter wishes to broadcast a common message to two legitimate receivers while keeping an eavesdropper ignorant of it. The transmitter shares independent secret keys of arbitrary rates with both receivers. Now, secure communication can be realized by different approaches. Secret keys can be used as *one-time pads* to encrypt the common message as in [10]. The drawback of this is that each receiver knows only its own secret key. Thus, the more one secret key is used, the more the other receiver is hurt as the encrypted message becomes useless for it. Another approach is to interpret the secret keys as fictitious messages used as randomization resources for *wiretap codes* [1, 2]. The drawback of this approach is that allocating certain resources for randomization reduces the available resources for the actual message transmission. Both approaches are conceptually different and surprisingly neither of them is superior to the other one. In fact, depending on the channel qualities of the legitimate receivers and the eavesdropper, either the one-time pad approach, the fictitious message approach, or a combination of both is needed to achieve capacity [8, 9].

In this paper we study the parallel BC with independent secret keys. In this model, the transmitter communicates with the legitimate receivers (and the eavesdropper) via multiple independent subchannels. We establish the secrecy capacity for the special cases of degraded subchannels, in which all subchannels follow the same order of degradation. The general model of parallel channels is of particular interest as it includes fading channels as a special case. Secure communication over parallel and fading channels for the BC (without secret keys) has been studied in [11] and [12].[1]

## II. BC with Independent Secret Keys

Let $\mathcal{X}$, $\mathcal{Y}_1$, $\mathcal{Y}_2$, and $\mathcal{Z}$ be finite input and output sets. For input and output sequences $x^n \in \mathcal{X}^n$, $y_1^n \in \mathcal{Y}_1^n$, $y_2^n \in \mathcal{Y}_2^n$, and $z^n \in \mathcal{Z}^n$ of length $n$, the discrete memoryless broadcast channel is described by the transition probability $P^n_{Y_1Y_2Z|X}(y_1^n, y_2^n, z^n|x^n) = \prod_{i=1}^n P_{Y_1Y_2Z|X}(y_{1,i}, y_{2,i}, z_i|x_i)$.

The BC with independent secret keys models the communication scenario in which a transmitter broadcasts a common message $M$ to legitimate receivers 1 and 2, while keeping

This research was supported in part by the U. S. National Science Foundation under Grant CMMI-1435778.

[1] *Notation:* The set of non-negative integers $\{1, 2, ..., L\}$ is denoted by $[1, L]$; $X^n_{[1,L]}$ denotes the group of length $n$ vectors $X_1^n, X_2^n, ..., X_L^n$ where $X_l^n = (X_{l,1}, X_{l,2}, ..., X_{l,n})$, $l \in [1, L]$.

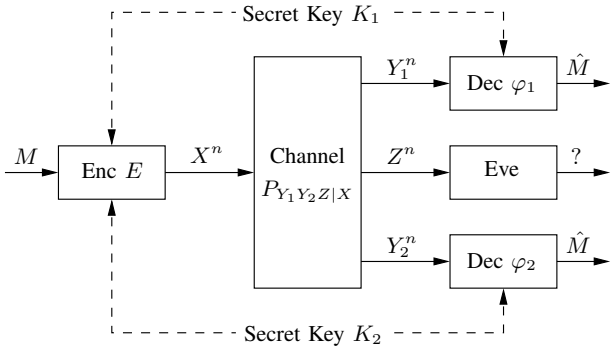



Fig. 1. Broadcast channel with independent secret keys.

an eavesdropper ignorant of it. The transmitter shares secret keys $K_1$ and $K_2$ of arbitrary rates with receivers 1 and 2 as shown in Fig. 1. The message $M$ and the secret keys $K_1$ and $K_2$ are assumed to be independent of each other and uniformly distributed over the sets $\mathcal{M} \coloneqq \{1, ..., M_n\}$ and $\mathcal{K}_i \coloneqq \{1, ..., K_{i,n}\}$, $i = 1, 2$. This has been studied in [8] and [9] and the secrecy capacity has been established for several cases.

Depending on the channel qualities of the legitimate receivers and the eavesdropper, the optimal use of the secret keys varies: Either they must be used as one-time pads to encrypt the message, must be interpreted as fictitious messages used as randomization resources for wiretap coding, or a combination of both. If the eavesdropper has the "strongest" channel in the sense that the Markov chain relationship $X - Z - Y_1 - Y_2$ is satisfied, the secret keys are used as one-time pads to create two encrypted messages based on a bit-wise XOR operation. Capacity is then achieved by superposition coding of these two messages.

*Theorem 1 ([8]). The secrecy capacity $C$ of the BC with independent secret keys and reversely degraded channels $X - Z - Y_1 - Y_2$ is*

$$C = \max_{P_{UX}} \min \left\{ I(X;Y_1|U), I(U;Y_2) \right\} \tag{1}$$

*where the* max *is over all input distributions $P_{UX}(u, x)$ such that $U - X - Z - Y_1 - Y_2$ form a Markov chain. Further, the cardinality of the range of $U$ can be bounded by $|\mathcal{U}| \leq |\mathcal{X}| + 1$.*

If the eavesdropper has the "weakest" channel in the sense that $X - Y_1 - Z$ and $X - Y_2 - Z$ are satisfied, then the secret keys are used as fictitious messages playing the role of randomization resources for wiretap codes.

*Theorem 2 ([8]). The secrecy capacity $C$ of the BC with independent secret keys and degraded channels $X - Y_1 - Z$ and $X - Y_2 - Z$ is*

$$C = \max_{P_X} \min \left\{ \begin{array}{l} I(X;Y_1) \\ I(X;Y_2) \\ \frac{1}{2}[I(X;Y_1) + I(X;Y_2) - I(X;Z)] \end{array} \right\}. \tag{2}$$

If the eavesdropper has neither the strongest nor the weakest channel, a combination of both approaches is needed.

*Theorem 3 ([9]). The secrecy capacity $C$ of the BC with independent secret keys for $X - Y_1 - Z - Y_2$ is*

$$C = \max_{P_{UX}} \min \left\{ \begin{array}{l} I(U;Y_2) \\ \frac{1}{2}\big[I(X;Y_1) + I(U;Y_2) - I(U;Z)\big] \end{array} \right\} \tag{3}$$

*where the* max *is over all input distributions $P_{UX}(u, x)$ such that $U - X - Y_1 - Z - Y_2$ form a Markov chain. The cardinality of the range of $U$ can be bounded by $|\mathcal{U}| \leq |\mathcal{X}| + 1$.*

## III. Parallel Channels

In the following we consider a parallel BC with $L$ independent subchannels and each of them is a BC as introduced above. Now the transmitter wants to transmit the common message over these $L$ subchannels. Accordingly, the parallel BC consists of $L$ finite input alphabets $\mathcal{X}_{[1,L]}$ and $3L$ finite output alphabets $\mathcal{Y}_{1,[1,L]}$, $\mathcal{Y}_{2,[1,L]}$, and $\mathcal{Z}_{[1,L]}$. The transition probability is then

$$P^n_{Y_{1,[1,L]}Y_{2,[1,L]}Z_{[1,L]}|X_{[1,L]}}(y^n_{1,[1,L]}, y^n_{2,[1,L]}, z^n_{[1,L]}|x^n_{[1,L]}) \tag{4}$$

$$= \prod_{l=1}^{L} P^n_{Y_{1,l}Y_{2,l}Z_l|X_l}(y^n_{1,l}, y^n_{2,l}, z^n_l|x^n_l) \tag{5}$$

$$= \prod_{l=1}^{L} \prod_{i=1}^{n} P_{Y_{1,l}Y_{2,l}Z_l|X_l}(y_{1,l,i}, y_{2,l,i}, z_{l,i}|x_{l,i}) \tag{6}$$

where $x^n_l \in \mathcal{X}^n_l$, $y^n_{1,l} \in \mathcal{Y}^n_{1,l}$, $y^n_{2,l} \in \mathcal{Y}^n_{2,l}$, and $z^n_l \in \mathcal{Z}^n_l$ are the input and output sequences of length $n$ on subchannel $l \in [1, L]$.

*Definition 1.* An $(n, M_n, K_{1,n}, K_{2,n})$-*code* for the parallel BC with independent secret keys consists of a (stochastic) encoder

$$E : \mathcal{M} \times \mathcal{K}_1 \times \mathcal{K}_2 \to \mathcal{P}(\mathcal{X}^n_1 \times ... \times \mathcal{X}^n_L) \tag{7}$$

with $\mathcal{P}(\cdot)$ the set of all probability distributions, and decoders

$$\varphi_1 : \mathcal{Y}^n_{1,1} \times ... \times \mathcal{Y}^n_{1,L} \times \mathcal{K}_1 \to \mathcal{M} \tag{8}$$

$$\varphi_2 : \mathcal{Y}^n_{2,1} \times ... \times \mathcal{Y}^n_{2,L} \times \mathcal{K}_2 \to \mathcal{M}. \tag{9}$$

We denote the average probability of decoding error at receiver $i$ by $\bar{e}_{i,n}$, $i = 1, 2$. The secrecy criterion is

$$I(M; Z^n_{[1,L]}) = I(M; Z^n_1, Z^n_2, ..., Z^n_L) \leq \delta_n \tag{10}$$

for $\delta_n > 0$ with $M$ the random variable uniformly distributed over the set of messages $\mathcal{M}$ and $Z^n_l = (Z_{l,1}, Z_{l,2}, ..., Z_{l,n})$ the output at the eavesdropper on subchannel $l$. This condition is termed *strong secrecy*.

*Definition 2.* A rate $R > 0$ is an *achievable secrecy rate* for the parallel BC with independent secret keys if for any $\tau > 0$ there exist an $n(\tau) \in \mathbb{N}$ and a sequence of $(n, M_n, K_{1,n}, K_{2,n})$-codes such that for all $n \geq n(\tau)$ we have $\frac{1}{n} \log M_n \geq R - \tau$ and $I(M; Z^n_{[1,L]}) \leq \delta_n$ while $\bar{e}_{1,n}, \bar{e}_{2,n}, \delta_n \to 0$ as $n \to \infty$. The *secrecy capacity* $C$ is the supremum of all achievable rates $R$.

### A. Strongest Eavesdropper

We start with the scenario in which the eavesdropper is the "strongest" receiver for all parallel subchannels, i.e., we have the Markov chain relationships $X_l - Z_l - Y_{1,l} - Y_{2,l}$ for all $l \in [1, L]$.

*Theorem 4. The secrecy capacity $C$ of the parallel BC with independent secret keys for reversely degraded channels is*

$$C = \max \min \left\{ \begin{array}{l} \sum_{l=1}^{L} I(X_l; Y_{1,l}|U_l) \\ \sum_{l=1}^{L} I(U_l; Y_{2,l}) \end{array} \right\} \tag{11}$$

*where the* max *is over all input distributions $\prod_{l=1}^{L} P_{U_l X_l}(u, x)$ such that $U_l - X_l - Z_l - Y_{1,l} - Y_{2,l}$, $l \in [1, L]$, form Markov chains, i.e., superposition coding on each subchannel is optimal. Further, the cardinality of the range of $U_l$ can be bounded by $|\mathcal{U}_l| \leq |\mathcal{X}_l| + 1$, $l \in [1, L]$.*

*Proof:* The achievability is based on [8, Theorem 2] by using superposition coding on each subchannel. More specifically, the secret keys $K_1$ and $K_2$ are used as one-time pads to encrypt the common message $M$ into two individual messages $M_1 = M \oplus K_1$ and $M_2 = M \oplus K_2$. Both messages are split up into $L$ submessages $M_1 = (M_{1,1}, M_{1,2}, ..., M_{1,L})$ and $M_2 = (M_{2,1}, M_{2,2}, ..., M_{2,L})$.

Now, on each subchannel $l \in [1, L]$ the message pair $(M_{1,l}, M_{2,l})$ is individually encoded and transmitted using superposition coding. Thereby, subchannel $l \in [1, L]$ is a degraded BC supporting the rates $R_{1,l} = I(X_l; Y_{1,l}|U_l)$ and $R_{2,l} = I(U_l; Y_{2,l})$. In total, we achieve with this strategy rates $R_1 = \sum_{i=1}^{L} R_{1,l}$ and $R_2 = \sum_{i=1}^{L} R_{2,l}$ yielding the desired achievable rate as in (11).

It remains to show the converse. At receiver $i$, $i = 1, 2$, we have the following version of Fano's inequality:

$$H(M|Y_{i,[1,L]}^n, K_i) \leq n\epsilon_{i,n} \tag{12}$$

with $\epsilon_{i,n} \to 0$ as $n \to \infty$. We define the auxiliary random variables

$$U_{l,i} = (M, K_2, Y_{1,[1,l-1]}^n, Y_{1,l}^{i-1}) \tag{13}$$

for all $l \in [1, L]$. For the weaker receiver we obtain

$$\begin{aligned}
nR &= H(M) = H(M|K_2) && (14)\\
&= I(M; Y_{2,[1,L]}^n|K_2) + n\epsilon_{i,n} && (15)\\
&\leq I(M, K_2; Y_{2,[1,L]}^n) + n\epsilon_{i,n} && (16)\\
&= \sum_{l=1}^{L} I(M, K_2; Y_{2,l}^n|Y_{2,[1,l-1]}^n) + n\epsilon_{i,n} && (17)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{n} I(M, K_2; Y_{2,l,i}|Y_{2,[1,l-1]}^n, Y_{2,l}^{i-1}) + n\epsilon_{i,n} && (18)\\
&\leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(M, K_2, Y_{2,[1,l-1]}^n, Y_{2,l}^{i-1}; Y_{2,l,i}) + n\epsilon_{i,n} && (19)\\
&\leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(M, K_2, Y_{1,[1,l-1]}^n, Y_{1,l}^{i-1}; Y_{2,l,i}) + n\epsilon_{i,n} && (20)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{n} I(U_{l,i}; Y_{2,l,i}) + n\epsilon_{i,n} && (21)
\end{aligned}$$

where (15) follows from Fano's inequality (12), (20) from the degradedness $X_l - Y_{1,l} - Y_{2,l}$ for all $l \in [1, L]$, and (21) from the definition of $U_{l,i}$, cf. (13).

With the same definition for $U_{l,i}$ we obtain for the stronger receiver

$$\begin{aligned}
nR &\leq I(M, K_1; Y_{1,[1,L]}^n) + n\epsilon_{i,n} && (22)\\
&= I(K_1; Y_{1,[1,L]}^n|M) + I(M; Y_{1,[1,L]}^n) \\
&\qquad - I(M; Z_{[1,L]}^n) + n\epsilon_{i,n} + n\delta_n && (23)\\
&\leq I(K_1; Y_{1,[1,L]}^n|M) + n\epsilon'_{i,n} && (24)\\
&\leq I(K_1; Y_{1,[1,L]}^n|M, K_2) + n\epsilon'_{i,n} && (25)\\
&= \sum_{l=1}^{L} I(K_1; Y_{1,l}^n|M, K_2, Y_{1,[1,l-1]}^n) + n\epsilon'_{i,n} && (26)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{n} I(K_1; Y_{1,l,i}|M, K_2, Y_{1,[1,l-1]}^n, Y_{1,l}^{i-1}) + n\epsilon'_{i,n} && (27)\\
&\leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(K_1, X_{l,i}; Y_{1,l,i}|M, K_2, Y_{1,[1,l-1]}^n, Y_{1,l}^{i-1}) + n\epsilon'_{i,n} && (28)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{n} I(X_{l,i}; Y_{1,l,i}|M, K_2, Y_{1,[1,l-1]}^n, Y_{1,l}^{i-1}) + n\epsilon'_{i,n} && (29)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{n} I(X_{l,i}; Y_{1,l,i}|U_{l,i}) + n\epsilon'_{i,n} && (30)
\end{aligned}$$

with $\epsilon'_{i,n} = \epsilon_{i,n} + \delta_n$. Here, (23) follows from the secrecy criterion (10) and (24) follows from the fact that $I(M; Y_{1,[1,L]}^n) \leq I(M; Z_{[1,L]}^n)$ due to the degradedness.

Now, let $Q$ be a time-sharing random variable independent of all others and uniformly distributed over $\{1, ..., n\}$. We set $U_l = (U_{l,Q}; Q)$, $X_l = X_{l,Q}$, $Y_{1,l} = Y_{1,l,Q}$, and $Y_{2,l} = Y_{2,l,Q}$ for all $l = 1, ..., L$ and obtain for the rate of the weaker receiver in (21)

$$\begin{aligned}
R &\leq \sum_{i=1}^{n} I(U_{l,Q}; Y_{2,l,Q}|Q) + \epsilon_{i,n} && (31)\\
&\leq \sum_{i=1}^{n} I(U_l; Y_{2,l}) + \epsilon_{i,n} && (32)
\end{aligned}$$

and for the rate of the stronger receiver in (30)

$$\begin{aligned}
R &\leq \sum_{l=1}^{L} I(X_{l,Q}; Y_{1,l,Q}|U_{l,Q}, Q) + \epsilon'_{i,n} && (33)\\
&= \sum_{l=1}^{L} I(X_l; Y_{1,l}|U_l) + \epsilon'_{i,n}. && (34)
\end{aligned}$$

This yields the desired bounds in (11) proving the desired converse. The cardinality bounds follow then immediately from [8, Theorem 2]. ■

### B. Weakest Eavesdropper

Now we look at the scenario in which the eavesdropper is the "weakest" receiver for all parallel subchannels, i.e., we have the Markov chain relationships $X_l - Y_{1,l} - Z_l$ and $X_l - Y_{2,l} - Z_l$ for all $l \in [1, L]$.

*Theorem 5. The secrecy capacity $C$ of the parallel BC with independent secret keys and degraded channels $X_l - Y_{1,l} - Z_l$ and $X_l - Y_{2,l} - Z_l$ is*

$$C = \max \min \left\{ \begin{array}{l} \sum_{l=1}^{L} I(X_l; Y_{1,l}) \\ \sum_{l=1}^{L} I(X_l; Y_{2,l}) \\ \sum_{l=1}^{L} \frac{1}{2}[I(X_l; Y_{1,l}) + I(X_l; Y_{2,l}) - I(X_l; Z_l)] \end{array} \right\} \tag{35}$$

*where the* max *is over all input distributions* $\prod_{l=1}^{L} P_{X_l}(x)$.

*Proof:* The achievability follows from [8, Theorem 3] by extending this coding scheme to the vector-valued case here, specifically, by setting $X = (X_1, X_2, ..., X_L)$, $Y_1 = (Y_{1,1}, Y_{1,2}, ..., Y_{1,L})$, $Y_2 = (Y_{2,1}, Y_{2,2}, ..., Y_{2,L})$, and $Z = (Z_1, Z_2, ..., Z_L)$ with $X$ having independent components.

For the converse we observe that the first two bounds in (35) are single-user bounds. The more interesting bound is the third, sum-rate-like bound which is as follows:

$$\begin{aligned} n2R &= H(M) + H(M) && (36)\\ &= H(M|K_1) + H(M|K_2) && (37)\\ &\le I(M; Y_{1,[1,L]}^n|K_1) + I(M; Y_{2,[1,L]}^n|K_2) + n\epsilon_n && (38)\\ &\le I(M; Y_{1,[1,L]}^n|K_1) + I(M; Y_{2,[1,L]}^n|K_2) \\ &\quad - I(M; Z_{[1,L]}^n) + n\epsilon_n' && (39)\\ &\le I(M, K_1; Y_{1,[1,L]}^n) + I(M, K_2; Y_{2,[1,L]}^n) \\ &\quad - I(M; Z_{[1,L]}^n) + n\epsilon_n' && (40)\\ &\le I(M, K_{12}; Y_{1,[1,L]}^n) + I(M, K_{12}; Y_{2,[1,L]}^n) \\ &\quad - I(M, K_{12}; Z_{[1,L]}^n) + n\epsilon_n' && (41) \end{aligned}$$

with $\epsilon_n' = \epsilon_n + \delta_n = \epsilon_{i,n} + \epsilon_{2,n} + \delta_n$. Here, (38) follows from Fano's inequality, (39) from the secrecy criterion (10), and (41) follows from the same steps used in [8, Eqs. (26)-(27)]. Due to the degradedness we can write the first and third term in (41) as

$$\begin{aligned} &I(M, K_{12}; Y_{1,[1,L]}^n) - I(M, K_{12}; Z_{[1,L]}^n) \\ &\quad = I(M, K_{12}; Y_{1,[1,L]}^n|Z_{[1,L]}^n) && (42)\\ &\quad = H(Y_{1,[1,L]}^n|Z_{[1,L]}^n) - H(Y_{1,[1,L]}^n|M, K_{12}, Z_{[1,L]}^n) && (43)\\ &\quad \le H(Y_{1,[1,L]}^n|Z_{[1,L]}^n) - H(Y_{1,[1,L]}^n|X_{[1,L]}^n, Z_{[1,L]}^n) && (44)\\ &\quad = H(Y_{1,[1,L]}^n|Z_{[1,L]}^n) - \sum_{l=1}^{L} H(Y_{1,l}^n|X_l^n, Z_l^n) && (45)\\ &\quad \le \sum_{l=1}^{L} H(Y_{1,l}^n|Z_l^n) - \sum_{l=1}^{L} H(Y_{1,l}^n|X_l^n, Z_l^n) && (46)\\ &\quad = \sum_{l=1}^{L} I(X_l^n; Y_{1,l}^n|Z_l^n). && (47) \end{aligned}$$

Similarly we get for the second term

$$I(M, K_{12}; Y_{2,[1,L]}^n) \le \sum_{l=1}^{L} I(X_l^n; Y_{2,l}^n) \tag{48}$$

so that we can upper bound (41) by

$$\begin{aligned} n2R &\le \sum_{l=1}^{L} \left[ I(X_l^n; Y_{1,l}^n|Z_l^n) + I(X_l^n; Y_{2,l}^n) \right] + n\epsilon_n' && (49)\\ &\le n \sum_{l=1}^{L} \left[ I(X_l; Y_{1,l}|Z_l) + I(X_l; Y_{2,l}) \right] + n\epsilon_n' && (50)\\ &= n \sum_{l=1}^{L} \left[ I(X_l; Y_{1,l}) + I(X_l; Y_{2,l}) - I(X_l; Z_l) \right] + n\epsilon_n' && (51) \end{aligned}$$

which gives the desired sum-rate-like bound in (35). This completes the converse. ■

### C. Eavesdropper in the Middle

Finally, we study the scenario in which the eavesdropper is neither the strongest nor the weakest receiver for all parallel subchannels. In particular, we assume that the Markov chain relationships $X_l - Y_{1,l} - Z_l - Y_{2,l}$ hold for all $l \in [1, L]$.

*Theorem 6. The secrecy capacity $C$ of the parallel BC with independent secret keys for $X_l - Y_{1,l} - Z_l - Y_{2,l}$ is*

$$C = \max \min \left\{ \begin{array}{l} \sum_{l=1}^{L} I(U_l; Y_{2,l}) \\ \sum_{l=1}^{L} \frac{1}{2}[I(X_l; Y_{1,l}) + I(U_l; Y_{2,l}) - I(U_l; Z_l)] \end{array} \right\} \tag{52}$$

*where the* max *is over all input distributions* $\prod_{l=1}^{L} P_{U_l X_l}(u, x)$ *such that* $U_l - X_l - Y_{1,l} - Z_l - Y_{2,l}$, $l \in [1, L]$, *form Markov chains. Further, the cardinality of the range of* $U_l$ *can be bounded by* $|\mathcal{U}_l| \le |\mathcal{X}_l| + 1$, $l \in [1, L]$.

*Proof:* The achievability follows from [9, Theorem 4] by extending this coding scheme to the vector-valued case here, specifically, by setting $U = (U_1, U_2, ..., U_L)$, $X = (X_1, X_2, ..., X_L)$, $Y_1 = (Y_{1,1}, Y_{1,2}, ..., Y_{1,L})$, $Y_2 = (Y_{2,1}, Y_{2,2}, ..., Y_{2,L})$, and $Z = (Z_1, Z_2, ..., Z_L)$ with $U$ and $X$ having independent components.

For the converse we define the auxiliary random variables

$$U_{l,i} = (M, K_2, Z_{[1,l-1]}^n, Z_l^{i-1}) \tag{53}$$

for all $l \in [1, L]$. For the first bound in (52) we obtain

$$\begin{aligned} nR &= H(M) = H(M|K_2) && (54)\\ &= I(M; Y_{2,[1,L]}^n|K_2) + n\epsilon_{i,n} && (55)\\ &\le I(M, K_2; Y_{2,[1,L]}^n) + n\epsilon_{i,n} && (56)\\ &= \sum_{l=1}^{L} I(M, K_2; Y_{2,l}^n|Y_{2,[1,l-1]}^n) + n\epsilon_{i,n} && (57)\\ &= \sum_{l=1}^{L} \sum_{i=1}^{n} I(M, K_2; Y_{2,l,i}|Y_{2,[1,l-1]}^n, Y_{2,l}^{i-1}) + n\epsilon_{i,n} && (58) \end{aligned}$$

$$= \sum_{l=1}^{L}\sum_{i=1}^{n} I(M, K_2, Y_{2,[1,l-1]}^n, Y_{2,l}^{i-1}; Y_{2,l,i}) + n\epsilon_{i,n} \quad (59)$$

$$\leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(M, K_2, Z_{[1,l-1]}^n, Z_l^{i-1}; Y_{2,l,i}) + n\epsilon_{i,n} \quad (60)$$

$$= \sum_{l=1}^{L}\sum_{i=1}^{n} I(U_{l,i}; Y_{2,l,i}) + n\epsilon_{i,n} \quad (61)$$

where (55) follows from Fano's inequality, (60) from the degradedness $X_l - Y_{1,l} - Z_l - Y_{2,l}$ for all $l \in [1, L]$, and (61) from the definition of $U_{l,i}$.

For the second bound we proceed as follows. We know from the proof of Theorem 5, in particular (40), that

$$\begin{aligned} n2R &\leq I(M, K_1; Y_{1,[1,L]}^n) + I(M, K_2; Y_{2,[1,L]}^n) \\ &\quad - I(M; Z_{[1,L]}^n) + n\epsilon_n' \quad (62) \\ &\leq I(M, K_{12}; Y_{1,[1,L]}^n) + I(M, K_2; Y_{2,[1,L]}^n) \\ &\quad - I(M, K_2; Z_{[1,L]}^n) + n\epsilon_n' \quad (63) \end{aligned}$$

with $\epsilon_n' = \epsilon_{1,n} + \epsilon_{2,n} + \delta_n$, where the last step follows similarly as in [9, Eqs. (25)-(27)]. Using the chain rule for mutual information we obtain

$$\begin{aligned} n2R &\leq I(M, K_{12}; Y_{1,[1,L]}^n) + I(M, K_2; Y_{2,[1,L]}^n) \\ &\quad - I(M, K_{12}; Z_{[1,L]}^n) + I(K_1; Z_{[1,L]}^n | M, K_2) + n\epsilon_n'. \end{aligned} \quad (64)$$

From the proof of Theorem 5 we know that we can bound the first and third term as

$$\begin{aligned} &I(M, K_{12}; Y_{1,[1,L]}^n) - I(M, K_{12}; Z_{[1,L]}^n) \\ &\quad \leq n \sum_{l=1}^{L} \big[ I(X_l; Y_{1,l}) - I(X_l; Z_l) \big], \end{aligned} \quad (65)$$

cf. (42)-(51). The second term is upper bounded by

$$I(M, K_2; Y_{2,[1,L]}^n) \leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(U_{l,i}; Y_{2,l,i}) \quad (66)$$

which follows as in (56)-(61). It remains to bound the last term as follows:

$$\begin{aligned} &I(K_1; Z_{[1,L]}^n | M, K_2) \\ &\quad \leq \sum_{l=1}^{L} I(K_1; Z_l^n | M, K_2, Z_{[1,l-1]}^n) \quad (67) \\ &\quad \leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(K_1; Z_{l,i} | M, K_2, Z_{[1,l-1]}^n, Z_l^{l-1}) \quad (68) \\ &\quad \leq \sum_{l=1}^{L}\sum_{i=1}^{n} I(K_1, X_{l,i}; Z_{l,i} | M, K_2, Z_{[1,l-1]}^n, Z_l^{l-1}) \quad (69) \\ &\quad = \sum_{l=1}^{L}\sum_{i=1}^{n} I(X_{l,i}; Z_{l,i} | M, K_2, Z_{[1,l-1]}^n, Z_l^{l-1}) \quad (70) \\ &\quad = \sum_{l=1}^{L}\sum_{i=1}^{n} I(X_{l,i}; Z_{l,i} | U_{l,i}). \quad (71) \end{aligned}$$

Now by inserting (65), (66), (71) into (64) and introducing a time-sharing random variable independent of all others and uniformly distributed over $\{1, ..., n\}$, it is straightforward to obtain the desired bounds in (52), thereby proving the converse. The cardinality bounds follow then immediately from [9, Theorem 4] ■

## IV. Conclusion

In this paper we have studied the parallel BC with independent secret keys. Here, the transmitter sends a common message to two legitimate receivers via multiple independent subchannels. Parallel channels are of particular interest as they include fading channels as a special case. We have established the secrecy capacity of the parallel BC with independent secret keys for the cases in which the channels of the legitimate receivers and the eavesdropper all follow the same order of degradation.

If the eavesdropper is weakest receiver for all subchannels, the optimal coding scheme is the vector-valued extension of the single-channel case. The same is true for the case in which the eavesdropper is neither the weakest nor the strongest for all subchannels. However, if the eavesdropper is the strongest receiver for all subchannels, the optimal coding scheme is based on individual superposition coding on each subchannel. All considered cases have in common that independent inputs for all subchannels are optimal which does not immediately follow from [8] and [9]. Moreover, for all considered cases the secrecy capacity of the parallel BC is in general larger than the sum of the secrecy capacities of all individual subchannels.

## References


[1] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Information Theoretic Security," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 4-5, pp. 355–580, 2009.

[2] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[3] A. D. Wyner, "The Wire-Tap Channel," *Bell Syst. Tech. J.*, vol. 54, pp. 1355–1387, Oct. 1975.

[4] H. Yamamoto, "Rate-Distortion Theory for the Shannon Cipher System," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 827–835, May 1997.

[5] N. Merhav, "Shannon's Secrecy System with Informed Receivers and Its Application to Systematic Coding for Wiretapped Channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2723–2734, Jun. 2008.

[6] W. Kang and N. Liu, "Wiretap Channel with Shared Key," in *Proc. IEEE Inf. Theory Workshop*, Dublin, Ireland, Aug. 2010, pp. 1–5.

[7] R. F. Schaefer and H. V. Poor, "Robust Transmission over Wiretap Channels with Secret Keys," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, USA, Nov. 2014, pp. 60–64.

[8] R. F. Schaefer and A. Khisti, "Secure Broadcasting of a Common Message with Independent Secret Keys," in *Proc. Conf. Inf. Sciences and Systems*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[9] R. F. Schaefer, A. Khisti, and H. V. Poor, "How to Use Independent Secret Keys for Secure Broadcasting of Common Messages," in *Proc. IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015, pp. 1971–1975.

[10] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell Syst. Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.

[11] A. Khisti, A. Tchamkerten, and G. W. Wornell, "Secure Broadcasting Over Fading Channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2453–2469, Jun. 2008.

[12] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Secure Communication Over Fading Channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2470–2492, Jun. 2008.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Privacy Preserving Rechargeable Battery Policies for Smart Metering Systems

*(Invited Paper)*

Simon Li
Department of Electrical and Computer Engineering
University of Toronto
Email: simonli@ece.utoronto.ca

Ashish Khisti
Department of Electrical and Computer Engineering
University of Toronto
Email: akhisti@ece.utoronto.ca

Aditya Mahajan
Department of Electrical and Computer Engineering
McGill University
Email: aditya.mahajan@mcgill.ca

***Abstract*—We consider a setup where a rechargeable battery is used to partially mask the load profile of a user from the utility provider in a smart-metered electrical system. We focus on the case of i.i.d. load profile, use mutual information as our privacy metric, and characterize the optimal policy as well as the associated leakage rate.**

**Our approach is based on obtaining single-letter expression for the leakage rate for a class of battery policies and providing a converse argument for establishing the optimality.**

## I. Introduction

Smart meters are becoming a critical part of modern electrical grids. They deliver fine-grained household power usage measurements to utility providers. This information allows them to implement changes to improve the efficiency of the electrical grid. However, despite the promise of savings in energy and money, there is potentially a loss of privacy. Anyone with access to the load profile may employ data mining algorithms to infer details about the private activities of the user [1]–[4].

In this paper, we investigate one possible solution to the privacy problem. Using a rechargeable battery, the user can distort the load profile generated by the appliances by charging and discharging the battery. Due to the proliferation of rechargeable batteries, energy harvesting devices and electric vehicles, the strategy of using these devices to partially obfuscate the user's load profile is becoming more feasible. As we discuss below, a number of recent works have studied this approach in the literature.

### A. Related Works

We consider a similar setup to [7] which introduces using mutual information as a privacy metric then considers an instance of the problem with binary alphabets. The setup is extended in [8], [9] where the multi-letter mutual information optimization problem is reformulated as a Markov Decision Process. The results in this paper mirror that of [10] where the optimal single-letter information leakage rate and policy is characterized using Markov Decision Theory. In this paper, we provide the proofs using purely information theoretic arguments which may be of interest in its own right. In other related works, rate-distortion type approaches for studying privacy-utility tradeoffs in smart grid systems have been studied in [11]–[14]. These works are not directly related to the present setup.

## II. Problem definition

We consider a smart metering system as shown in Fig. 1 where at each time a residence generates an aggregate demand that must either be satisfied by charges in the battery or by drawing power from the grid. $\{X_t\}_{t\geq 1}$, $X_t \in \mathcal{X}$ where $\mathcal{X} := \{0, 1, 2, \dots, m_x\}$ denotes the (exogeneous) i.i.d. power demand process distributed according to $Q_X$. $\{Y_t\}_{t\geq 1}$, $Y_t \in \mathcal{Y}$, denotes the energy consumed from the grid where $\mathcal{Y} := \{0, 1, 2, \dots, m_y\}$ and $\{S_t\}_{t\geq 1}$, $S_t \in \mathcal{S}$ denotes the energy stored in the battery where $\mathcal{S} := \{0, 1, 2, \dots, m_s\}$ and the initial charge $S_1$ of the battery is distributed according to probability mass function $P_{S_1}$.

We assume that $m_x \leq m_y$ so that the system is guaranteed to be able to satisfy the demand at any time by drawing solely from the grid i.e. $Y_t = X_t,\ \forall t$. While in general, the alphabets $\mathcal{X}$ and $\mathcal{Y}$ can be any finite subset of the integers – where negative values of $X$ and $Y$ would model a situation where energy (possibly generated from an alternative energy source) is sold back to the utility provider – it is more realistic to for them to be a contiguous interval. In this case, without further assumptions on the battery size, the alphabets would have to satisfy $\mathcal{X} \subset \mathcal{Y}$ in order to guarantee that energy is not wasted and the power demand can always be satisfied. Nonetheless, our results generalize to these cases.

We assume an ideal battery that has no conversion losses or other inefficiencies. Therefore, the following conservation equation must be satisfied at all time instances:

$$S_{t+1} = S_t - X_t + Y_t. \tag{1}$$

The energy management system observes the power demand and battery charge and consumes energy from the grid according to a randomized *charging policy* $\mathbf{q} = (q_1, q_2, \dots)$. In particular, at time $t$, given $(x^t,\ s^t, y^{t-1})$, the history of demand, battery charge, and past consumption, the battery policy chooses the level of current consumption $Y_t$ to be $y$ with probability $q_t(y \mid x^t, s^t, y^{t-1})$. For a randomized charging policy to be feasible, it must satisfy the conservation

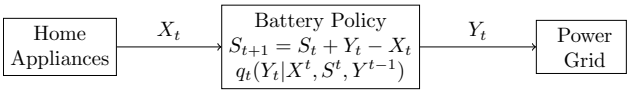



Fig. 1: System Diagram. The user demand is denoted by $X_t$, the grid consumption by $Y_t$, and the battery state by $S_t$. The battery policy is denoted by the conditional distribution $q(Y_t|X^t, S^t, Y^{t-1})$. The battery policy effectively defines a channel with memory from the residence to the utility provider.

equation (1), so given the current power demand and battery charge $(x_t, s_t)$, the feasible values of grid consumption are defined by

$$\mathcal{Y}_\circ(s_t - x_t) = \{y \in \mathcal{Y} : s_t - x_t + y \in \mathcal{S}\}.$$

Thus, we require that

$$\begin{aligned} & q_t(\mathcal{Y}_\circ(s_t - x_t) \mid x^t, s^t, y^{t-1}) \\ & := \sum_{y \in \mathcal{Y}_\circ(s_t - x_t)} q_t(y \mid x^t, s^t, y^{t-1}) \\ & = 1. \end{aligned}$$

The set of all such feasible strategies is denoted by $\mathcal{Q}_A$. A battery policy effectively defines a channel with memory between a residence and the utility provider (as portrayed in Fig. 1).

The quality of a charging policy depends on the amount of information leaked under that policy. This notion is captured by mutual information $I^{\mathbf{q}}(S_1, X^T; Y^T)$ evaluated according to the joint probability distribution on $(S^T, X^T, Y^T)$ induced by the sequence $\mathbf{q}$:

$$\begin{aligned} & \mathbb{P}^{\mathbf{q}}(S^T = s^T, X^T = x^T, Y^T = y^T) \\ & \quad = P_{S_1}(s_1) P_{X_1}(x_1) q_1(y_1 \mid x_1, s_1) \\ & \qquad \times \prod_{t=2}^{T} \Big[ \mathbb{1}_{s_t}\{s_{t-1} - x_{t-1} + y_{t-1}\} \\ & \qquad\quad Q(x_t) q_t(y_t \mid x^t, s^t, y^{t-1}) \Big]. \end{aligned} \tag{2}$$

Given a policy $\mathbf{q} = (q_1, q_2, \dots) \in \mathcal{Q}_A$, we define the worst case information leakage *rate* as follows:

$$L_\infty(\mathbf{q}) := \limsup_{T \to \infty} \frac{1}{T} I^{\mathbf{q}}(S_1, X^T; Y^T). \tag{3}$$

*Remark* II.1. The random variable $S_1$ in the mutual information terms do not affect the asymptotic rate. It will be clear in the sequel that this simplifies the analysis.

We are interested in the following optimization problem:

**Problem A.** *Given the alphabet $\mathcal{X}$ and distribution $Q_X$ of the power demand, the alphabet $\mathcal{S}$ of the battery, the initial distribution $P_{S_1}$ of the battery state, and the alphabet $\mathcal{Y}$ of the demand: find a battery charging policy $\mathbf{q} = (q_1, q_2, \dots) \in \mathcal{Q}_A$ that minimizes the leakage rate $L_\infty(\mathbf{q})$ given by* (3)*.*

## III. Stationary Posterior Policies

The simplest class of policies are stationary and memoryless, conditioning only on the current battery state and power demand:

$$q(y|x, s). \tag{4}$$

As such evaluating the leakage rate (3) even for this simplified class of policies requires numerical approaches, see e.g., [7], [13]. Our key insight is that if we further impose a certain invariance condition we can obtain a closed form expression for the leakage rate. Interestingly we will see that this class of policies also includes a globally optimal policy. Our proposed class preserves the following property:

$$\mathbb{P}(S_2 = s_2|Y_1 = y_1) = \mathbb{P}(S_1 = s_2), \ \forall s_2 \in \mathcal{S}, y_1 \in \hat{\mathcal{Y}} \tag{5}$$

where $\hat{\mathcal{Y}} := \{y : P_{Y_1}(y_1) > 0\}$ for some initial battery state distribution $\mathbb{P}_{S_1}$. This invariance condition implies that $S_t \perp Y_{t-1}$ and also that $\mathbb{P}_{S_t} = \mathbb{P}_{S_1}$, $\forall t$. By exploiting this property, we can obtain single-letter achievable leakage rates as follows:

**Lemma III.1.** *Given an instance of Problem A with i.i.d. power demand $Q_X(x)$ and initial battery state distribution $\mathbb{P}_{S_1}$, if the stationary memoryless policy $\mathbf{q} = (q, q, \dots) \in \mathcal{Q}_A$ satisfies the invariance property* (5)*, then*

$$L_\infty(\mathbf{q}) = I^{\mathbf{q}}(S_1, X_1; Y_1),$$

*where $(S_1, X_1, Y_1) \sim \mathbb{P}_{S_1}(s_1) Q(x_1) q(y_1|x_1, s_1)$.*

*Proof.* The invariance property and the memorylessness of $\mathbf{q}$ implies that $(Y_t, X_t, S_t) \perp Y^{t-1}$, $\forall t$. Therefore we have

$$\begin{aligned} \frac{1}{T} I^{\mathbf{q}}(S_1, X^T; Y^T) & \overset{(a)}{=} \sum_{t=1}^{T} \frac{1}{T} I^{\mathbf{q}}(S^t, X^T; Y_t|Y^{t-1}) \\ & \overset{(b)}{=} \sum_{t=1}^{T} \frac{1}{T} I^{\mathbf{q}}(S_t, X_t; Y_t|Y^{t-1}) \\ & \overset{(c)}{=} I^{\mathbf{q}}(S_1, X_1; Y_1), \ \forall T \end{aligned}$$

where (a) is due to the chain rule of mutual information and the fact that $S^t$ is a deterministic function of $(S_1, X^{t-1}, Y^{t-1})$ given by the battery update equation (1), (b) is due to the memoryless condition (4), and (c) is due to the invariance property (5). □

We will next develop some further properties of the invariance condition (5). Let us define an auxillary random variable $W_t := S_t - X_t$ where $W_t \in \mathcal{W} := \mathcal{S} - \mathcal{X}$ and for $w \in \mathcal{W}$, let

$$\mathcal{D}(w) := \{(x, s) \in \mathcal{X} \times \mathcal{S} : s - x = w\}.$$

**Lemma III.2.** *An initial battery distribution $\mathbb{P}_{S_1}$ and a stationary memoryless policy $\mathbf{q} = (q, q, \dots)$ satisfies the invariance property* (5) *iff for each $(s_2, y_1) \in \mathcal{S} \times \mathcal{X}$, we have*

$$\mathbb{P}_{S_1}(s_2) Q(y_1) = \sum_{(\tilde{x}, \tilde{s}) \in \mathcal{D}(s_2 - y_1)} q(y_1|\tilde{x}_1, \tilde{s}_1) Q(\tilde{x}_1) \mathbb{P}_{S_1}(\tilde{s}_1). \tag{6}$$

*Proof.* (If) Note that since the rhs is equal to the joint $\mathbb{P}^{\mathbf{q}}(S_2 = s_2, Y_1 = y_1)$, the systems of equations in the Lemma implies that $S_2 \perp Y_1$ and $\mathbb{P}^{\mathbf{q}}_{S_2} = \mathbb{P}_{S_1}$ which is the invariance property (5).

(Only if) Assuming the invariance property to be true, since $S_1 - X_1 = S_2 - Y_1$ given by the battery update equation (1) we must have $\mathbb{P}^{\mathbf{q}}_{Y_1}(y_1) = Q(y_1),\ \forall y_1 \in \mathcal{X}$. Using Bayes rule and the definition of the joint distribution we recover the statement in the Lemma. □

Lemma III.2 implies that the alphabet for $\{Y_t\}_{t>0}$ must be limited to $\mathcal{X}$ and $\mathbb{P}^{\mathbf{q}}_{Y_t} = Q$. In addition, Eq. (6) provides an explicit condition that must be satisfied by the stationary memoryless policies for any fixed $\mathbb{P}_{S_1} \in \mathcal{P}_S$. Note that these are essentially $|\mathcal{W}|$ linear constraints. It should be clear that these constraints are always feasible. For example, using the policy $Y_t = X_t$, any $\mathbb{P}_{S_1}$ will satisfy the invariance property (5). However, this will maximize the leakage rate. We next discuss a policy that turns out be optimal.

### A. Optimal Policy

**Lemma III.3.** *Given a fixed $\mathbb{P}_{S_1}$ and $W_1 = S_1 - X_1$, the optimal policy $\mathbf{q}^* = (q^*, q^*, \dots)$ satisfying the invariance property III.2 is*

$$q^*(y|x,s) = \begin{cases} \frac{Q(y)P_{S_1}(y+s-x)}{P_{W_1}(s-x)} & \text{if } y \in \mathcal{X} \cap \mathcal{Y}_\circ(s-x) \\ 0 & \text{otherwise} \end{cases}$$

*achieving a leakage rate of*

$$L_\infty(\mathbf{q}^*) = I(S_1 - X_1; X_1)$$

*where $(S_1, X_1) \sim \mathbb{P}_{S_1}(s_1)Q(x_1)$.*

*Proof.* By definition, $q^*(y|x,s) \geq 0,\ \forall s \in \mathcal{S}, x \in \mathcal{X}, y \in \mathcal{X} \cap \mathcal{Y}_\circ(s-x)$. Next, we show that $q^*$ is properly normalized.

$$\begin{aligned} &\sum_{\tilde{y} \in \mathcal{X} \cap \mathcal{Y}_\circ(s-x)} Q(\tilde{y}) P_{S_1}(\tilde{y} + s - x) \\ &= \sum_{(\tilde{x}, \tilde{s}) \in \mathcal{D}(s-x)} Q(\tilde{x}) P_{S_1}(\tilde{s}) \\ &= \text{Denominator of } q^*(\mathcal{Y}_\circ(s-x)|x,s), \end{aligned}$$

where the second step follows by substituting $\tilde{x} = \tilde{y}$ and $\tilde{s} = \tilde{y} + s - x$ and observing that $\tilde{s} - \tilde{x} \in \mathcal{D}(s-x)$. Therefore, $\mathbf{q}^*$ is admissible. The invariance property can be verified using Lemma III.2 or as follows:

$$\begin{aligned} &\mathbb{P}^{\mathbf{q}^*}(S_2 = s_2, Y_1 = y_1) \\ &\overset{(a)}{=} \mathbb{P}^{\mathbf{q}^*}(S_2 = s_2, Y_1 = y_1, W_1 = s_2 - y_1) \\ &\overset{(b)}{=} \mathbb{P}^{\mathbf{q}^*}(Y_1 = y_1, W_1 = s_2 - y_1) \\ &= \mathbb{P}^{\mathbf{q}^*}(Y_1 = y_1 | W_1 = s_2 - y_1)\mathbb{P}(W_1 = s_2 - y_1) \\ &\overset{(c)}{=} \mathbb{P}^{\mathbf{q}^*}(Y_1 = y_1 | X_1 = y_1, S_1 = s_2)\mathbb{P}(W_1 = s_2 - y_1) \\ &= q(y_1|y_1, s_2)\mathbb{P}(W_1 = s_2 - y_1) \\ &= Q(y_1)\mathbb{P}_{S_1}(s_2) \end{aligned}$$

where (a) and (b) use the fact that $S_2 - Y_1 = W_1$ holds from the battery update equation, (c) is because $q^*(y|x,s)$ only depends on $(x,s)$ via $s-x$ and the last equality follows from the definition of $q^*$. The last equality shows that the invariance property is satisfied.

To show optimality, fix $\mathbb{P}_{S_1}$ and let $\mathbf{q}$ be any policy satisfying Lemma III.2 and consider the following inequalities:

$$\begin{aligned} L_\infty(\mathbf{q}) &\overset{(a)}{=} I(S_1, X_1; Y_1) \\ &\overset{(b)}{\geq} I(W_1; Y_1) \\ &= H(W_1) - H(W_1 + Y_1 | Y_1) \\ &\overset{(c)}{=} H(W_1) - H(S_2) \\ &\overset{(d)}{=} H(W_1) - H(S_1) \\ &\overset{(e)}{=} H(S_1 - X_1) - H(S_1 - X_1 | X_1) \\ &= I(S_1 - X_1; X_1) \end{aligned}$$

(a) is due to Lemma III.2, (b) is due to the data processing inequality, (c) and (d) are due to the battery update equation (1) and the invariance property of $\mathbf{q}$, and (e) is by definition.

The achievability proof is completed by noting that under $\mathbf{q}^*$, we have $Y_t - W_t - (X_t, S_t)$ and so the lower bound is obtained. □

**Proposition III.1.** *Minimizing over the initial battery distribution $\mathbb{P}_{S_1}$ in Lemma III.3 we obtain the optimal leakage rate in the class of policies satisfying the invariance property III.2.*

*Remark* III.1. The limitation of this achievability scheme requires that the battery have a specific distribution over the battery's initial states. However, this loss of generality is operationally insignificant since the user can start off by randomly charging the battery from an external source.

### B. Converse

So far we have shown that the policy in Lemma III.3, is optimal for the class of invariance policies that satisfy (5). We will now prove an information theoretic converse that establishes that the stated policy is globally optimal among all policies in $\mathcal{Q}_A$. This provides the counterpart of the result in [10], but avoids the use of the dynamic programming framework. Consider the following inequalities: for any admissible policy $\mathbf{q} \in \mathcal{Q}_A$ we have

$$\begin{aligned} &I(S_1, X^T; Y^T) \geq \sum_{t=1}^{T} I(S_t, X_t; Y_t | Y^{t-1}) \geq \sum_{t=1}^{T} I(W_t; Y_t | Y^{t-1}) \\ &= H(W_1) - H(W_1|Y_1) + H(W_2|Y_1) - H(W_2|Y^2) + \cdots \\ &\overset{(a)}{=} H(W_1) - H(S_2|Y_1) + H(S_2 - X_2|Y_1) - H(S_3|Y^2) + \cdots \\ &= H(W_1) + \sum_{t=2}^{T} I(W_t; X_t | Y^{t-1}) - H(W_T|Y^T) \end{aligned}$$

where (a) is because $S_{t+1}$ is an invertible function of $W_t$ given $Y_t$. Now, taking the limit $T \to \infty$ to obtain a lower bound to the leakage rate we have

$$\begin{aligned} L_\infty(\mathbf{q}) &= \lim_{T\to\infty} \frac{1}{T} I(S_1, X^T; Y^T) \\ &\geq \lim_{T\to\infty} \frac{1}{T} \left[ H(W_1) + \sum_{t=2}^{T} I(W_t; X_t|Y^{t-1}) - H(W_T|Y^T) \right] \\ &\overset{(a)}{=} \lim_{T\to\infty} \frac{1}{T} \left[ \sum_{t=2}^{T} I(W_t; X_t|Y^{t-1}) \right] \\ &\overset{(b)}{\geq} \min_{P_S \in \mathcal{P}_S} I(S - X; X). \end{aligned}$$

(a) is because the entropy of any discrete random variable is bounded and (b) follows from the observation that every term in the summation is only a function of the posterior $P(S_t|Y^{t-1})$. Therefore, minimizing each term over a $P_S \in \mathcal{P}_S$ results in a lower bound to the optimal leakage rate which is achievable using Proposition III.1.

## IV. Conclusions

In this paper, we provide a single-letter characterization of the optimal private information leakage rate using information theoretic arguments. While the result was already established in [10], the proof provided in this paper is based on more elementary arguments and avoids the use of the dynamic programming framework. Our proof shows that the optimal leakage rate is achieved using a class of stationary memoryless policies that preserve the posterior distribution of the battery state. We believe that the techniques discussed here also extend to continuous valued input and output alphabets.

## References


[1] A. Predunzi, "A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel," in *Proc. IEEE Power Eng. Society Winter Meeting, New York*, Jan. 2002.

[2] G. W. Hart, "Nonintrusive appliance load monitoring," Proceedings IEEE, vol. 80, no. 12, pp. 1870-1891, Dec. 1992.

[3] H. Y. Lam, G. S. K. Fung, and W. K. Lee, "A novel method to construct taxonomy of electrical appliances based on load signatures," IEEE Trans. user Electronics, vol. 53, no. 2, pp. 653-660, May 2007.

[4] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in Proc. 2nd ACM Workshop Embedded Sensing Systems for Energy-Efficiency in Building, ser. BuildSys 10. New York, NY, USA: ACM, 2010, pp. 61-66.

[5] P. Harsha and M. Dahleh, "Optimal management and sizing of energy storage under dynamic pricing for the efficient integration of renewable energy," in Power Systems, IEEE Transactions on, 30(3):1164-1181, May 2015.

[6] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. A. Lewis, and R. Cepeda, "Privacy for smart meters: towards undetectable appliance load signatures," in Proc. IEEE Smart Grid Commun. Conf., Gaithersburg, Maryland, 2010.

[7] D. Varodayan and A. Khisti, "Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage," in *Proc. IEEE Int'l Conf. Acoust. Speech Sig. Proc. Prague, Czech Republic*, May 2011.

[8] S. Li, A. Khisti, and A. Mahajan, "Structure of optimal privacy-preserving policies in smart-metered systems with a rechargeable battery," Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1-5, Stockholm, Sweden, June 28-July 1, 2015.

[9] J. Yao and P. Venkitasubramaniam, "On the Privacy of an In-Home Storage Mechanism," 52nd Allerton Conference on Communication, Computation and Control, Monticello, IL, October 2013.

[10] S. Li, A. Khisti, A. Mahajan, "Privacy-Optimal Strategies for Smart Metering Systems with a Rechargeable Battery," online, http://arxiv.org/abs/1510.07170,

[11] S. Rajagopalan, L. Sankar, S. Mohajer, and V. Poor, "Smart meter privacy: A utility-privacy framework," in 2nd IEEE International Conference on Smart Grid Communications, 2011.

[12] L. Sankar, S.R. Rajagopalan, and H.V. Poor, "An Information-Theoretic Approach To Privacy," in Proc. 48th Annual Allerton Conf. on Commun., Control, and Computing, Monticello, IL, Sep. 2010, pp. 1220-1227.

[13] D. Gunduz and J.Gomez-Vilardebo, "Smart meter privacy in the presence of an alternative energy source," in *2013 IEEE International Conference on Communications (ICC)*, June 2013.

[14] J. Gomez-Vilardebo and D. Gunduz, "Privacy of smart meter systems with an alternative energy source," in Proc. IEEE Int'l Symp. on Inform. Theory, Istanbul, Turkey, Jul. 2013, pp. 2572-2576.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# A Rate-Distortion Approach to Caching

Roy Timo[†], Shirin Saeedi Bidokhti[*], Michèle Wigger[‡] and Bernhard C. Geiger[†]
[†]Institute for Communications Engineering, Technische Universität München
[*]Department of Electrical Engineering, Stanford University
[‡]Communications and Electronics Department, Telecom ParisTech
{roy.timo, bernhard.geiger}@tum.de, saeedi@stanford.edu, michele.wigger@telecom-paristech.fr

***Abstract*—This paper takes a rate-distortion approach to the caching problem of Maddah-Ali and Niesen. We characterise the optimal tradeoffs between compression rate, reconstruction distortion and cache capacity for a single-user problem and special cases of a two-user problem. These tradeoffs illustrate some interesting connections between optimal caching strategies, Gács-Körner common information, and Wyner's common information.**

## I. Introduction and Setup

We address a communication scenario where users request files from a server during peak-traffic periods. The server reduces the peak-traffic by pre-placing information in cache memories close to the users during prior periods of low traffic. In these low-traffic periods, communication rate is not a limiting resource and the amount of pre-placed information is mainly restricted by the cache memory sizes.

More specifically, in this paper we consider the scenario in Figure 1. The server has access to a library with $L$ files:

$$\text{Library} \quad \mathbf{X} := (X_1^n, X_2^n, \ldots, X_L^n),$$

where each file is a sequence of $n$ symbols

$$X_\ell^n := (X_{\ell,1}, X_{\ell,2}, \ldots, X_{\ell,n})$$

taking value in a finite alphabet $\mathcal{X}_\ell$. For simplicity, we assume that each file is a sequence of independent and identically distributed (i.i.d.) symbols, where symbols pertaining to different files can be correlated:

$$(X_{1,1}, \ldots, X_{L,1}), \ldots, (X_{1,n}, \ldots, X_{L,n}) \quad \text{i.i.d.} \sim P_{\mathbf{X}}, \quad (1)$$

for some given joint law $P_{\mathbf{X}}$ over $\boldsymbol{\mathcal{X}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$.

Assume that there is a single user, which selects an index

$$\ell \in \mathcal{L} := \{1, 2, \ldots, L\}$$

arbitrarily and requests the corresponding file $X_\ell^n$ from the server. The user has a local cache memory of size $nC$ bits where the server can pre-place information $M_c$, and which the user can access to reconstruct its requested file $X_\ell^n$. A central assumption in our work is that the server has to place the information in the cache *before* it learns the user's request. The information $M_c$ stored in the cache should thus be chosen such that it is useful for (or common to) as many files as possible.

Once the server learns the user's request $\ell \in \mathcal{L}$, it sends an $nR$-bit *delivery message* $M$ to the user. Based on this message $M$ and the cache content $M_c$, the user attempts to reconstruct its requested file $X_\ell^n$. Hence, the delivery message $M$ should contain all the information about $X_\ell^n$ that is relevant to the user and that is not yet stored in the cache memory.

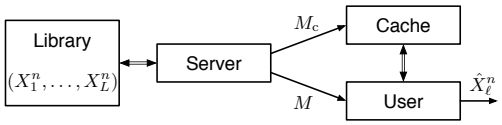



Fig. 1. Single-user RD cache problem.

Such a cache-aided setup was first considered by Maddah-Ali and Niesen in [1, 2] and triggered a series of fruitful results [3]–[9]. The works in [1]–[7] studied the problem where *independent* files $X_1^n, \ldots, X_L^n$ had to be reconstructed losslessly by *multiple* users. More specifically, these works presented various upper and lower bounds on the minimum required delivery-rate $R$ for given cache capacity $C$. While we limit ourselves to a single user with cache memory, we extend the analysis to lossy reconstruction of potentially correlated files, cf. (1). We furthermore analyse the problem when a second user without cache memory is present, see the setup in Figure 4.

The main problem of interest in this paper is thus the optimal tradeoff between the *delivery rate* $R$, the *cache capacity* $C$, and the user's *reconstruction distortion*. Notice that the delivery rate $R$ is a *worst-case* rate (or a compound rate) in the sense that it has to be sufficiently large so that the user can reconstruct every file $X_\ell^n$, $\ell \in \mathcal{L}$, with desired accuracy. The problem setup by Wang, Lim and Gastpar [9], can be considered as an ergodic average-case equivalent of our worst-case (or compound) setup.

## II. Single User

### A. Formal Problem Definition

Let $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_L$ be given reconstruction sets. A *joint rate-distortion-cache (RDC) code* for a given blocklength $n$ consists of $(2L+1)$-mappings:

(i) A *cache encoder* $f_{\mathrm{c}} : \boldsymbol{\mathcal{X}}^n \to \mathcal{M}_{\mathrm{c}}$, where $\mathcal{M}_{\mathrm{c}}$ is finite.
(ii) A *file encoder* $f_\ell : \boldsymbol{\mathcal{X}}^n \to \mathcal{M}$ for each $\ell \in \mathcal{L}$, where $\mathcal{M}$ is finite.
(iii) A *file decoder* $g_\ell : \mathcal{M} \times \mathcal{M}_{\mathrm{c}} \to \hat{\mathcal{X}}_\ell^n$ for each $\ell \in \mathcal{L}$.

For brevity, we will call the above collection of encoders and decoders an $(n, \mathcal{M}, \mathcal{M}_{\mathrm{c}})$-code. Given demand $\ell \in \mathcal{L}$, the cache content and the delivery message are

$$M_{\mathrm{c}} := f_{\mathrm{c}}(\mathbf{X}^n) \qquad \text{and} \qquad M := f_\ell(\mathbf{X}^n);$$

and the user's reconstruction is

$$\hat{X}_\ell^n := g_\ell(M, M_\mathrm{c}) \in \hat{\mathcal{X}}_\ell^n.$$

As per the usual rate-distortion (RD) paradigm, let us assume that the quality of $\hat{X}_\ell^n$ can be meaningfully measured using average per-letter distortions. Specifically, for each $\ell \in \mathcal{L}$, let

$$\delta_\ell : \hat{\mathcal{X}}_\ell \times \mathcal{X}_\ell \to [0, \infty)$$

be a bounded *distortion function*. For simplicity, we assume that for each symbol $x_\ell \in \mathcal{X}_\ell$ there always exists an $\hat{x}_\ell$ in $\hat{\mathcal{X}}_\ell$ such that $\delta_\ell(\hat{x}_\ell, x_\ell) = 0$.

*Definition 1:* Let $\mathbf{D} := (D_1, D_2, \ldots, D_L)$ and $C$ be arbitrary nonnegative reals. We say that a delivery rate $R \geq 0$ is $(\mathbf{D}, C)$*-admissible* if for every $\epsilon > 0$ there exists a sufficiently large blocklength $n$ and an $(n, \mathcal{M}, \mathcal{M}_\mathrm{c})$-code such that

$$\forall\ \ell \in \mathcal{L}\colon \quad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \delta_\ell(\hat{X}_{\ell,i}, X_{\ell,i})\right] \leq D_\ell + \epsilon,$$

and

$$|\mathcal{M}_\mathrm{c}| \leq 2^{n(C+\epsilon)} \quad \text{and} \quad |\mathcal{M}| \leq 2^{n(R+\epsilon)}. \tag{2}$$

We call $C$ the *cache capacity* and $\mathbf{D}$ the *distortion constraints*. The optimal RDC tradeoff for blocklengths $n \to \infty$ is characterised by the following function.

*Definition 2:* The RDC function is

$$\mathsf{R}(\mathbf{D}, C) := \inf\big\{R \geq 0 : R \text{ is } (\mathbf{D}, C)\text{-admissible}\big\}.$$

### B. Main Results

The RDC function has the following properties:

*Proposition 1:*

(i) $\mathsf{R}(\mathbf{D}, C)$ is jointly convex and non-increasing in $\mathbf{D}$ and $C$.

(ii) If $C \geq H(\mathbf{X})$, then $\mathsf{R}(\mathbf{D}, C) = 0$ for all $\mathbf{D}$.

(iii) If $C = 0$, then

$$\mathsf{R}(\mathbf{D}, 0) = \max_{\ell \in \mathcal{L}} \mathsf{R}_{X_\ell}(D_\ell),$$

where $\mathsf{R}_{X_\ell}(D_\ell)$ is the usual RD function for $X_\ell$,

$$\mathsf{R}_{X_\ell}(D_\ell) := \min_{\substack{p_{\hat{X}_\ell|X_\ell}\colon\ \mathcal{X}_\ell \to \hat{\mathcal{X}}_\ell \\ \text{s.t. } \mathbb{E}[\delta_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell}} I(X_\ell; \hat{X}_\ell).$$

Let

$$\mathsf{R}^*(\mathbf{D}, C) := \min \max_{\ell} I(X_\ell; \hat{X}_\ell | U), \tag{3}$$

where the minimum is taken over all $(U, \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_L)$ jointly distributed with $\mathbf{X}$ such that $I(\mathbf{X}; U) \leq C$ and $\mathbb{E}[\delta_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell$ for all $\ell \in \mathcal{L}$.

*Theorem 1:*

$$\mathsf{R}(\mathbf{D}, C) = \mathsf{R}^*(\mathbf{D}, C).$$

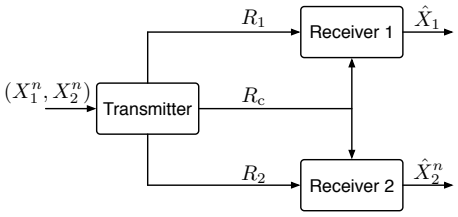



Fig. 2. Lossy Source Coding for a Simple Network.

### C. Connections to the Gray-Wyner Network

For the case of $L = 2$ files, $\mathbf{X}^n = (X_1^n, X_2^n)$, there is a close connection between the RDC function and Gray and Wyner's classic *"source coding for a simple network"* problem [10]. The Gray-Wyner network is illustrated in Figure 2: A transmitter is connected to two different receivers via a common link of rate $R_\mathrm{c}$ and two private links of rates $R_1$ and $R_2$. The set of all achievable rate tuples $(R_\mathrm{c}, R_1, R_2)$ for which receivers 1 and 2 can respectively reconstruct $X_1^n$ and $X_2^n$ to within distortions $D_1$ and $D_2$ is given by [10, Thm. 8]

$$\mathcal{R}_\mathrm{GW}(D_1, D_2) := \bigcup \left\{ (R_\mathrm{c}, R_1, R_2) : \begin{array}{r} R_\mathrm{c} \geq I(X_1, X_2; U) \\ R_1 \geq I(X_1; \hat{X}_1 | U) \\ R_2 \geq I(X_2; \hat{X}_2 | U) \end{array} \right\},$$

where the union is over all tuples $(X_1, X_2, U, \hat{X}_1, \hat{X}_2)$ satisfying $\mathbb{E}[\delta_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell$, for $\ell \in \{1, 2\}$. The next proposition can be proved by associating the common rate $R_c$ of the Gray-Wyner problem with the rate of the caching message $M_c$, and the two private rates $R_1$ and $R_2$ of the Gray-Wyner problem with the rates of our delivery message $M$ when the user demands $X_1^n$ and $X_2^n$, respectively.

*Proposition 2:*

$$\mathsf{R}((D_1, D_2), C) = \min_{(C, R_1, R_2) \in \mathcal{R}_\mathrm{GW}(D_1, D_2)} \max\big\{R_1, R_2\big\}.$$

### D. Almost Lossless Compression

Let us now restrict attention to the case where the user wants to reconstruct $X_\ell^n$ (almost) losslessy. Specifically, suppose that $\hat{\mathcal{X}}_\ell = \mathcal{X}_\ell$ and $\delta_\ell(\hat{x}_\ell, x_\ell) = \mathbb{1}\{\hat{x}_\ell \neq x_\ell\}$ for all $\ell \in \mathcal{L}$ are Hamming distortion functions; and $\mathbf{0} := (0, \ldots, 0)$ is a tuple of $L$ zeros. Given these assumptions, define the rate-cache (RC) function

$$\mathsf{R}_\mathbf{0}(C) := \mathsf{R}(\mathbf{0}, C).$$

From Theorem 1 we have the next corollary.

*Corollary 1.1:*

$$\mathsf{R}_\mathbf{0}(C) = \mathsf{R}^*(\mathbf{0}, C) = \min_{U} \max_{\ell} H(X_\ell | U),$$

where the minimum is taken over all auxiliary random variables $U$, jointly distributed with $\mathbf{X}$, satisfying $I(\mathbf{X}; U) \leq C$.

Figure 3 shows the typical behaviour of $\mathsf{R}_\mathbf{0}(C)$. To obtain better understanding, we propose two lower bounds and study conditions when they are tight.

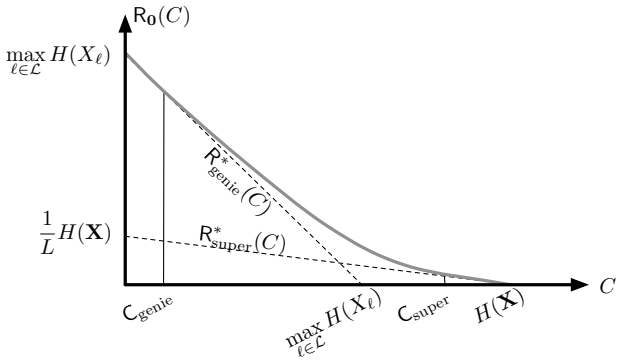



Fig. 3. An illustration of a typical RC function $\mathsf{R}_{\mathbf{0}}(C)$ and the lower bounds in Propositions 3 and 6.

*1) Lower Bound $\mathsf{R}^*_{\mathbf{0},\text{Genie}}(C)$ on $\mathsf{R}_{\mathbf{0}}(C)$:* Imagine that, before the caching phase, a genie tells the server which $\ell \in \mathcal{L}$ the user will select in the future. The optimal caching strategy for this hypothetical *genie-aided* problem is obvious, because for each $\ell \in \mathcal{L}$ we have a standard RD problem: The server uses an optimal code to losslessly compress the source $X_\ell^n$, stores the first $nC$ bits produced by this code in the user's cache memory, and sends the remaining bits as the delivery message. The user assembles the bits from the cache memory and the delivery message and reconstructs the requested file. The RC function of this genie-aided system, $\mathsf{R}_{\mathbf{0},\text{Genie}}(C)$, hence equals[1]

$$\mathsf{R}_{\mathbf{0},\text{Genie}}(C) = \mathsf{R}^*_{\mathbf{0},\text{Genie}}(C) := \max\big\{0,\ \max_{\ell\in\mathcal{L}} H(X_\ell) - C\big\}.$$

Since the server can always choose to ignore the genie-information, the RC function of the genie-aided system cannot exceed the RC function of the original scenario:

*Proposition 3:*

$$\mathsf{R}_{\mathbf{0}}(C) \geq \mathsf{R}_{\mathbf{0},\text{Genie}}(C).$$

For degraded file sets, above lower bound is tight.

*Example 1:* Let the DMS $\mathbf{X}$ be given by $X_\ell := (A_1, \ldots, A_\ell)$ for all $\ell \in \mathcal{L}$, where $(A_1, \ldots, A_L)$ have an arbitrary joint distribution. Then,

$$\mathsf{R}_{\mathbf{0}}(C) = \mathsf{R}^*_{\mathbf{0},\text{Genie}}(C) = \max\{0, H(X_L) - C\}.$$

*2) Connection to the Gács-Körner Common Information:* The lower bound $\mathsf{R}^*_{\mathbf{0},\text{Genie}}(C)$ is also trivially tight at zero cache capacity, $C = 0$; for example, see Assertion (ii) in Proposition 1. It is therefore natural to define

$$C_{\text{Genie}} := \sup\Big\{C \leq H(\mathbf{X}) : \mathsf{R}_{\mathbf{0}}(C) = \mathsf{R}^*_{\mathbf{0},\text{Genie}}(C)\Big\}$$

to be the largest cache capacity for which there is *no rate loss* with respect to the optimal genie-aided system.

Define the subset $\mathcal{L}^* \subseteq \mathcal{L}$ as

$$\mathcal{L}^* := \arg\max_{\ell\in\mathcal{L}} H(X_\ell).$$

[1]The maximum over $\ell \in \mathcal{L}$ is needed because we again consider a worst-case (compound) setup over all possible demands $\ell \in \mathcal{L}$.

Further, let

$$C^*_{\text{Genie}} := \max_U\ I(\mathbf{X}; U),$$

where the maximum is taken over all auxiliary random variables $U$ jointly distributed with $\mathbf{X}$ for which the following statements hold:

(i) For every $\ell^* \in \mathcal{L}^*$, we have $U \leftrightarrow X_{\ell^*} \leftrightarrow X_{\mathcal{L}\setminus\ell^*}$, where $X_{\mathcal{L}\setminus\ell^*} := \big(X_1, X_2, \ldots, X_{\ell^*-1}, X_{\ell^*+1}, \ldots, X_L\big)$.

(ii) For every $\ell^* \in \mathcal{L}^*$,

$$H(X_{\ell^*}|U) = \max_{\ell\in\mathcal{L}} H(X_\ell|U),$$

(iii) $U$ is defined on an alphabet $\mathcal{U}$ with $|\mathcal{U}| \leq |\boldsymbol{\mathcal{X}}| + |\mathcal{L}^*| + L$.

*Proposition 4:*

$$C_{\text{Genie}} = C^*_{\text{Genie}}.$$

The critical cache capacity $C^*_{\text{Genie}}$ is related to the natural $L$-variable generalisation [12] of Gács and Körner's *common information*:

$$\mathsf{K}^*_{\text{GK}} := \max_{U:\ H(U|X_\ell)=0,\ \forall\ \ell\in\mathcal{L}} H(U).$$

*Proposition 5:*

$$C^*_{\text{Genie}} \geq \mathsf{K}^*_{\text{GK}}. \tag{4}$$

If $H(X_1) = \cdots = H(X_L)$, then (4) holds with equality.

*3) Lower Bound $\mathsf{R}^*_{\mathbf{0},\text{Super}}(C)$ on $\mathsf{R}_{\mathbf{0}}(C)$:* Now imagine a situation where we have a *superuser* that requests all the $L$ sources $X_1^n, \ldots, X_L^n$ and that obtains $L$ *delivery messages* of rate $R$ each. Moreover, suppose that as before this superuser has a local cache memory of size $nC$ bits that can be filled by the server. The optimal strategy for this superuser problem is again obvious, since it is equivalent to a standard RD problem with a single compression message of rate $LR+C$: The server takes an optimal code to compress the entire library $\mathbf{X}^n$ and distributes the produced bits in the cache memory and over the $L$ delivery messages. The RC function of this superuser system, $\mathsf{R}_{\mathbf{0},\text{Super}}(C)$, hence is:

$$\mathsf{R}_{\mathbf{0},\text{Super}}(C) = \mathsf{R}^*_{\mathbf{0},\text{Super}}(C) := \max\left\{0,\ \frac{1}{L}\big(H(\mathbf{X}) - C\big)\right\}.$$

If one limits the superuser to reconstruct each source $X_\ell^n$, $\ell \in \mathcal{L}$, solely based on the content in the cache memory and the $\ell$-th delivery message, one obtains our original setup. The RC function of the superuser system thus can not exceed the RC function of the original setup:

*Proposition 6:*

$$\mathsf{R}_{\mathbf{0}}(C) \geq \mathsf{R}_{\mathbf{0},\text{Super}}(C).$$

For independent and identically distributed files, above lower bound is tight:

*Example 2:* Let the DMS $\mathbf{X}$ follow the product distribution $P_{\mathbf{X}} = \prod_{\ell=1}^{L} P_X$. In this case,

$$\mathsf{R}_{\mathbf{0}}(C) = \mathsf{R}^*_{\mathbf{0},\text{Super}}(C) = \max\left\{0, H(X) - \frac{C}{L}\right\}.$$

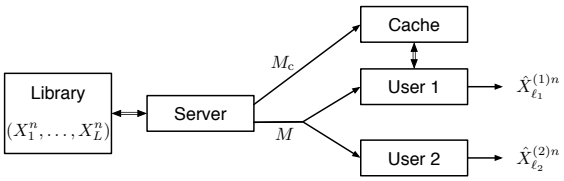



Fig. 4. Two-user RD cache problem.

*4) Connection to Wyner's Common Information:* The superuser lower bound is trivially tight when $C \geq H(\mathbf{X})$. So it is natural to consider the smallest cache capacity for which there is *no rate loss* with respect to the optimal superuser system,

$$C_{\text{Super}} := \inf\big\{C \geq 0 : \mathsf{R}_{\mathbf{0}}(C) = \mathsf{R}_{\mathbf{0},\text{Super}}(C)\big\}.$$

Let

$$C^*_{\text{Super}} := \min_U I(\mathbf{X}; U),$$

where the minimum is taken over all auxiliary random variables $U$ jointly distributed with $\mathbf{X}$ such that

(i) $X_\ell \leftrightarrow U \leftrightarrow X_{\mathcal{L}\setminus\ell}$ for all $\ell \in \mathcal{L}$;
(ii) $H(X_1|U) = \cdots = H(X_L|U)$; and
(iii) $U$ is defined on $\mathcal{U}$ with $|\mathcal{U}| \leq |\boldsymbol{\mathcal{X}}| + 2L$.

*Proposition 7:*

$$C_{\text{Super}} = C^*_{\text{Super}}.$$

The critical cache capacity $C^*_{\text{Super}}$ is related to the natural $L$-variable generalisation [11] of *Wyner's common information*:

$$\mathsf{K}^*_{\mathsf{W}}(\mathbf{X}) := \min_U \; I(\mathbf{X}; U),$$

where the minimum is taken over all $U$ jointly distributed with $\mathbf{X}$ for which

(i) $X_\ell \leftrightarrow U \leftrightarrow X_{\mathcal{L}\setminus\ell}$ for all $\ell \in \mathcal{L}$; and
(ii) $U$ is defined on an alphabet $\mathcal{U}$ with $|\mathcal{U}| \leq |\boldsymbol{\mathcal{X}}| + L$.

*Proposition 8:*

$$C^*_{\text{Super}} \geq \mathsf{K}^*_{\mathsf{W}}.$$

If the source $\mathbf{X}$ is sufficiently symmetric, above inequality holds with equality.

## III. Two-Users with One Cache

### A. Setup

We now consider a two-user extension of the problem in Section II. Let us assume that user 1 has a cache with capacity $C$, while user 2 does not have a cache; see Figure 4. The library consists of the same $L$ files $\mathbf{X}^n := (X_1^n, \ldots, X_L^n)$ used in Section II, and communication again takes place in two phases — a *caching phase* and a *delivery phase*. Let $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{L}$ denote those indices that can be potentially selected by users 1 and 2, respectively. That is, user $k$ (for $k = 1, 2$) will request a file from $\big\{X_{\ell_k}^n : \ell_k \in \mathcal{L}_k\big\}$. Let $L_1 := |\mathcal{L}_1|$ and $L_2 := |\mathcal{L}_2|$.

A *two-user joint RDC code* with blocklength $n$ consists of

(i) A *cache encoder*

$$f_{\text{c}} \colon \boldsymbol{\mathcal{X}}^n \to \mathcal{M}_{\text{c}}.$$

(ii) A *file encoder*

$$f_{(\ell_1,\ell_2)} \colon \boldsymbol{\mathcal{X}}^n \to \mathcal{M}, \quad (\ell_1, \ell_2) \in \mathcal{L}_1 \times \mathcal{L}_2.$$

(iii) A *user-1 file decoder*

$$g^{(1)}_{\ell_1,\ell_2} : \mathcal{M} \times \mathcal{M}_{\text{c}} \to \hat{\mathcal{X}}_{\ell_1}^{(1),n}, \quad (\ell_1, \ell_2) \in \mathcal{L}_1 \times \mathcal{L}_2.$$

(iv) A *user-2 file decoder*

$$g^{(2)}_{\ell_1,\ell_2} : \mathcal{M} \to \hat{\mathcal{X}}_{\ell_2}^{(2),n}, \quad (\ell_1, \ell_2) \in \mathcal{L}_1 \times \mathcal{L}_2.$$

Notice that we allow the decoders to depend on the demands of both users. We call the above collection of encoders and decoders an $(n, \mathcal{M}, \mathcal{M}_{\text{c}})$-two-user-code.

During the caching phase, the server pre-places the message $M_{\text{c}} := f_{\text{c}}(\mathbf{X}^n)$ in the cache of user 1. After the demands $(\ell_1, \ell_2) \in \mathcal{L}_1 \times \mathcal{L}_2$ are revealed to the server and both users, the server sends the message $M := f_{(\ell_1,\ell_2)}(\mathbf{X}^n)$ to both users. Users 1 and 2 respectively output

$$\hat{X}_{\ell_1}^{(1),n} := g^{(1)}_{\ell_1,\ell_2}(M, M_{\text{c}}) \quad \text{and} \quad \hat{X}_{\ell_2}^{(2),n} := g^{(2)}_{\ell_1,\ell_2}(M).$$

For convenience, we index user 1's reconstruction sequence only with its own demand $\ell_1$; it can however also depend on user 2's demand $\ell_2$. Similarly, for user 2's reconstruction.

The users might have differing exigencies regarding the files in the library. To account for this, we admit both users to measure reconstruction accuracy with different bounded per-letter distortion functions $\delta_{\ell_1}^{(1)} \colon \hat{\mathcal{X}}_{\ell_1}^{(1)} \times \mathcal{X}_{\ell_1} \to [0, \infty)$ and $\delta_{\ell_2}^{(2)} \colon \hat{\mathcal{X}}_{\ell_2}^{(2)} \times \mathcal{X}_{\ell_2} \to [0, \infty)$ (for indices $\ell_1 \in \mathcal{L}_1$ and $\ell_2 \in \mathcal{L}_2$).

*Definition 3:* Let $C$ be a nonnegative real number, and let $\mathbf{D}^{(1)} := \big\{D_{\ell_1}^{(1)}\big\}_{\ell_1 \in \mathcal{L}_1}$ and $\mathbf{D}^{(2)} := \big\{D_{\ell_2}^{(2)}\big\}_{\ell_2 \in \mathcal{L}_2}$ be $L_1$- and $L_2$-tuples of nonnegative real numbers.

We say that a compression rate $R \geq 0$ is $(\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, C)$-*admissible* if for any $\epsilon > 0$ there exists a sufficiently large blocklength $n$ and an $(n, \mathcal{M}, \mathcal{M}_{\text{c}})$-code satisfying (2) and

$$\forall\, k \in \{1,2\}\colon \forall\, \ell \in \mathcal{L}_k\colon \quad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \delta_{\ell_k}^{(k)}\Big(\hat{X}_{\ell,i}^{(k)}, X_{\ell,i}\Big)\right] \leq D_\ell^{(k)} + \epsilon. \quad (5)$$

*Definition 4:* The *two-user RDC function* is

$$\mathsf{R}_{\text{2user}}(\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, C) := \inf\big\{R \geq 0 : R \text{ is } \big(\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, C\big)\text{-admissible}\big\}.$$

### B. Genie-Aided Lower Bound on the RDC Function

If both users' demands were revealed by a genie to the server even before the caching phase, our setup would coincide with a "worst-case" (or compound) successive-refinement setup. The rate-distortions function of this worst-demands successive refinement problem thus forms a lower bound on $\mathsf{R}_{\text{2user}}(\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, C)$.

*Definition 5:* Let $\mathsf{R}^*_{\text{SuccRef}}(\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, C)$ be the RDC function defined in (6) on top of the next page, where the minimum is taken over all tuples $(\mathbf{X}, \hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)})$ such that for $k \in \{1, 2\}$:

$$\forall\, \ell \in \mathcal{L}_k\colon \quad \mathbb{E}\Big[\delta_\ell^{(k)}\big(\hat{X}_\ell^{(k)}, X_\ell\big)\Big] \leq D_\ell^{(k)}. \quad (7)$$

$$\mathsf{R}^*_{\text{SuccRef}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big) := \max_{(\ell_1,\ell_2)\in\mathcal{L}_1\times\mathcal{L}_2} \min_{P_{\mathbf{X},\hat{\mathbf{X}}^{(1)},\hat{\mathbf{X}}^{(2)}}} \max\Big\{I\big(\mathbf{X};\hat{X}^{(2)}_{\ell_2}\big),\ I\big(\mathbf{X};\hat{X}^{(1)}_{\ell_1},\hat{X}^{(2)}_{\ell_2}\big)-C\Big\} \tag{6}$$

$$\mathsf{R}_{\text{2user,Ach}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big) := \min \max_{(\ell_1,\ell_2)\in\mathcal{L}_1\times\mathcal{L}_2} \max\Big\{I\big(\mathbf{X};\hat{X}^{(2)}_{\ell_2}\big)+I\big(\mathbf{X};\hat{X}^{(1)}_{\ell_1}|U,\hat{X}^{(2)}_{\ell_2}\big),\ I\big(\mathbf{X};U,\hat{X}^{(1)}_{\ell_1},\hat{X}^{(2)}_{\ell_2}\big)-C\Big\} \tag{8}$$

$$\mathsf{R}_{\text{2user}}(\mathbf{0},\mathbf{0},C) \le \min_{P_{U|\mathbf{X}}} \max_{(\ell_1,\ell_2)\in\mathcal{L}_1\times\mathcal{L}_2} \max\Big\{H(X_{\ell_2})+H(X_{\ell_1}|U,X_{\ell_2}),\ H(U,X_{\ell_1},X_{\ell_2})-C\Big\} \tag{9}$$

*Theorem 2:*

$$\mathsf{R}_{\text{2user}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big) \ge \mathsf{R}^*_{\text{SuccRef}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big).$$

### C. Upper Bound on the RDC Function

We have the following upper bound on the RDC function.

*Definition 6:* Let $\mathsf{R}_{\text{2user,Ach}}(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C)$ be defined as in (8) on top of the next page, where the minimum is taken over all tuples $\big(U,\ \hat{\mathbf{X}}^{(1)}{:=}\{\hat{X}^{(1)}_{\ell_1}\}_{\ell_1\in\mathcal{L}_1},\hat{\mathbf{X}}^{(2)}{:=}\{\hat{X}^{(2)}_{\ell_2}\}_{\ell_2\in\mathcal{L}_2}\big)$ jointly distributed with $\mathbf{X}$ such that (7) holds for $k\in\{1,2\}$.

*Theorem 3:*

$$\mathsf{R}_{\text{2user}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big) \le \mathsf{R}_{\text{2user,Ach}}\big(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C\big).$$

Theorem 3 can equivalently be stated as follows: a rate $R>0$ is $(\mathbf{D}^{(1)},\mathbf{D}^{(2)},C)$-admissible whenever there is a tuple $\big(U,\ \hat{\mathbf{X}}^{(1)},\hat{\mathbf{X}}^{(2)}\big)$ and a collection of auxiliary rates $\{\tilde{R}_{\ell_2}\}_{\ell_2\in\mathcal{L}_2}$ such that for every pair $(\ell_1,\ell_2)\in\mathcal{L}_1\times\mathcal{L}_2$:

$$C+\tilde{R}_{\ell_2} \ge I\big(U;\mathbf{X},\hat{X}^{(2)}_{\ell_2}\big)-I\big(U;\hat{X}^{(2)}_{\ell_2}\big)=I\big(U;\mathbf{X}\big|\hat{X}^{(2)}_{\ell_2}\big)$$
$$R-\tilde{R}_{\ell_2} \ge I\big(\mathbf{X};\hat{X}^{(2)}_{\ell_2}\big)+I\big(\mathbf{X};\hat{X}^{(1)}_{\ell_1}|U,\hat{X}^{(2)}_{\ell_2}\big).$$

These rates are achieved by the following scheme. The server compresses the entire library $\mathbf{X}^n$ into $U^n$ using the *adaptive conditional RD code* for side-information $\hat{X}^n_{\ell_2}$ that we describe in the next paragraph. Our adaptive RD code produces a first message of $nC$ bits which the server stores in user 1's cache, and a second message of $n\tilde{R}_{\ell_2}$ bits which the server sends as part of the delivery message. In the delivery message it also sends a standard RD message that allows both users to reconstruct $\hat{X}^{(2),n}_{\ell_2}$, and a standard conditional RD message that allows user 1 to reconstruct $\hat{X}^{(1),n}_{\ell_1}$ given that it already knows $(U^n,\hat{X}^{(2),n}_{\ell_2})$. Both users first reconstruct $\hat{X}^{(2),n}_{\ell_2}$. User 1 subsequently reconstructs $U^n$ and $\hat{X}^{(1),n}_{\ell_1}$, always using previously reconstructed sequences as side-information.

Our adaptive conditional RD code uses a codebook $\mathcal{C}:=\{U^n(m_u)\}$ with a nested binning structure: it contains $\approx 2^{nC}$ outer bins that each consist of $\approx 2^{n\tilde{R}_{\ell_2}}$ inner bins. The outer binning rate $C$ is fixed in advanced; the inner binning rate however adapts to the quality of the side-information $\hat{X}^{(2),n}_{\ell_2}$ and is fixed only after the demand $\ell_2$ is revealed. Encoding is in two steps. In a first step the server picks the unique codeword $U^n(m^*_u)$ that for every $\ell_2\in\mathcal{L}_2$ is jointly typical with the pair $\big(\mathbf{X}^n,\hat{X}^{(2),n}_{\ell_2}\big)$. The outer bin index of $U^n(m^*_u)$ is immediately available and the server stores the $nC$ bits representing this index in user 1's cache. Once the demand $\ell_2$ is fixed, also the inner bin index is available and the server sends it as part of the delivery message. Decoding is standard using both bin indices and the side-information $\hat{X}^{(2),n}_{\ell_2}$.

### D. Almost Lossless Reconstructions

Let now both users reconstruct their demanded files $X^n_{\ell_1}$ and $X^n_{\ell_2}$ (almost) losslessy. From Theorem 3:

*Corollary 3.1:* The RC-function for the lossless setup satisfies the upper bound in (9) on top of this page.

*Corollary 3.2:* Bound (9) holds with equality when

1) $\mathcal{L}_1=\mathcal{L}_2=\{\ell,\ell'\}$ for $\ell,\ell'\in\mathcal{L}$;
2) $\mathcal{L}_1=\{\ell\}$ for some $\ell\in\mathcal{L}$; or
3) $\mathcal{L}_2=\{\ell\}$ for some $\ell\in\mathcal{L}$.

*Proof:* To prove cases 1.) and 2.), specialise the lower bound in Theorem 2 to the lossless case and to $U=(X_\ell,X_{\ell'})$ and $U=X_\ell$, respectively. For case 3.) a new converse is required. ∎

Interestingly, in the first two cases there is no penalty for not knowing the demands during the caching phase.


## ACKNOWLEDGEMENTS

The work of R. Timo was supported by the Alexander von Humboldt Foundation. The work of S. Saeedi Bidokhti was supported by the Swiss National Science Foundation fellowship no. 158487.


## REFERENCES


[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[2] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *arXiv:1308.0178v3*, 2013.

[3] S. Sahraei and M. Gastpar, "K users caching two files: an improved achievable rate", *online* http://arxiv.org/abs/1512.06682v1.

[4] Z. Chen, P. Fan, and K. Ben Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *online* http://arXiv.org/abs/1407.1935v2.pdf, Nov. 2015.

[5] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *online* http://arxiv.org/abs/1501.06003.

[6] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *IEEE Intl. Symp. Inform. Theory*, Hong Kong, China, 2015.

[7] C. Tian, "A note on the fundamental limits of coded caching," *online* http://arXiv.org/abs/1503.00010.

[8] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in *IEEE Intl. Symp. Wireless Commun. Systems*, Brussels, Belgium, 2015.

[9] C. Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: sequential coding for computing," *arXiv:1504.00553*, 2015.

[10] R. Gray and A. Wyner, "Source coding for a simple network," *Bell Sys. Tech. J.*, vol. 53, no. 9, pp. 1681–1721, 1974.

[11] W. Liu, G. Xu, and B. Chen, "The common information of N dependent random variables," in *Allerton Conf. Commun. Control Comp.*, 2010.

[12] R. Tandon, L. Sankar, and H.V. Poor, "Multi-user privacy: The Gray-Wyner system and generalized common information" in *Proc. IEEE Int. Sym. on Information Theory*, 2011

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Improving on the Cut-Set Bound via a Geometric Analysis of Typical Sets

Xiugang Wu
Stanford University
x23wu@stanford.edu

Ayfer Özgür
Stanford University
aozgur@stanford.edu

***Abstract*—We consider the discrete memoryless symmetric primitive relay channel, where, a source $X$ wants to send information to a destination $Y$ with the help of a relay $Z$ and the relay can communicate to the destination via an error-free digital link of rate $R_0$, while $Y$ and $Z$ are conditionally independent and identically distributed given $X$. We develop two upper bounds on the capacity of this channel that are tighter than existing bounds, including the celebrated cut-set bound. Our approach significantly differs from the standard information-theoretic approach for proving upper bounds on the capacity of multi-user channels. We build on the blowing-up lemma to analyze the probabilistic geometric relations between the typical sets of the $n$-letter random variables associated with a reliable code for communicating over this channel. These relations translate to new entropy inequalities between the $n$-letter random variables involved.**

## I. Introduction

Characterizing the capacity of relay channels [1] has been a long-standing open problem in network information theory. The seminal work of Cover and El Gamal [2] has introduced two basic achievability schemes: Decode-and-Forward and Compress-and-Forward, and derived a general upper bound on the capacity, now known as the cut-set bound. Over the last decade, significant progress has been made on the achievability side: these schemes have been extended and unified to multi-relay networks [3]–[4] and many new relaying strategies have been discovered, such as Amplify-and-Forward, Compute-and-Forward, Noisy Network Coding etc. [5]–[7]. However, the progress on developing upper bounds that are tighter than the cut-set bound has been relatively limited. In particular, in most of the special cases where the capacity is known, the upper bound is given by the cut-set bound [2], [8]–[10].

In general, however, the cut-set bound is known to be not tight. Specifically, consider the primitive relay channel depicted in Fig. 1, where the source's input $X$ is received by the relay $Z$ and the destination $Y$ through a channel $p(y, z|x)$, and the relay $Z$ can communicate to the destination $Y$ via an error-free digital link of rate $R_0$. When $Y$ and $Z$ are conditionally independent given $X$, and $Y$ is a stochastically degraded version of $Z$, Zhang [11] used the blowing up lemma to show that the inequality between the capacity and the cut-set bound is indeed strict in certain regimes of this channel. However, Zhang's result does not provide any information regarding the gap or suggest a way to compute it. For a special case of the primitive relay channel where the noise is modulo additive and $Z$ is a corrupted version of the noise for the $X$-$Y$ link, Aleksic, Razaghi and Yu characterize the capacity and show that it is strictly lower than the cut-set bound [12]. While this result provides an exact capacity characterization for a non-trivial special case, it builds strongly on the peculiarity of the channel model and in this respect its scope is more limited than Zhang's result.

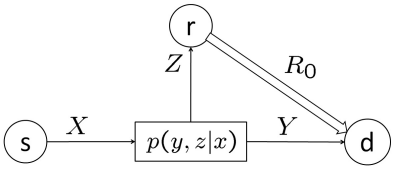



Fig. 1. Primitive relay channel.

More recently, a new upper bound demonstrating an explicit gap to the cut-set bound was developed by Xue [13] for general primitive relay channels. Xue's bound relates the gap of the cut-set bound to the reliability function of the $X$-$Y$ link. Unlike Zhang's result, Xue's bound can be numerically computed. While it is strictly tighther than the cut-set bound in certain regimes of the primitive relay channel, with an explicit computable gap, it can also be looser than the cut-set bound.

In [14], we presented two new upper bounds on the capacity of the primitive relay channel. The first of these bounds can be regarded as a direct improvement of Xue's bound. It is indeed strictly tighter than both Xue's bound and the cut-set bound and like Xue's bound involves the reliability function of the $X$-$Y$ link. Our second bound was based on a new set of arguments and is structurally different than the first one. It can be significantly tighter than our first bound as demonstrated in [14] for the binary symmetric relay channel, however it can be also looser than it for some other channel models.

The current paper is a continuation of our work in [14]. We present a new bound which is strictly tighter than our first bound in [14] and is also structurally different from it. In particular, it does not involve the reliability function of the $X$-$Y$ link but is structurally closer to our second bound in [14]. The more important contribution of this paper is to distill a new proof technique which significantly differs from existing

This work was supported in part by the NSF CAREER award 1254786 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

converse approaches in the literature and can be potentially useful for other multi-user problems. In general, proving an upper bound on the capacity of a multi-user channel involves dealing with entropy relations between the various $n$-letter random variables induced by the reliable code and the channel structure (together with using Fano's inequality). In order to prove the desired relations between the entropies of the $n$-letter random variables involved, in this paper we consider their $B$-letter i.i.d. extensions (leading to length $B$ i.i.d. sequences of $n$-letter random variables). We then use the blowing up lemma to analyze the geometry of the typical sets associated with these $B$-letter sequences. We present two different ways to translate the (probabilistic) geometric relations between these typical sets into new entropy relations between the random variables involved. This leads to two different bounds on the capacity of the primitive relay channel which do not include each other in general. As pointed out before, the first of these bounds is new to this paper, the second one recovers the second bound we presesented in [14].

## II. Preliminaries

Consider a primitive relay channel as depicted in Fig. 1. The source's input $X$ is received by the relay $Z$ and the destination $Y$ through a channel

$$(\Omega_X, p(y,z|x), \Omega_Y \times \Omega_Z)$$

where $\Omega_X, \Omega_Y$ and $\Omega_Z$ are finite sets denoting the alphabets of the source, the destination and the relay, respectively, and $p(y,z|x)$ is the channel transition probability; the relay $Z$ can communicate to the destination $Y$ via an error-free digital link of rate $R_0$.

For this channel, a code of rate $R$ for $n$ channel uses, denoted by

$$(\mathcal{C}_{(n,R)}, f_n(z^n), g_n(y^n, f_n(z^n))), \text{ or simply, } (\mathcal{C}_{(n,R)}, f_n, g_n),$$

consists of the following:

1) A codebook at the source $X$,
$$\mathcal{C}_{(n,R)} = \{x^n(m) \in \Omega_X^n, m \in \{1, 2, \ldots, 2^{nR}\}\};$$
2) An encoding function at the relay $Z$,
$$f_n : \Omega_Z^n \to \{1, 2, \ldots, 2^{nR_0}\};$$
3) A decoding function at the destination $Y$,
$$g_n : \Omega_Y^n \times \{1, 2, \ldots, 2^{nR_0}\} \to \{1, 2, \ldots, 2^{nR}\}.$$

The average probability of error of the code is defined as

$$P_e^{(n)} = \Pr(g_n(Y^n, f_n(Z^n)) \neq M),$$

where the message $M$ is assumed to be uniformly drawn from the message set $\{1, 2, \ldots, 2^{nR}\}$. A rate $R$ is said to be achievable if there exists a sequence of codes

$$\{(\mathcal{C}_{(n,R)}, f_n, g_n)\}_{n=1}^{\infty}$$

such that the average probability of error $P_e^{(n)} \to 0$ as $n \to \infty$.

The capacity of the primitive relay channel is the supremum of all achievable rates. The well-known cut-set bound on the capacity of the primitive relay channel is stated as the following.

*Proposition 2.1 (Cut-set Bound):* For the general primitive relay channel, if a rate $R$ is achievable, then there exists some $p(x)$ such that

$$\begin{cases} R \leq I(X; Y, Z) & (1) \\ R \leq I(X; Y) + R_0. & (2) \end{cases}$$

Inequalities (1) and (2) are generally known as the broadcast bound and multiple-access bound, since they correspond to the broadcast channel $X$-$YZ$ and multiple-access channel $XZ$-$Y$, respectively.

### A. Symmetric Primitive Relay Channel

To simplify the exposition, in this paper, we only concentrate on the symmetric case of the primitive relay channel, that is, when $Y$ and $Z$ are conditionally independent and identically distributed given $X$, however our results can be extended to the asymmetric case by using channel simulation arguments. Formally, a primitive relay channel is said to be symmetric if

1) $p(y,z|x) = p(y|x)p(z|x)$,
2) $\Omega_Y = \Omega_Z := \Omega$, and $\Pr(Y = \omega|X = x) = \Pr(Z = \omega|X = x)$ for any $\omega \in \Omega$ and $x \in \Omega_X$.

In this case, we also use $p(\omega|x)$ to denote the transition probability of both the $X$-$Y$ and $X$-$Z$ channels.

## III. Main Results

This section presents two new upper bounds on the capacity of symmetric primitive relay channels that are generally tighter than the cut-set bound. Before stating our main theorems, in the following section we first explain the relation of our new bounds to the cut-set bound.

### A. Improving on the Cut-Set Bound

Let the relay's transmission be denoted by $I_n = f_n(Z^n)$. Let us recall the derivation of the cut-set bound. The first step in deriving (1)–(2) is to use Fano's inequality to conclude that

$$nR \leq I(X^n; Y^n, I_n) + n\epsilon.$$

We can then either proceed as

$$\begin{aligned} nR &\leq I(X^n; Y^n, I_n) + n\epsilon \\ &\leq I(X^n; Y^n, Z^n) + n\epsilon \\ &\leq nI(X; Y, Z) + n\epsilon \end{aligned}$$

to obtain the broadcast bound (1), where the second inequality follows from the data processing inequality and the single letterization in the third line can be either done with a time-sharing or fixed composition code argument[1]; or we can

[1]Note that the time-sharing or the fixed composition code argument for single letterization is needed to preserve the coupling to the second inequality in (4) via $X$.

proceed as

$$\begin{aligned} nR &\le I(X^n;Y^n,I_n)+n\epsilon \\ &\le I(X^n;Y^n)+H(I_n|Y^n)-H(I_n|X^n)+n\epsilon \quad (3)\\ &\le nI(X;Y)+nR_0+n\epsilon \quad (4)\end{aligned}$$

to obtain the multiple-access bound (2), where to obtain the last inequality we upper bound $H(I_n|Y^n)$ by $nR_0$ an use the fact that $H(I_n|X^n)$ is non-negative.

Instead of simply lower bounding $H(I_n|X^n)$ by 0 in the last step, our bounds presented in the next two subsections are based on letting $H(I_n|X^n) = na_n$, and prove a third inequality that forces $a_n$ to be strictly positive. Intuitively, it is easy to see that $a_n$ cannot be arbitrarily small. Specifically, suppose $a_n \approx 0$, then roughly speaking, this implies that given the transmitted codeword $X^n$, there is no ambiguity about $I^n$, or equivalently, all the $Z^n$ sequences jointly typical with $X^n$ are mapped to the same $I_n$. Since $Y^n$ and $Z^n$ are statistically equivalent given $X^n$ (they share the same typical set given $X^n$) this would further imply that $I_n$ can be determined based on $Y^n$ and therefore the transmitted codeword can be decoded based solely on $Y^n$, which forces the rate to be even smaller than $I(X;Y)$. In general, there is a trade-off between how close the rate can get to the multiple-access bound $I(X;Y)+R_0$ and how much it can exceed the point-to-point capacity $I(X;Y)$ of the $X$-$Y$ link.

In our new bounds, we capture this trade-off by leaving $a_n$ as it is in (3), yielding

$$R \le I(X;Y)+R_0-a_n+\epsilon$$

and proving a new constraint on the rate involving $a_n$. This new constraint is obtained by writing

$$\begin{aligned} nR &\le I(X^n;Y^n,I_n)+n\epsilon \\ &= H(Y^n,I_n)-H(Y^n|X^n)-H(I_n|X^n)+n\epsilon, \quad (5)\end{aligned}$$

and upper bounding $H(Y^n,I_n)$ in terms of $a_n$. We do this in two different ways corresponding to the two different ways of expanding $H(Y^n,I_n)$, i.e.

$$\begin{aligned} H(Y^n,I_n) &= H(Y^n)+H(I_n|Y^n) \\ &= H(I^n)+H(Y_n|I^n).\end{aligned}$$

Our first bound attacks the first conditional entropy term and is based on proving that

$$H(I_n|Y^n) \le H\left(\sqrt{\frac{a_n\ln 2}{2}}\right)+\sqrt{\frac{a_n\ln 2}{2}}\log(|\Omega|-1), \quad (6)$$

while our second bound attacks the second conditional entropy term and is based on proving that

$$\begin{aligned} &H(Y_n|I^n) \\ &\le H(X^n|I_n)-H(X^n|Z^n)+n(H(Z|X)+\Delta(p(x),a_n)),\end{aligned}$$

where $\Delta(p(x),a_n)$ is a quantity that depends on the input distribution $p(x)$ and $a_n$, which we formally define in Section III-C. Once these entropy relations are proved, it is not difficult to plug them in (5) and see how they lead to the theorems stated in the next two sections. The heart of our argument is therefore to prove these two entropy inequalities. To accomplish this, we suggest a new set of proof techniques. In particular, we look at the $B$-letter i.i.d. extensions of the random variables $X^n, Y^n$ and $I_n$ and study the geometric relations between their typical sets by using the generalized blowing-up lemma. While we use this same general approach for bounding the two entropy terms, we build on different arguments in each case, which eventually leads to two different bounds on the capacity of the relay channel that do not include each other in general.

### B. Bounding $H(I_n|Y^n)$

Our first bound builds on bounding $H(I_n|Y^n)$ and it is given by the following theorem that will be proved in Section IV. This bound is new and in particular strictly tighter than our first bound in [14].

*Theorem 3.1:* For the symmetric primitive relay channel, if a rate $R$ is achievable, then there exists some $p(x)$ and

$$a \in \left[0, \min\left\{R_0, H(Z|X), \frac{2}{\ln 2}\left(\frac{|\Omega|-1}{|\Omega|}\right)^2\right\}\right] \quad (7)$$

such that

$$\begin{cases} R \le I(X;Y,Z) & (8)\\ R \le I(X;Y)+R_0-a & (9)\\ R \le I(X;Y)+H\left(\sqrt{\frac{a\ln 2}{2}}\right) \\ \qquad\qquad +\sqrt{\frac{a\ln 2}{2}}\log(|\Omega|-1)-a. & (10)\end{cases}$$

Clearly our bound in Theorem 3.1 implies the cut-set bound in Proposition 2.1. In fact, it can be checked that our bound is *strictly* tighter than the cut-set bound for any $R_0 > 0$. For this, note that (9) will reduce to (2) only if $a = 0$; however, if $a = 0$ then (10) will constrain $R$ by the rate $I(X;Y)$ which is lower than the cut-set bound.

### C. Bounding $H(Y^n|I_n)$

Before presenting our second upper bound, we first define a parameter $\Delta(p(x),a)$ that will be used in stating the theorem. This bound is equivalent our second bound in [14], however we provide an alternative definition for $\Delta(p(x),a)$ in terms of information-theoretic quantitites.

*Definition 3.1:* Given a fixed channel transition probability $p(\omega|x)$, for any $p(x)$ and $a \ge 0$, $\Delta(p(x),a)$ is defined as

$$\begin{aligned} \Delta(p(x),a) := \max_{\tilde{p}(\omega|x)} & H(\tilde{p}(\omega|x)|p(x))+D(\tilde{p}(\omega|x)||p(\omega|x)|p(x)) \\ & -H(p(\omega|x)|p(x)) \quad (11)\end{aligned}$$

$$\text{s.t.} \quad \sum_{(x,\omega)}|p(x)\tilde{p}(\omega|x)-p(x)p(\omega|x)| \le 2\sqrt{\frac{a\ln 2}{2}}. \quad (12)$$

In the above, we adopt the notation in [16]. Specifically, $D(\tilde{p}(\omega|x)||p(\omega|x)|p(x))$ is the conditional relative entropy

defined as

$$D(\tilde{p}(\omega|x)||p(\omega|x)|p(x)) := \sum_{(x,\omega)} p(x)\tilde{p}(\omega|x)\log\frac{\tilde{p}(\omega|x)}{p(\omega|x)},$$

$H(\tilde{p}(\omega|x)|p(x))$ is the conditional entropy defined with respect to the joint distribution $p(x)\tilde{p}(\omega|x)$, i.e.,

$$H(\tilde{p}(\omega|x)|p(x)) := -\sum_{(x,\omega)} p(x)\tilde{p}(\omega|x)\log\tilde{p}(\omega|x),$$

and $H(p(\omega|x)|p(x))$ is the conditional entropy similarly defined with respect to $p(x)p(\omega|x)$.

It can be easily seen that $\Delta(p(x),a) \geq 0$ for all $p(x)$ and $a \geq 0$, and $\Delta(p(x),a) = 0$ when $a = 0$. Moreover, for any fixed $p(x)$ and $a > 0$, $\Delta(p(x),a) = \infty$ if and only if there exists some $x$ with $p(x) > 0$, and some $\omega$ and $\tilde{\omega}$ such that $p(\omega|x) = 0$ and $p(\tilde{\omega}|x) > 0$. Thus, a sufficient condition for $\Delta(p(x),a) < \infty$ for all $p(x)$ and $a > 0$ is that the channel transition matrix is *fully connected*, i.e., $p(\omega|x) > 0, \forall(x,\omega) \in \Omega_X \times \Omega$. In this case, $\Delta(p(x),a) \to 0$ as $a \to 0$ for any $p(x)$.

We are now ready to state our second new upper bound, which is proved by bounding $H(Y^n|I_n)$.

*Theorem 3.2:* For the symmetric primitive relay channel, if a rate $R$ is achievable, then there exists some $p(x)$ and $a \in [0, \min\{R_0, H(Z|X)\}]$ such that

$$\begin{cases} R \leq I(X;Y,Z) & (13)\\ R \leq I(X;Y) + R_0 - a & (14)\\ R \leq I(X;Y) + \Delta(p(x),a) & (15)\end{cases}$$

Theorem 3.2 also implies the cut-set bound in Propositions 2.1. In particular, when the channels $X$-$Y$ and $X$-$Z$ have a fully connected transition matrix, our new bound is *strictly* tighter than the cut-set bound since $\Delta(p(x),a) \to 0$ as $a \to 0$ for any $p(x)$ in this case.

In the remaining space we provide the proof of Theorem 3.1.

## IV. Proof of Theorem 3.1

Based on the discussion in Section III-A, to show Theorem 3.1, it suffices to prove the entropy inequality (6) between various $n$-letter random variables. For this, we go to the higher dimensional, say $nB$ dimensional space, to invoke the concepts of typical sets, and resort to a result on measure concentration, namely, the generalized blowing-up lemma.

Specifically, consider the $B$-length i.i.d. extensions of the random variables $X^n, Y^n, Z^n$ and $I_n$, i.e.,

$$\{(X^n(b), Y^n(b), Z^n(b), I_n(b))\}_{b=1}^{B}, \quad (16)$$

where for any $b \in [1:B]$, $(X^n(b), Y^n(b), Z^n(b), I_n(b))$ has the same distribution as $(X^n, Y^n, Z^n, I_n)$. For notational convenience, in the sequel we write the $B$-length vector $[X^n(1), X^n(2), \ldots, X^n(B)]$ as $\mathbf{X}$ and similarly define $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{I}$; note here we have $\mathbf{I} = [f_n(Z^n(1)), f_n(Z^n(2)), \ldots, f_n(Z^n(B))] =: f(\mathbf{Z})$.

The following lemma is critical for establishing inequality (6). Its own proof is given at the end of the paper.

*Lemma 4.1:* Let $f^{-1}(\mathbf{i}) := \{\underline{\omega} \in \Omega^{nB} : f(\underline{\omega}) = \mathbf{i}\}$ and $\Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+\delta)}(f^{-1}(\mathbf{i}))$ be its blown-up set defined as

$$\Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+\delta)}(f^{-1}(\mathbf{i})) := \Bigg\{ \underline{\omega} \in \Omega^{nB} : \exists\ \underline{\omega}' \in f^{-1}(\mathbf{i}) \text{ s.t. } d(\underline{\omega}, \underline{\omega}') \leq nB\left(\sqrt{\frac{a_n \ln 2}{2}} + \delta\right) \Bigg\}$$

where $d(\underline{\omega}, \underline{\omega}')$ denotes the Hamming distance between $\underline{\omega}$ and $\underline{\omega}'$. Then for any $\delta > 0$ and $B$ sufficiently large,

$$\Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+\delta)}(f^{-1}(\mathbf{I}))) \geq 1 - \delta.$$

With the above lemma, we now upper bound $H(\mathbf{I}|\mathbf{Y})$. Let

$$E = \mathbb{I}(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+\delta)}(f^{-1}(\mathbf{I})))$$

where $\mathbb{I}(\cdot)$ is the indicator function defined as

$$\mathbb{I}(A) = \begin{cases} 1 & \text{if } A \text{ holds} \\ 0 & \text{otherwise.}\end{cases}$$

We have

$$\begin{aligned} &H(\mathbf{I}|\mathbf{Y}) \leq H(\mathbf{I}, E|\mathbf{Y})\\ &= H(E|\mathbf{Y}) + H(\mathbf{I}|\mathbf{Y}, E)\\ &\leq H(\mathbf{I}|\mathbf{Y}, E) + 1\\ &= \Pr(E=1)H(\mathbf{I}|\mathbf{Y}, E=1) + \Pr(E=0)H(\mathbf{I}|\mathbf{Y}, E=0) + 1\\ &\leq H(\mathbf{I}|\mathbf{Y}, E=1) + \delta nBR_0 + 1. \quad (17)\end{aligned}$$

To bound $H(\mathbf{I}|\mathbf{Y}, E=1)$, consider a Hamming ball[2] centered at $\mathbf{Y}$ of radius $nB\left(\sqrt{\frac{a_n \ln 2}{2}} + \delta\right)$. The condition $E = 1$, i.e., $\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+\delta)}(f^{-1}(\mathbf{I}))$, ensures that there is at least one point $\underline{\omega} \in f^{-1}(\mathbf{I})$ belonging to this ball, and therefore, given $E = 1$ and $\mathbf{Y}$ there are at most $\left|\text{Ball}\left(nB\left(\sqrt{\frac{a_n \ln 2}{2}} + \delta\right)\right)\right|$ possibilities of $\mathbf{I}$, leading to the following upper bound on $H(\mathbf{I}|\mathbf{Y}, E=1)$,

$$\begin{aligned} H(\mathbf{I}|\mathbf{Y}, E=1) &\leq \log\left|\text{Ball}\left(nB\left(\sqrt{\frac{a_n \ln 2}{2}} + \delta\right)\right)\right|\\ &\leq nB\left[H\left(\sqrt{\frac{a_n \ln 2}{2}}\right) + \sqrt{\frac{a_n \ln 2}{2}}\log(|\Omega|-1) + \delta_1\right] \quad (18)\end{aligned}$$

for some $\delta_1 \to 0$ as $\delta \to 0$, where (18) follows from the characterization of the volume of a Hamming ball. Plugging (18) into (17), we have

$$H(\mathbf{I}|\mathbf{Y}) \leq nB\left[H\left(\sqrt{\frac{a_n \ln 2}{2}}\right) + \sqrt{\frac{a_n \ln 2}{2}}\log(|\Omega|-1) + \delta_1\right] + \delta nBR_0 + 1.$$

[2] A Hamming ball centered at $\mathbf{c}$ of radius $r$, denoted by $\text{Ball}(\mathbf{c}, r)$, is defined as the set of points that are within Hamming distance $r$ of $\mathbf{c}$. The center $\mathbf{c}$ can be omitted in the notation when it becomes irrelevant.

Dividing $B$ at both sides of the above inequality and noting that

$$H(\mathbf{I}|\mathbf{Y}) = \sum_{b=1}^{B} H(I_n(b)|Y^n(b)) = BH(I_n|Y^n),$$

we have

$$H(I_n|Y^n) \le n\left[H\left(\sqrt{\frac{a_n \ln 2}{2}}\right) + \sqrt{\frac{a_n \ln 2}{2}}\log(|\Omega|-1) + \delta_1 + \delta R_0 + \frac{1}{nB}\right]. \quad (19)$$

Since $\delta, \delta_1$ and $\frac{1}{nB}$ in (19) can all be made arbitrarily small by choosing $B$ sufficiently large, we obtain (6).

We next prove Lemma 4.1.

*Proof of Lemma 4.1:* Consider any $(\mathbf{x}, \mathbf{i}) \in \mathcal{T}_\epsilon^{(B)}(X^n, I_n)$, where $\mathcal{T}_\epsilon^{(B)}(X^n, I_n)$ denotes the $\epsilon$-jointly typical sets[3] with respect to $(X^n, I_n)$. From [17, Sec. 2.5], we have for some $\epsilon_1 \to 0$ as $\epsilon \to 0$,

$$p(\mathbf{i}|\mathbf{x}) \ge 2^{-B(H(I_n|X^n)+\epsilon_1)} \ge 2^{-nB(a_n+\epsilon_1)},$$

i.e.,

$$\Pr(\mathbf{Z} \in f^{-1}(\mathbf{i})|\mathbf{x}) \ge 2^{-nB(a_n+\epsilon_1)}.$$

We now apply the generalized blowing-up lemma as stated in the following.

*Lemma 4.2 (Generalized Blowing-Up Lemma):* Let $U_1, U_2, \ldots, U_n$ be $n$ independent random variables taking values in a finite set $\mathcal{U}$. Then, for any $A \subseteq \mathcal{U}^n$ with $\Pr(U^n \in A) \ge 2^{-na_n}$,

$$\Pr(U^n \in \Gamma_{n(\sqrt{\frac{a_n \ln 2}{2}}+r)}(A)) \ge 1 - e^{-2nr^2}, \forall r > 0.$$

With Lemma 4.2, we have

$$\begin{aligned}
&\Pr(\mathbf{Z} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x}) \\
&= \Pr(\mathbf{Z} \in \Gamma_{nB(\sqrt{\frac{(a_n+\epsilon_1)\ln 2}{2}}+[\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1}-\sqrt{\frac{(a_n+\epsilon_1)\ln 2}{2}}])}(f^{-1}(\mathbf{i}))|\mathbf{x}) \\
&\ge \Pr(\mathbf{Z} \in \Gamma_{nB(\sqrt{\frac{(a_n+\epsilon_1)\ln 2}{2}}+[\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1}-\sqrt{\frac{a_n \ln 2}{2}}-\sqrt{\frac{\epsilon_1 \ln 2}{2}}])}(f^{-1}(\mathbf{i}))|\mathbf{x}) \\
&\ge \Pr(\mathbf{Z} \in \Gamma_{nB(\sqrt{\frac{(a_n+\epsilon_1)\ln 2}{2}}+\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x}) \\
&\ge 1 - e^{-2nB\epsilon_1} \\
&\ge 1 - \sqrt{\epsilon_1}
\end{aligned}$$

for sufficiently large $B$. Noting that $\mathbf{Y}$ and $\mathbf{Z}$ are identically distributed given $\mathbf{X}$, we obtain

$$\Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x}) \ge 1 - \sqrt{\epsilon_1},$$

and thus,

$$\begin{aligned}
&\Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{I}))) \\
&= \sum_{(\mathbf{x},\mathbf{i})} \Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x},\mathbf{i})p(\mathbf{x},\mathbf{i}) \\
&= \sum_{(\mathbf{x},\mathbf{i})} \Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x})p(\mathbf{x},\mathbf{i}) \quad (20) \\
&\ge \sum_{(\mathbf{x},\mathbf{i})\in\mathcal{T}_\epsilon^{(B)}(X^n,I_n)} \Pr(\mathbf{Y} \in \Gamma_{nB(\sqrt{\frac{a_n \ln 2}{2}}+2\sqrt{\epsilon_1})}(f^{-1}(\mathbf{i}))|\mathbf{x})p(\mathbf{x},\mathbf{i}) \\
&\ge (1-\sqrt{\epsilon_1}) \sum_{(\mathbf{x},\mathbf{i})\in\mathcal{T}_\epsilon^{(B)}(X^n,I_n)} p(\mathbf{x},\mathbf{i}) \\
&\ge (1-\sqrt{\epsilon_1})^2 \\
&\ge 1 - 2\sqrt{\epsilon_1}
\end{aligned}$$

for sufficiently large $B$, where (20) follows due to the Markov chain: $\mathbf{Y} \leftrightarrow \mathbf{X} \leftrightarrow \mathbf{Z} \leftrightarrow \mathbf{I}$. Finally, choosing $\delta$ to be $2\sqrt{\epsilon_1}$ concludes the proof of Lemma 4.1. ■

[3]This paper adopts the same definitions and notations for typical sets as those in [17].

## REFERENCES


[1] E. C. van der Meulen, "Three-terminal communication channels," *Adv. Appl. Prob.*, vol. 3, pp. 120–154, 1971.

[2] T. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inform. Theory*, vol. 25, pp. 572–584, 1979.

[3] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3037–3063, September 2005.

[4] X. Wu and L.-L. Xie, " A unified relay framework with both D-F and C-F relay nodes," *IEEE Trans. Inform. Theory*, vol. 60, no. 1, pp. 586–604, January 2014.

[5] B. Schein and R. Gallager, "The Gaussian parallel relay network," in *Proc. of IEEE International Symposium on Informaiton Theory*, pp. 22, June 2000.

[6] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, 2011.

[7] S. H. Lim, Y.-H. Kim, A. El Gamal, S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inform. Theory*, vol. 57, no. 5, pp. 3132–3152, May 2011.

[8] S. Zahedi, "On reliable communication over relay channels," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2005.

[9] A. El Gamal and M. Aref, "The capacity of the semideterministic relay channel," *IEEE Trans. Inform. Theory*, vol. 28, no. 3, pp. 536, May 1982.

[10] Y.-H. Kim, "Capacity of a class of deterministic relay channels," *IEEE Trans. Inform. Theory*, vol. 54, no. 3, pp.1328–1329, Mar. 2008.

[11] Z. Zhang, "Partial converse for a relay channel," *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 1106–1110, Sept. 1988.

[12] M. Aleksic, P. Razaghi, and W. Yu, "Capacity of a class of modulo-sum relay channels," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 921–930, 2009.

[13] F. Xue, "A new upper bound on the capacity of a primitive relay channel based on channel simulation," *IEEE Trans. Inform. Theory*, vol. 60, pp. 4786–4798, Aug. 2014.

[14] X. Wu, L.-L. Xie, and A. Ozgur, "Upper bounds on the capacity of symmetric primitive relay channels," in *Proc. IEEE Int. Symposium on Information Theory*, Hong Kong, June 14–19, 2015.

[15] M. Raginsky, I. Sason, Concentration of Measure Inequalities in Information Theory, Communications and Coding Oct. 2013 [Online]. Available: http://arxiv.org/abs/1212.4663

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

[17] A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge, U.K.: Cambridge University Press, 2012.

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# The Capacity of the State-Dependent Semideterministic Relay Channel

Ritesh Kolte, Ayfer Özgür
Department of Electrical Engineering
Stanford University, CA 94305, USA
Email: {rkolte,aozgur}@stanford.edu

Haim Permuter
Dept. of Electrical and Computer Engineering
Ben-Gurion University, Beer-Sheva 84105, Israel
Email: haimp@bgu.ac.il

***Abstract*—The capacity region of the semideterministic relay channel has been characterized using the idea of partial-decode-forward. However, the requirement to explicitly decode part of the message at the relay can be restrictive, for example, when nodes have different side information regarding the state of the channel. In this paper, we generalize this scheme to *cooperative-bin-forward* by building on the observation that explicit recovering of part of the message is not needed to induce cooperation. Instead, the relay can bin its received signal and the bin index is cooperatively forwarded to the decoder. The main advantage of this new scheme is illustrated by considering a state-dependent extension. While partial-decode-forward is suboptimal in the new setup, cooperative-bin-forward continues to achieve capacity.**

***Index Terms*—Cooperative-bin-forward, Relay channel, State, Semideterministic**

## I. Introduction

The capacity region of the semideterministic relay channel, depicted in Figure 1, is characterized in [1] using the partial-decode-forward scheme. In this scheme, the source splits its message into two parts and encodes them using superposition coding. The relay decodes one part of the message, and maps this to a codeword to be transmitted in the next block. The codebooks at the source are generated conditioned on the relay's transmission, which results in coherent transmissions from the source and the relay.

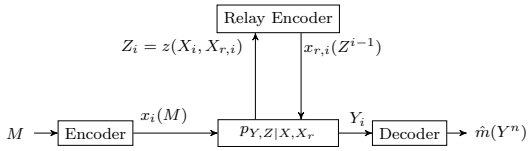



Fig. 1: Semideterministic Relay Channel

Consider now the extension of this model depicted in Figure 2, which corresponds to a state-dependent semideterministic relay channel where the state information is causally available only at the source and the destination. This model captures the natural cellular downlink scenario, in which training enables the source and the destination to learn the channel gain between them (state = channel gain), while a relay could be potentially available to assist the communication, e.g. a wifi access point. In this scenario, it is typically unrealistic to assume that the relay is also able to obtain timely information about the channel state between the source and the destination. As such, requiring the relay to decode part of the source message, without any state information, is unduly restrictive and to the best of our knowledge, the capacity has not been characterized previously.

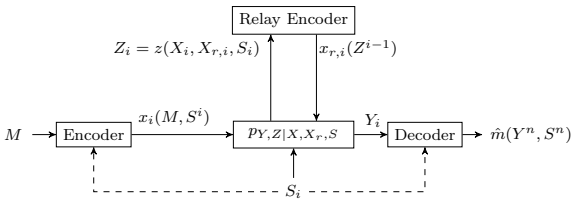



Fig. 2: State-dependent Semideterministic Relay Channel with Causal State Information at Source and Destination

The main contribution of this paper is to develop a new scheme which we call *cooperative-bin-forward*. This new scheme does not require the relay to decode part of the message; instead, the relay simply bins its received signal and maps the bin-index to a codeword to be transmitted in the next block. As in partial-decode-forward, the codebooks at the source are generated conditioned on the relay's transmission, resulting in coherent cooperation. This cooperative aspect of the scheme distinguishes it from bin-forward (a.k.a. hash-forward) that has been considered previously for primitive relay channels in [2]. For the vanilla semideterministic relay channel in Figure 1, cooperative-bin-forward recovers the capacity achieved by partial-decode-forward. However, while partial-decode-forward is suboptimal for the state-dependent semideterministic relay channel in Figure 2, cooperative-bin-forward continues to achieves the capacity.

Another setup we consider, motivated by the relay-without-delay channel considered in [3], is the "without-delay" variation of the state-dependent setup described above, depicted in Figure 3. In this setup, the transmission of the relay is allowed to depend on its past and current received signal. The capacity region for this setup without state is characterized in [3], using partial-decode-forward combined with instantaneous relaying (a.k.a. codetrees or Shannon strategies). We show that cooperative-bin-forward combined with instantaneous relaying achieves the capacity regions of this setup too, while partial-decode-forward suffers from the same shortcoming mentioned in the previous paragraphs.

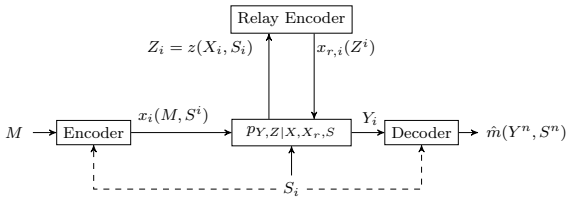



Fig. 3: State-dependent Semideterministic Relay-Without-Delay Channel with Causal State Information at Source and Destination. Since the current transmission of the relay is allowed to depend on its current received signal, we need to specify the model so that the latter does not depend on the former, in contrast to Figure 2.

*Related Work*

Various cases of state-dependent relay channels have been considered in [4]–[10]. The achievability schemes in these works combine well-known block-Markov relaying ideas such as partial-decode-forward and compress-forward with Shannon strategies or (Gelfand-Pinsker) multicoding. A class of state-dependent orthogonal relay channels with state information only at decoder was considered in [11], and optimality of a partial-decode-compress-forward scheme was proved. To the best of our knowledge, the state-dependent relay channel considered in this paper has not been previously studied, and as mentioned previously, standard combinations of available ideas are not sufficient to obtain good achievability schemes.

*Organization*

The following section describes the models and notation. Section III contains the formal statements of the main results described in the introduction. A toy example is considered in Section IV for the purpose of explicitly illustrating the advantage of cooperative-bin-forward over partial-decode-forward. The following sections contain the proof of the main result. We conclude by describing some open problems in section VI.

## II. System Models

As standard, capital letters denote random variables, small letters denote realizations, and calligraphic letters denote the alphabet of the corresponding random variable. The notation $\mathcal{T}_\epsilon^{(n)}$ stands for the $\epsilon$-strongly typical set of sequences for the random variables in context.

### A. State-Dependent Semideterministic Relay Channels

The state-dependent semideterministic relay channel is depicted in Figure 2, and described by the pmf $p_S(s)p_{Y|X,X_r,S}(y|x, x_r, s)$ and $Z = z(X, X_r, S)$. The encoder and decoder have causal state information. So a $(n, 2^{nR}, \epsilon)$ code for the above channel consists of the source encoding, relay encoding and decoding functions:

$$\begin{aligned} x_i &: [1 : 2^{nR}] \times \mathcal{S}^i \to \mathcal{X}, \quad 1 \le i \le n, \\ x_{r,i} &: \mathcal{Z}^{i-1} \to \mathcal{X}_r, \quad 1 \le i \le n, \\ \hat{m} &: \mathcal{Y}^n \times \mathcal{S}^n \to [1 : 2^{nR}], \end{aligned}$$

such that

$$\Pr\{\hat{m}(Y^n, S^n) \neq M\} \le \epsilon,$$

where $M \in [1 : 2^{nR}]$ denotes the transmitted message. A rate $R$ is said to be *achievable* if for every $\epsilon > 0$, there exists a $(n, 2^{nR}, \epsilon)$ code for sufficiently large $n$. The capacity is defined to be the supremum of achievable rates.

The state-dependent semideterministic relay-without-delay channel is depicted in Figure 3, and described by the pmf $p_S(s)p_{Y|X,X_r,S}(y|x, x_r, s)$ and $Z = z(X, S)$. The difference from the previous setup is that the relay encoding function is now allowed to depend also on $Z_i$:

$$x_{r,i} : \mathcal{Z}^i \to \mathcal{X}_r, \quad 1 \le i \le n.$$

Note that here we need to restrict $Z$ to be $z(X, S)$, instead of $z(X, X_r, S)$.

## III. Main Results

The first result provides an expression for the capacity of the state-dependent semideterministic relay channel.

**Theorem 1.** The capacity of the state-dependent semideterministic relay channel, shown in Figure 2, is given by

$$\min\{I(X, X_r; Y|S),\ H(Z|S, X_r) + I(X; Y|S, X_r, Z)\}, \quad (1)$$

maximized over distributions that can be factorized as $p_{X_r}(x_r)p_{X|X_r,S}(x|x_r, s)$.

The proof of the achievability part is presented in the next section. We refer the reader to the longer version of this paper [12] for all the missing proofs (including converses).

The difference between the capacity expression in Theorem 1 and the capacity of the semideterministic relay channel [13, Eq. (16.8)] is that the mutual information and entropy terms involve a conditioning on $S$. Such an expression would also characterize the capacity if the relay is provided with the state information, and it would be achievable by performing partial-decode-forward while treating the state as a time-sharing sequence. It is quite interesting then that the capacity expression remains the same even when the relay does not have state information. However, the limitation is reflected in the fact that the choice of pmf is restricted to be $p_{X_r}(x_r)p_{X|X_r,S}(x|x_r, s)$, instead of $p_{X,X_r|S}(x, x_r|s)$. So, the cost of not having state information at the relay is reflected entirely in the limited choice of pmf.

The following theorem states the capacity of the without-delay variation of the above case. The expression involves an auxiliary random variable, which allows the relay to perform instantaneous relaying on top of the binning.

**Theorem 2.** The capacity of the state-dependent semideterministic relay-without-delay channel, shown in Figure 3, is given by

$$\min\{I(U, X; Y|S),\ H(Z|U, S) + I(X; Y|U, Z, S)\}, \quad (2)$$

maximized over distributions of the form $p_U(u)p_{X|U,S}(x|u, s)$ and $X_r = x_r(U, Z)$, and $|\mathcal{U}| \le |\mathcal{S}|\,(|\mathcal{X}||\mathcal{X}_r| - 1) + 2$.

The capacity region for the setup of Theorem 2 in the absence of states is characterized in [3, Proposition 7]. Setting $S$ to be the empty random variable in Theorem 2 recovers this

result. Note that the objective in (2) is the same as that in (1) with $X_r$ being replaced by $U$. However, the optimization in (2) is over a different domain since the dependence of $X_r$ on $Z$ can now be chosen and is not specified by the channel.

## IV. Illustrative Example

Consider the following special case of Figure 2. Let the state $S$ be the ternary random variable

$$p_S(s) = \begin{cases} p/2, & \text{if } s = 0, \\ p/2, & \text{if } s = 1, \\ 1-p, & \text{if } s = 2, \end{cases}$$

where $p < 1/2$. The other variables are all binary. The channel $z(X, S)$ is the memory-with-stuck-at-faults channel considered in [13, Figure 7.7], while the channel $p_{Y|X,X_r,S}$ is specialized to be a noiseless channel from $X_r$ to $Y$. Formally,

$$z(X, S) = \begin{cases} 0, & \text{if } S = 0, \\ 1, & \text{if } S = 1, \\ X, & \text{if } S = 2, \end{cases}$$
$$Y = X_r.$$

Recall that the source and the destination know the state information causally while the relay has no state information.

If, motivated by the optimality of decode-forward in the case of a line network with no state, the relay is required to decode the message, then the achievable rate is limited to be no more than the capacity of the memory-with-stuck-at-faults channel when the state is known causally *only* to the source, which is $1 - H_2\left(\frac{p}{2}\right)$. We point out that this cannot be improved by using partial-decode-forward, because the absence of a direct link between the source and destination means that any part of the message that is not forwarded by the relay cannot be communicated to the destination in any manner. However, a higher rate can be achieved if the relay simply forwards its received signal, resulting in an effective channel between the source and the destination that is the memory with stuck-at faults channel with state known causally *both* to the source and the destination. The capacity of this channel is $1 - p$, which is achieved by multiplexing at the source and demultiplexing at the destination according to the observed state. Thus, a rate $1 - p$, which is higher than $1 - H_2\left(\frac{p}{2}\right)$, can be achieved.

What if the channel from the relay to destination is not a noiseless bit-pipe, but a general noisy channel with capacity at least $1-p$? The rate $1-p$ can still be achieved if the operation at the relay is changed from simply forwarding to randomly binning its received signal into $\approx 2^{n(1-p)}$ bins and forwarding a codeword corresponding to the chosen bin. To recover the message, the destination can first decode the bin-index. Since the destination has state information, it can reconstruct the state-multiplexed codebook at the source. Hence, it can recover the message by finding the unique source codeword, if any, that results in the received signal at the relay falling in the correct bin.

The above example serves to illustrate the limitation of partial-decode-forward when nodes have different side-information. This example did not require cooperative transmissions from the source and the relay, because the source transmission did not directly affect the received signal at the destination. When there is also a direct link between the source and the destination, as allowed in the general models that we consider in this paper, the source and relay need to perform the bin-forward operation in a cooperative fashion.

## V. Proof of Theorem 1

Due to the availability of causal state information at the source encoder and the decoder, the source encoder constructs codebooks for each state symbol and treats the state sequence as a time-sharing sequence (i.e. it performs multiplexing). Note that since the relay does not have state information, it might not be able to decode part of the message. However, it can still perform the bin-forward operation, allowing us to establish coherence between the source and the relay transmissions without sacrificing unnecessarily on the rate.

*Proof:*

Fix a pmf $p_{X_r}(x_r)p_{X|X_r,S}(x|x_r, s)$ and $\epsilon > 0$. Split $R$ as $R' + R''$, with the message $M$ denoted accordingly as $(M', M'')$. Divide the total communication time into $B$ blocks, each of length $n$.

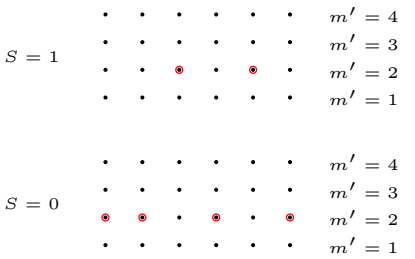



Fig. 4: The figure depicts the cribbed codewords generated for encoding $m'$ for a given $x_r^n(l)$. Each node corresponds to a $z$ symbol that is generated independently according to $p_{Z|X_r,S}(\cdot|x_{r,i}(l), s)$. The red circles show how encoder 1 chooses the codeword if it wants to transmit $m' = 2$ and observes $s^n = (0, 0, 1, 0, 1, 0)$. This construction is not identical but equivalent to that described in [13, Section 7.4.1].

*Codebook Generation:*

For each block $b \in [1 : B]$, a codebook is generated independently of the other blocks as follows.

- *Cooperation codewords*
  Generate $2^{n\widetilde{R}}$ codewords $x_{rb}^n(l_{b-1})$, i.i.d. according to $p_{X_r}$, where $l_{b-1} \in [1 : 2^{n\widetilde{R}}]$.
- *Cribbed codewords*
  For each $l_{b-1}$ and each $s \in \mathcal{S}$, generate a codebook of $2^{nR'}$ codewords. The $i$th symbol of such a codeword is chosen independently according to $p_{Z|X_r,S}(\cdot|x_{rbi}(l_{b-1}), s)$. The result of this is that for each $l_{b-1}$, each $m'_b \in [1 : 2^{nR'}]$ and each $s_b^n = (s_{b1}, s_{b2}, \ldots, s_{bn})$, the source encoder can form an effective codeword $z_b^n(m'_b|l_{b-1}, s_b^n)$, whose $i$th

symbol can be causally chosen as the $i$th symbol of the $m'_b$-th codeword from the codebook corresponding to $l_{b-1}$ and $s_{bi}$. See Figure 4.

- *Transmission codewords*
  For each $l_{b-1}$, each $m'_b \in [1 : 2^{nR'}]$ and each $s \in \mathcal{S}$, generate a codebook of $2^{nR''}$ codewords. The $i$th symbol of such a codeword is generated independently according to $p_{X|X_r,Z,S}(\cdot|x_{rbi}(l_{b-1}), z_{bi}(m'_b|l_{b-1}, s), s)$. The result of this construction is that for each $l_{b-1}$, each $m'_b \in [1 : 2^{nR'}]$, each $m''_b \in [1 : 2^{nR''}]$ and each $s^n_b$, the source encoder can form an effective codeword $x^n_b(m''_b|l_{b-1}, m'_b, s^n_b)$, whose $i$th symbol can be causally chosen as the $i$th symbol of the $m''_b$-th codeword from the codebook corresponding to $l_{b-1}$, $m'_b$ and $s_{bi}$.
- *Binning*
  Partition the set $\mathcal{Z}^n$ into $2^{n\widetilde{R}}$ bins, by choosing a bin for each $z^n$ independently and uniformly at random. Denote the index of the chosen bin for $z^n$ by $\mathsf{bin}_b(z^n)$.

*Encoding:*

Fix $l_0 = 1$ and $(m'_B, m''_B) = (1, 1)$. Since the message in the last block is fixed, the effective rate of communication is $\frac{B-1}{B}R$, which can be made as close as desired to $R$ by choosing a sufficiently large $B$.

In block $b$, $l_{b-1}$ is known to the source encoder. To communicate message $m_b = (m'_b, m''_b)$, it transmits $x^n_b(m''_b|l_{b-1}, m'_b, s^n_b)$. The relay transmits $x^n_{rb}(l_{b-1})$. Due to the deterministic link from source to relay and the codebook construction, the received signal at the relay in block $b$ is the codeword $z^n_b(m'_b|l_{b-1}, s^n_b)$. The source and the relay set $l_b$ to be the index of the bin containing $z^n_b(m'_b|l_{b-1}, s^n_b)$.

From the encoding operation described above, we can see that the label $l_b$ depends on $(l_{b-1}, m'_b, s^n_b)$. We do not require the relay to decode $m'_b$, but the source and the relay can still establish cooperation by directly performing a binning on the $z^n_b$ codeword to agree on the $u^n_{b+1}$ codeword to be used in the next block, thus providing the scheme with the title "cooperative-bin-forward". The term *cooperative* is added to emphasize that the source and the relay agree on the binning and transmit coherently. Thus, the scheme achieves cooperation by communicating $l_b$ via the relay, instead of $m'_b$. While the relay is not required to decode the partial message, we still need the destination to be able to decode all parts of the transmitted message successfully. In the following, appropriate conditions are imposed so that the destination can utilize the state information at its disposal to achieve successful decoding.

*Decoding:*

The decoder performs backward decoding, starting from block $B$ and moving towards block 1, performing the following two steps for each block $b$:

(1) Assuming that $l_b$ is known from previous operations, the decoder, for each $l_{b-1} \in [1 : 2^{nR}]$, finds the unique $m'_b$ such that

$$\mathsf{bin}_b(z^n_b(m'_b|l_{b-1}, s^n_b)) = l_b.$$

Whenever a unique $m'_b$ cannot be found for some $l_{b-1}$, the decoder chooses any $m'_b$ arbitrarily. So after this operation, the decoder has chosen one $m'_b$ for each $l_{b-1}$, given its knowledge of $l_b$ and $s^n_b$. We will signify this explicitly by denoting the chosen message as $\hat{m}'_b(l_{b-1}, s^n_b, l_b)$.

(2) Now it looks for the unique $(\hat{l}_{b-1}, \hat{m}''_b)$ such that

$$\Big(x^n_{rb}(\hat{l}_{b-1}),\, z^n_b(\hat{m}'_b(\hat{l}_{b-1}, s^n_b, l_b)|\hat{l}_{b-1}, s^n_b),\, x^n_b(\hat{m}''_b|\hat{l}_{b-1}, \hat{m}'_b(\hat{l}_{b-1}, s^n_b, l_b), s^n_b),\, s^n_b, y^n_b\Big) \in \mathcal{T}^{(n)}_\epsilon. \quad (3)$$

*Probability of Error:*

In the following error analysis, we will observe that in order to achieve the largest rate, the scheme will set $R' \approx H(Z|X_r, S)$. The causal multiplexing-demultiplexing strategy proposed above effectively creates a different codebook for $m'_b$ for each $s^n_b$ sequence. The total number of $z^n_b$ codewords constructed by the source encoder considering only the typical $s^n_b$ sequences is therefore $\approx 2^{nH(S)} \cdot 2^{nR'} \approx 2^{nH(S,Z|X_r)}$. However, these codewords cannot be distinct since there are only $\approx 2^{nH(Z|X_r)}$ distinct typical sequences $z^n_b$ (conditioned on $l_{b-1}$). This implies that multiple $(s^n_b, m'_b)$ pairs will be mapped to the same codeword $z^n_b$ and therefore, the relay will be not be able to decode $m'_b$ due to the lack of state information. In the absence of state, the scheme will set $\widetilde{R} \approx R' \approx H(Z|X_r)$. In this case, it is easy to see that given its knowledge of $l_{b-1}$, the relay can indeed recover $m'_b$, since each message $m'_b$ is mapped to a different bin. Thus, cooperatively communicating the bin index becomes equivalent to cooperatively communicating the partial message $m'_b$, so cooperative-bin-forward for setup without states is indeed equivalent to partial-decode-forward. When we have states, even though we still set $\widetilde{R} \approx R'$, the relay is not be able to decode any part of the message, so the binning aspect of the scheme is instrumental.

By symmetry, we can assume without loss of generality that the true messages and bin-indices corresponding to the current block are all 1, i.e.

$$(L_{b-1}, M'_b, M''_b) = (1, 1, 1).$$

We bound the probability of decoding error in block $b$ conditioned on successful decoding for blocks $\{B, B-1, \ldots, b+1\}$, averaged over the randomness in the messages and codebook generation. In particular, successful decoding in block $b+1$ means that $L_b$ has been decoded successfully, where we remind ourselves that

$$L_b = \mathsf{bin}_b(Z^n_b(1|1, S^n_b)).$$

An error occurs in block $b$ only if any of the following events occur:

$(a)$ $\hat{M}'_b(1, S^n_b, L_b) \neq 1$,
$(b)$ $(\hat{L}_{b-1}, \hat{M}''_b) \neq (1, 1)$ given $\hat{M}'_b(1, S^n_b, L_b) = 1$.

*Event* $(a)$*:* $\hat{M}'_b(1, S_b^n, L_b) \neq 1$*:* We have

$$\begin{aligned}
&\Pr\left(\hat{M}'_b(1, S_b^n, L_b) \neq 1\right) \\
&= \Pr\left(\mathsf{Bin}_b(Z_b^n(m'_b|1, S_b^n)) = L_b \text{ for some } m'_b > 1\right) \\
&\leq \sum_{m'_b>1} \Pr\left(Z_b^n(m'_b|1, S_b^n) = Z_b^n(1|1, S_b^n)\right) \\
&\quad + \sum_{m'_b>1} \Pr\left(\mathsf{Bin}_b(Z_b^n) = \mathsf{Bin}_b(\widetilde{Z}_b^n) \,\middle|\, Z_b^n \neq \widetilde{Z}_b^n\right) \\
&\leq 2^{nR'} \cdot 2^{-n(H(Z|X_r,S)-\delta(\epsilon))} + 2^{nR'} \cdot 2^{-n\widetilde{R}},
\end{aligned}$$

where we use $\delta(\epsilon)$ to denote any function of $\epsilon$ for which $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Hence, we get that

$$\Pr\left(\hat{M}'_b(1, S_b^n, L_b) \neq 1\right) \to 0 \text{ as } n \to \infty,$$

if the following two constraints are satisfied:

$$R' < \widetilde{R}, \tag{4}$$

$$R' < H(Z|X_r, S) - \delta(\epsilon). \tag{5}$$

*Event* $(b)$*:* $(\hat{L}_{b-1}, \hat{M}''_b) \neq (1, 1)$ *given* $\hat{M}'_b(1, S_b^n, L_b) = 1$*:* For brevity, we do not mention the conditioning on $\{\hat{M}'_b(1, S_b^n, L_b) = 1\}$ in the following expressions. The probability of this event is upper bounded by

$$\begin{aligned}
&\Pr\left(\text{Condition (3) not satisfied by } (l_{b-1}, m''_b) = (1, 1)\right) \\
&+ \Pr\left(\text{Condition (3) satisfied by some } (l_{b-1}, m''_b) \neq (1, 1)\right).
\end{aligned}$$

The first term goes to zero as $n \to \infty$ by the law of large numbers. The second term can be analyzed as follows:

$$\begin{aligned}
&\Pr\left(\text{Condition (3) satisfied by some } (l_{b-1}, m''_b) \neq (1, 1)\right) \\
&\leq \sum_{l_{b-1}=1, m''_b>1} \Pr\left(\text{Condition (3) satisfied by } (1, m''_b)\right) \\
&\quad + \sum_{l_{b-1}>1, m''_b\geq 1} \Pr\left(\text{Condition (3) satisfied by } (l_{b-1}, m''_b)\right) \\
&\leq 2^{nR''} 2^{-n(I(X;Y|X_r,Z,S)-\delta(\epsilon))} \\
&\quad + 2^{n(\widetilde{R}+R'')} 2^{-n(I(X,X_r,Z;Y|S)-\delta(\epsilon))},
\end{aligned}$$

which follows by applying the packing lemma. Note that when $l_{b-1} > 1$, it so happens due to the codebook construction that $y_b^n$ is independent of all the other sequences for any value of $(m'_b, m''_b)$. So the joint distribution of the sequences has the same factorization no matter what $m'_b$ is chosen for $l_{b-1} > 1$. The only fact that matters for our analysis is that at most one $m'_b$ has been chosen somehow for each $l_{b-1} > 1$. This allows us to write the fourth event as the union of at most $2^{n(\widetilde{R}+R'')}$ events, where each corresponds to a different value of $(l_{b-1}, m''_b)$. Thus, as $n \to \infty$, we get that

$$\Pr\left((\hat{L}_{b-1}, \hat{M}''_b) \neq (1, 1) \,\middle|\, \hat{M}'_b(1, S_b^n, L_b) = 1\right) \to 0,$$

if

$$R'' < I(X; Y|X_r, Z, S) - \delta(\epsilon), \tag{6}$$

$$\widetilde{R} + R'' < I(X, X_r; Y|S) - \delta(\epsilon). \tag{7}$$

The proof is concluded by performing Fourier-Motzkin elimination, and letting $n \to \infty$, $B \to \infty$ and $\epsilon \to 0$. ■

## VI. Concluding Remarks and Some Open Problems

We presented the cooperative-bin-forward scheme and showed that it achieves the capacity region in a variety of semideterministic setups. While partial-decode-forward has been the scheme of interest in semideterministic setups, we demonstrated the advantages of cooperative-bin-forward by considering state-dependent setups, where partial-decode-forward is suboptimal, but cooperative-bin-forward is optimal. A number of interesting questions remain. Most importantly, how can the cooperative-bin-forward scheme be extended to, e.g. the model in Figure 2, when the source-relay link is not deterministic, but a general noisy link? Another interesting question is that of designing optimal achievability schemes for all the state-dependent setups considered in this paper when the state is known only to the source encoders, causally or strictly causally. Finally, the semideterministic relay channel with two state components, one known to the source and the other to the relay, with an uninformed destination, is also an interesting open question.

## Acknowledgments

The work of R. K. and A. Ö. was partly supported by a Stanford Graduate Fellowship, by NSF CAREER award 1254786 and by the Center for Science of Information (CSoI) under grant agreement CCF-0939370. The work of H. P. was supported by the Israel Science Foundation (grant no. 684/11) and the ERC starting grant.

## References

[1] A. Gamal and M. Aref, "The capacity of the semideterministic relay channel (corresp.)," *IEEE Trans. Inf. Theory*, vol. 28, no. 3, pp. 536–536, May 1982.

[2] Y.-H. Kim, "Capacity of a class of deterministic relay channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1328–1329, March 2008.

[3] A. El Gamal, N. Hassanpour, and J. Mammen, "Relay networks with delays," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3413–3431, Oct 2007.

[4] M. Li, O. Simeone, and A. Yener, "Message and state cooperation in a relay channel when only the relay knows the state," *CoRR*, vol. abs/1102.0768, 2011. [Online]. Available: http://arxiv.org/abs/1102.0768

[5] B. Akhbari, M. Mirmohseni, and M. Aref, "Compress-and-forward strategy for relay channel with causal and non-causal channel state information," *IET Comm.*, vol. 4, no. 10, pp. 1174–1186, Jul 2010.

[6] M. Khormuji and M. Skoglund, "The relay channel with partial causal state information," in *International Symposium on Information Theory and Its Applications*, Dec 2008, pp. 1–6.

[7] M. Khormuji, A. El Gamal, and M. Skoglund, "State-dependent relay channel: Achievable rate and capacity of a semideterministic class," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2629–2638, May 2013.

[8] Z. Deng, F. Lang, B.-Y. Wang, and S. mei Zhao, "Capacity of a class of relay channel with orthogonal components and non-causal channel state," in *IEEE International Symposium on Information Theory*, July 2013, pp. 2696–2700.

[9] A. Zaidi, S. Kotagiri, J. Laneman, and L. Vandendorpe, "Cooperative relaying with state available noncausally at the relay," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2272–2298, May 2010.

[10] A. Zaidi, S. Shamai, P. Piantanida, and L. Vandendorpe, "Bounds on the capacity of the relay channel with noncausal state at the source," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2639–2672, May 2013.

[11] I. Aguerri and D. Gunduz, "Capacity of a class of relay channels with state," in *IEEE Information Theory Workshop*, Sep 2012, pp. 277–281.

[12] R. Kolte, A. Ozgur, and H. Permuter, "Cooperative binning for semideterministic channels." [Online]. Available: http://arxiv.org/abs/1508.05149

[13] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2012.

# Random Coding Error Exponents for the Two-User Interference Channel

Wasim Huleihel and Neri Merhav
Technion - Israel Institute of Technology
Department of Electrical Engineering
Haifa 3200003, ISRAEL
E-mail: {wh@campus, merhav@ee}.technion.ac.il

***Abstract*—This paper is about deriving lower bounds on the error exponents for the two-user interference channel under the random coding regime for several ensembles. Specifically, we first analyze the standard random coding ensemble, where the codebooks are comprised of independently and identically distributed (i.i.d.) codewords. For this ensemble, we focus on optimum decoding, which is in contrast to other, suboptimal decoding rules that have been used in the literature (e.g., joint typicality decoding, treating interference as noise, etc.). The fact that the interfering signal is a codeword, rather than an i.i.d. noise process, complicates the application of conventional techniques of performance analysis of the optimum decoder. Also, unfortunately, these conventional techniques result in loose bounds. Using analytical tools rooted in statistical physics, as well as advanced union bounds, we derive single-letter formulas for the random coding error exponents. We compare our results with the best known lower bound on the error exponent, and show that our exponents can be strictly better. Then, in the second part of this paper, we consider more complicated coding ensembles, and find a lower bound on the error exponent associated with the celebrated Han-Kobayashi (HK) random coding ensemble, which is based on superposition coding.**

***Keywords*—Random coding, error exponent, interference channels, superposition coding, Han-Kobayashi scheme, statistical physics, optimal decoding, multiuser communication.**

## I. Introduction

### A. Previous Work

The two-user interference channel (IFC) models a general scenario of communication between two transmitters and two receivers (with no cooperation at either side), where each receiver decodes its intended message from an observed signal, which is interfered by the other user, and corrupted by channel noise. The information-theoretic analysis of this model has begun over more than four decades ago and has recently witnessed a resurgence of interest. Most of the previous work on multiuser communication, and specifically, on the IFC, has focused on obtaining inner and outer bounds to the capacity region (see, for example, [1, Ch. II.7]). In a nutshell, the study of this kind of channel has started in [2], and continued in [3], where simple inner and outer bounds to the capacity region were given. Then, in [4], by using the well-known superposition coding technique, the inner bound of [3] was strictly improved. In [5], various inner and outer bounds were obtained by transforming the IFC model into some multiple-access or broadcast channel. Unfortunately, the capacity region for the general interference channel is still unknown, although it has been solved for some very special cases [6, 7]. The best known inner bound is the Han-Kobayashi (HK) region, established in [8], and which will also be considered in this paper.

To our knowledge, [9, 10] are the only previous works which treat the error exponents for the IFC under optimal decoding. Specifically, [9] derives lower bounds on error exponents of random codebooks comprised of i.i.d. codewords uniformly distributed over a given type class, under maximum likelihood (ML) decoding at each user, that is, optimal decoding. Contrary to the error exponent analysis of other multiuser communication systems, such as the multiple access channel [11], the difficulty in analyzing the error probability of the optimal decoder for the IFC is due to statistical dependencies induced by the interfering signal. Indeed, for the IFC, the marginal channel determining each receiver's ML decoding rule is induced also by the codebook of the interfering user. This indeed extremely complicates the analysis, mostly because the interfering signal is a codeword and not an i.i.d. process. Another important observation, which was noticed in [9], is that the usual bounding techniques (e.g., Gallager's bounding technique) on the error probability fail to give tight results. To alleviate this problem, the authors of [9], combined some of the ideas from Gallager's bounding technique [12] to get an upper bound on the average probability of decoding error under ML decoding, the method of types [13], and used the method of distance enumerators, in the spirit of [14], which allows to avoid the use of Jensen's inequality in some steps. Finally, another relevant work is [15], where lower bounds on the error exponents of both standard and cognitive multiple-access channels (MACs), were derived assuming suboptimal successive decoding scheme.

### B. Contributions

The main purpose of this paper is to extend the study of achievability schemes to the more refined analysis of error exponents achieved by the two users, similarly as in [9]. Specifically, we derive single-letter expressions for the error exponents associated with the average error probability, for the finite-alphabet two-user IFC, under several random coding ensembles. The main contributions of this paper are as follows:

• Similarly as in recent works (see, e.g., [11, 16-19] and references therein) on the analysis of error exponents, we derive single-letter lower bounds for the random coding error exponents. For the standard random coding ensemble, considered in Subsection II-B, we analyze the optimal decoder

for each receiver, which is interested solely in its intended message. This is in contrast to usual decoding techniques analyzed for the IFC, in which each receiver decodes, in addition to its intended message, also part of (or all) the interfering codeword (that is, the other user's message), or other conventional achievability arguments [1, Ch. II.7], which are based on joint-typicality decoding, with restrictions on the decoder (such as, "treat interference as noise" or to "decode the interference"). This enables us to understand whether there is any significant degradation in performance due to the sub-optimality of the decoder. Also, since [9] analyzed the optimal decoder as well, we compare our formulas with those of [9], and show that our error exponent can be strictly better, which implies that the bounding technique in [9] is not tight. It is worthwhile to mention that the analytical formulas of our error exponents are simpler than the lower bound of [9].

- As was mentioned earlier, in [9] only random codebooks comprised of i.i.d. codewords (uniformly distributed over a type class) were considered. These ensembles are much simpler than the superposition codebooks of [8]. Unfortunately, it very tedious to analyze superposition codebooks using the methods of [9], and even if we do so, the tightness is questionable. In this paper, however, the new tools that we have derived enable us to analyze more involved random coding ensembles. Indeed, we can consider the coding ensemble used in the HK achievability scheme [8] and derive the respective error exponents. We also discuss an ensemble of hierarchical/tree codes [20].
- The analysis of the error exponents, carried out in this paper, turns out to be much more difficult than in previous works on point-to-point and multiuser communication problems, see, e.g., [11, 16-19]. Specifically, we encounter two main difficulties in our analysis: First, typically, when analyzing the probability of error, the first step is to apply the union bound. Usually, for point-to-point systems, under the random coding regime, the average error probability can be written as a union of pairwise independent error events. Accordingly, in this case, it is well known that the truncated union bound is exponentially tight [21, Lemma A.2]. This is no longer the case, however, when considering multiuser systems, and in particular, the IFC. For the IFC, the events comprising the union are strongly dependent, especially due to the fact that we are considering the optimal decoder. Indeed, recall that the optimal decoder for the first user, for example, declares that a certain message was transmitted if this message maximizes the likelihood pertaining to the marginal channel. This marginal channel[1] is the average of the actual channel over the messages of the interfering user, and thus depends on the whole codebook of the that user. Accordingly, the overall error event is the union of an exponential number of error events where each event depends on the marginal channel, and thus on the codebook of the interfering user. To alleviate this difficulty, following the ideas of [11], we derived new upper bounds on the probability of a union of events, which take into account the dependencies among the events. The second difficulty that we have encountered in our analysis is that in contrast to previous works, applying the type class enumerator method [14] is not simple, due to the reason mentioned above. Using some methods from large deviations theory, we were able to tackle this difficulty.

[1]The precise definition will be given in the sequel.

- Recently, in [15], the authors independently suggested lower bounds on the error exponents of both standard and cognitive multiple-access channels (MACs), assuming suboptimal successive decoding scheme, and using the standard random coding ensemble (considered in Subsection II-B). Although the motivation in [15] is different, the codebook construction and the decoding rule are the same as in the first part of this paper, and thus, essentially, their results apply also for the IFC. Now, despite the fact that the analysis in our paper is not the same as in [15], for the standard random coding ensemble, our lower bound coincides with that of [15]. More importantly, as was mentioned above, we consider also the more complicated ensemble pertaining to the HK scheme. Accordingly, the derivation of the lower bound on the error exponent of this ensemble is built upon the derivation of the lower bound on the error exponent of the standard random coding ensemble, and thus it makes useful and convenient to start with the analysis of the latter ensemble. We emphasize that the techniques used in [15] are not sufficient to analyze the ensemble pertaining to the HK scheme. Finally, we mention that the focus in [15] was on achievable rate region, rather than the error exponents, and thus no comparison to [9] was provided.
- We believe that by using the techniques and tools derived in this paper, other multiuser systems, such as the IFC with mismatched decoding, the MAC [11], the broadcast channel, the relay channel, etc., and accordingly, other coding schemes, such as binning [16], and hierarchical codes [20], can be analyzed.

### C. Notation Conventions

Throughout this paper, scalar random variables (RVs) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters, e.g. $X$, $x$, and $\mathcal{X}$, respectively. A similar convention will apply to random vectors of dimension $n$ and their sample values, which will be denoted with the same symbols in the boldface font. We also use the notation $X_i^j$ $(j > i)$ to designate the sequence of RVs $(X_i, X_{i+1}, \ldots, X_j)$. The set of all $n$-vectors with components taking values in a certain finite alphabet, will be denoted as the same alphabet superscripted by $n$, e.g., $\mathcal{X}^n$. Generic channels will be usually denoted by the letters $P$, $Q$, or $W$. We shall mainly consider joint distributions of two RVs $(X, Y)$ over the Cartesian product of two finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. For brevity, we will denote any joint distribution, e.g. $Q_{XY}$, simply by $Q$, the marginals will be denoted by $Q_X$ and $Q_Y$, and the conditional distributions will be denoted by $Q_{X|Y}$ and $Q_{Y|X}$. The joint distribution induced by $Q_X$ and $Q_{Y|X}$ will be denoted by $Q_X \times Q_{Y|X}$, and a similar notation will be used when the roles of $X$ and $Y$ are switched.

The expectation operator will be denoted by $\mathbb{E}\{\cdot\}$, and when we wish to make the dependence on the underlying distribution $Q$ clear, we denote it by $\mathbb{E}_Q\{\cdot\}$. Information measures induced by the generic joint distribution $Q_{XY}$, will be subscripted by $Q$, for example, $I_Q(X;Y)$ will denote the corresponding mutual information, etc. The divergence (or, Kullback-Liebler distance) between two probability measures $Q$ and $P$ will be denoted by $D(Q||P)$. The conditional information divergence between the conditional distributions

$Q_{Y|X}$ and $P_{Y|X}$, averaged over $P_X$, will be denoted by $D(Q_{Y|X}||P_{Y|X}|P_X)$. Logarithms are defined with respect to (w.r.t.) the natural basis, that is, $\log(\cdot) = \ln(\cdot)$, and finally, for a real number $x$, we denote $[x]_+ \triangleq \max\{0, x\}$.

## II. Problem Formulation and Main Results

### A. The IFC Model

Consider a two-user interference channel of two senders, two receivers, and a discrete memoryless channel (DMC), defined by a set of single-letter transition probabilities, $W_{Y_1Y_2|X_1X_2}(y_1y_2|x_1x_2)$, with finite input alphabets $\mathcal{X}_1, \mathcal{X}_2$ and finite output alphabets $\mathcal{Y}_1, \mathcal{Y}_2$. Here, each sender, $k \in \{1,2\}$, wishes to communicate an independent message $M_k \in \left\{1, 2, \ldots, 2^{nR_k}\right\}$ at rate $R_k$, and each receiver, $l \in \{1,2\}$, wishes to decode its respective message. Specifically, a $(2^{nR_1}, 2^{nR_2}, n)$ code $\mathcal{C}_n$ consists of:

- Two message sets $\mathcal{M}_1 \triangleq \left\{0, \ldots, 2^{nR_1} - 1\right\}$ and $\mathcal{M}_2 \triangleq \left\{0, \ldots, 2^{nR_2} - 1\right\}$ for the first and second users, respectively.
- Two encoders, where for each $k \in \{1,2\}$, the $k$-th encoder assigns a codeword $x_{k,i}^n \triangleq (x_{k,i,1}, x_{k,i,2}, \ldots, x_{k,i,n})$ to each message $i \in \mathcal{M}_k$.
- Two decoders, where each decoder $l \in \{1,2\}$ assigns an estimate $\hat{M}_l$ to $M_l$.

We assume that the message pair $(M_1, M_2)$ is uniformly distributed over $\mathcal{M}_1 \times \mathcal{M}_2$. It is clear that the *optimal decoder* of the first user, for this problem, is given by

$$\hat{M}_1 = \arg\max_{i \in \mathcal{M}_1} P\left(y_1^n | x_{1,i}^n\right) \tag{1}$$

$$= \arg\max_{i \in \mathcal{M}_1} e^{-nR_2} \sum_{j=1}^{M_2 - 1} P\left(y_1^n | x_{1,i}^n, x_{2,j}^n\right) \tag{2}$$

where $P\left(y_1^n | x_{1,i}^n, x_{2,j}^n\right)$ is the marginal channel defined as

$$P\left(y_1^n | x_{1,i}^n, x_{2,j}^n\right) \triangleq \prod_{k=1}^{n} W_{Y_1|X_1X_2}(y_{1k}|x_{1,i,k}x_{2,j,k}), \tag{3}$$

and

$$W_{Y_1|X_1X_2}(y_{1,k}|x_{1,i,k}x_{2,j,k}) \triangleq \sum_{y_{2,k} \in \mathcal{Y}_2} W_{Y_1Y_2|X_1X_2}(y_{1,k}y_{2,k}|x_{1,i,k}x_{2,j,k}). \tag{4}$$

The optimal decoder of the second user is defined similarly. Since there is no cooperation between the two receivers, the error probabilities for the code $\mathcal{C}_n$, are defined as:

$$P_{e,i}(\mathcal{C}_n) \triangleq 2^{-n(R_1+R_2)} \cdot \sum_{m_1, m_2} \mathbb{P}\left\{\hat{M}_i(Y_i^n) \neq m_i | M_1 = m_1, M_2 = m_2\right\}, \quad i = 1, 2. \tag{5}$$

### B. The Ordinary Random Coding Ensemble

In this subsection, we consider the ordinary random coding ensemble: For each $k \in \{1,2\}$, we select independently $M_k$ codewords $x_{k,i}^n$, for $i \in \mathcal{M}_k$, under the uniform distribution across the type class $T(P_{X_k})$, for a given distribution $P_{X_k}$ on $\mathcal{X}_k$. Our goal is to assess the exponential rate of $\bar{P}_{e,1}^{(n)} \triangleq \mathbb{E}\{P_{e,1}(\mathcal{C}_n)\}$, where the average is over the code ensemble, that is,

$$E_1^*(R_1, R_2) \triangleq \liminf_{n \to \infty} -\frac{1}{n} \log \bar{P}_{e,1}^{(n)}, \tag{6}$$

and similarly for the second user. Before stating the main result, we define some quantities. Given a joint distribution $Q_{X_1X_2Y_1}$ over $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_1$, consider the definitions in (7), shown at the top of the next page. Our main result is the following. Due to space limitation, the proofs of all the following results are omitted and can be found in [22].

*Theorem 1* Let $R_1$ and $R_2$ be given, and let $E^*(R_1, R_2)$ be defined as in (6). Consider the ensemble of fixed composition codes of types $P_{X_1}$ and $P_{X_2}$, for the first and second users, respectively. For a discrete memoryless two-user IFC, we have:

$$E_1^*(R_1, R_2) \geq \tilde{E}_1(R_1, R_2), \tag{8}$$

for any $R_1, R_2 \geq 0$.

Several remarks on Theorem 1 are in order.

- Due to symmetry, the error exponent for the second user, that is, $E_2^*(R_1, R_2)$ is simply obtained from Theorem 1 by swapping the roles of $X_1$, $Y_1$, and $R_1$, with $X_2$, $Y_2$, and $R_2$, respectively.
- An immediate byproduct of Theorem 1 is finding the set of rates $(R_1, R_2) \in \mathbb{R}_+^2$ for which $\tilde{E}_1(R_1, R_2) > 0$, namely, for which the probability of error vanishes exponentially as $n \to \infty$. It is not difficult to show that this set is given by:

$$\mathcal{R}_{\text{ordinary},1} = \hat{\mathcal{R}}_1 \cup \left\{(R_1, R_2) : \begin{array}{c} R_1 < I(X_1; Y_1|X_2) \\ R_1 + R_2 < I(X_1, X_2; Y_1) \end{array}\right\}, \tag{9}$$

evaluated with $P_{X_1X_2Y_1} = P_{X_1} \times P_{X_2} \times W_{Y_1|X_1X_2}$, where $\hat{\mathcal{R}}_1 \triangleq \{R_1 : R_1 < I(X_1; Y_1)\}$. Fig. 1 demonstrates a qualitative description of this region. The interpretation is as follows: The corner point $(I(X_1; Y_1|X_2), I(X_2; Y_1))$ is achieved by first decoding the interference (the second user), canceling it, and then decoding the first user. The sum-rate constraint can be achieved by joint decoding the two users (similarly to MAC), and thus, obviously, also by our optimal decoder. Finally, the region $R_1 < I(X_1; Y_1)$ and $R_2 \geq I(X_2; Y_1|X_1)$ means that we decode the first user while treating the interference as noise. Evidently, from the perspective of the first decoder, which is interested only in the message that is emitted from the first sender, the second sender can use any rate, and thus there is no bound on $R_2$ whenever $R_1 < I(X_1; Y_1)$. Note that this region was also obtained in [9], but from a lower bound on the error exponent. Accordingly, this means that according to [9], the achievable rate could be larger. Our results, however, show that one cannot do better when standard random coding is applied. Notice that $\mathcal{R}_{\text{ordinary},1}$ is well-known to be contained in the HK region [10, 23].
- *Existence of a single code:* our result holds true on the average, where the averaging is done over the random choice of codebooks. It can be shown (see, for example, [24, p. 2924]) that there exists deterministic sequence of fixed composition codebooks of increasing block length $n$ for which the same asymptotic error performance can be achieved for *both* users simultaneously.

$$f\left(Q_{X_1X_2Y_1}\right) \triangleq \mathbb{E}_Q\left[\log W_{Y_1|X_1X_2}(Y_1|X_1X_2)\right], \tag{7a}$$

$$t_0(Q_{X_1Y_1}) \triangleq R_2 + \max_{\hat{Q}\colon \hat{Q}_{X_2}=P_{X_2},\, \hat{Q}_{X_1Y_1}=Q_{X_1Y_1},\, I_{\hat{Q}}(X_2;X_1,Y_1)\le R_2} \left[f(\hat{Q}) - I_{\hat{Q}}(X_2;X_1,Y_1)\right], \tag{7b}$$

$$\mathcal{L}(\tilde{Q}_{X_1X_2Y_1}, Q_{X_1X_2Y_1}) \triangleq \Bigg\{\hat{Q}:\ \max\left[t_0(Q_{X_1X_2Y_1}), f(Q_{X_1X_2Y_1})\right] \le \max\left[f(\tilde{Q}_{X_1X_2Y_1}), f(\hat{Q}) + \left[R_2 - I_{\hat{Q}}(X_2;X_1,Y_1)\right]_+\right]\Bigg\}, \tag{7c}$$

$$E_1(\tilde{Q}_{X_1X_2Y_1}, Q_{X_1X_2Y_1}) \triangleq \min_{\hat{Q}\colon \hat{Q}_{X_2}=P_{X_2},\, \hat{Q}_{X_1Y_1}=\tilde{Q}_{X_1Y_1},\, \hat{Q}\in\mathcal{L}(\tilde{Q}_{X_1X_2Y_1}, Q_{X_1X_2Y_1})} \left[I_{\hat{Q}}(X_2;X_1,Y_1) - R_2\right]_+, \tag{7d}$$

$$\hat{E}_1(Q_{X_1X_2Y_1}, R_2) \triangleq \min_{\tilde{Q}\colon \tilde{Q}_{X_1}=P_{X_1},\, \tilde{Q}_{X_2Y_1}=Q_{X_2Y_1}} \left[I_{\tilde{Q}}(X_1;X_2,Y_1) + E_1(\tilde{Q}_{X_1X_2Y_1}, Q_{X_1X_2Y_1})\right], \tag{7e}$$

$$\tilde{E}_1(R_1,R_2) \triangleq \min_{Q_{Y_1|X_1X_2}\colon Q_{X_1}=P_{X_1}, Q_{X_2}=P_{X_2}} \left[D(Q_{Y_1|X_1X_2}||W_{Y_1|X_1X_2}|P_{X_1}\times P_{X_2}) + \left[\hat{E}_1(Q_{X_1X_2Y_1}, R_2) - R_1\right]_+\right]. \tag{7f}$$

$$\bar{P}_{e,1}^{(n)} = \Pr\left[\bigcup_{i=1}^{M_1-1}\left\{\sum_{j=0}^{M_2-1} P\left(Y_1^n|X_{1,i}^n, X_{2,j}^n\right) \ge \sum_{j=0}^{M_2-1} P\left(Y_1^n|X_{1,0}^n, X_{2,j}^n\right)\right\}\right], \tag{10}$$

$$= \mathbb{E}\left\{\Pr\left[\bigcup_{i=1}^{M_1-1}\left\{\sum_{j=0}^{M_2-1} P\left(Y_1^n|X_{1,i}^n, X_{2,j}^n\right) \ge \sum_{j=0}^{M_2-1} P\left(Y_1^n|X_{1,0}^n, X_{2,j}^n\right)\right\}\middle|\, \mathcal{F}_0\right]\right\}, \tag{11}$$

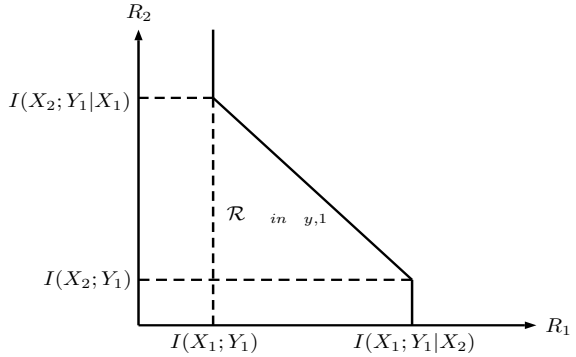



Fig. 1. Rate region $\mathcal{R}_{\text{ach},1}$ for which $E_1^*(R_1,R_2) > 0$.

• *On the proof:* it is instructive to discuss (in some more detail than earlier) one of the main difficulties in proving Theorem 1, which is customary to multiuser systems, such as the IFC. Without loss of generality, we assume throughout, that the transmitted codewords are $x_{1,0}^n$ and $x_{2,0}^n$. Accordingly, the average probability of error associated with the decoder (2) is given by (11), shown at the top of the next page, where $\mathcal{F}_0 \triangleq \left(X_{1,0}^n, X_{2,0}^n, Y_1^n\right)$. By the union bound and Shulman's inequality [21, Lemma A.2], we know that for a sequence of pairwise independent events, $\{\mathcal{A}_i\}_{i=1}^N$, the following holds:

$$\frac{1}{2}\min\left\{1, \sum_{i=1}^N \Pr\{\mathcal{A}_i\}\right\} \le \Pr\left\{\bigcup_{i=1}^N \mathcal{A}_i\right\} \le \min\left\{1, \sum_{i=1}^N \Pr\{\mathcal{A}_i\}\right\}, \tag{12}$$

which is a useful result when assessing the exponential behavior of such probabilities. Equation (12) is one of the building blocks of tight exponential analysis of previously considered point-to-point systems (see, e.g., [16-19], and many references therein). However, it is evident that in our case the various events are not pairwise independent, and therefore this result cannot be applied directly. Indeed, since we are interested in the optimal decoder, each event of the union in (11), depends on the whole codebook of the second user. One may speculate that this problem can be tackled by conditioning on the codebook of the second user, and then (12). However, the cost of this conditioning is a very complicated (if not intractable) large deviations analysis of some quantities. To alleviate this problem, we derived new upper bounds on the probability of union of events, which takes into account the dependencies among the events. This was done using the techniques of [11].

Another difficulty that arises in the error exponent analysis of the IFC model, is that in contrast to previous works, applying the distance enumerator[2] method [14], is not a simple task. Again, our optimal decoder compares two quantities (i.e., likelihoods) which are both depend on the whole codebook of the second user. The consequence of this situation, is that in order to analyze the probability of error, it is required to analyze the joint distribution of type class enumerators, and not just rely on their marginal distributions, as usually done, e.g., [16-19].

• *Comparison with [9]*: Similarly to [9], we present results

[2]For a given $y^n \in \mathcal{Y}^n$, and a given joint probability distribution $Q_{XY}$ on $\mathcal{X}\times\mathcal{Y}$, the *distance enumerator* (or, *type class enumerator*), $N(Q_{XY})$, is the number of codewords $\{x_i^n\}$ in $\mathcal{C}_n$ whose conditional empirical joint distribution with $y^n$ is $Q_{XY}$, namely, $N(Q_{XY}) = \left|x^n \in \mathcal{C}_n :\ \hat{Q}_{x^n y^n} = Q_{XY}\right|$, where $\hat{Q}_{x^n y^n}$ is the empirical joint distribution of $x^n$ and $y^n$, and $|\mathcal{A}|$ designates the cardinality of a finite set $\mathcal{A}$.

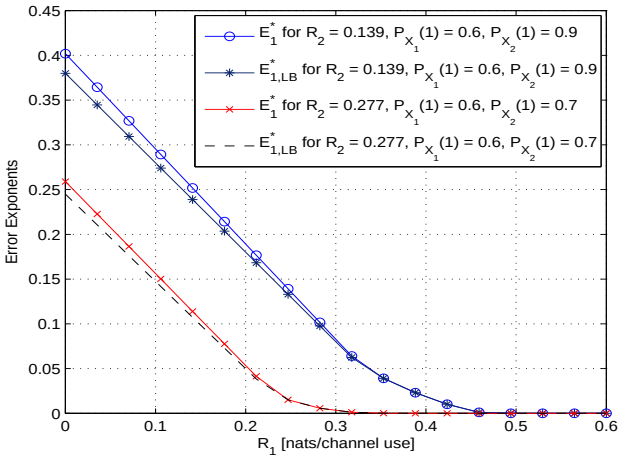



Fig. 2. Comparison between our error exponent $E_1^*(R_1, R_2)$ and the lower bound $E_{\text{LB}}(R_1, R_2)$ of [9], as a function of $R_1$ for two different values of $R_2$ and fixed choices of $P_{X_1}$ and $P_{X_2}$.

for the binary $Z$-channel model defined as follows: $Y_1 = X_1 \cdot X_2 \oplus Z$ and $Y_2 = X_2$, where $X_1, X_2, Y_1, Y_2 \in \{0, 1\}$, $Z \sim \text{Bern}(p)$, "$\cdot$" is multiplication, and "$\oplus$" is modulo-2 addition. In the numerical calculations, we fix $p = 0.01$. Fig. 2 presents the lower bound on the error exponent under optimal decoding, derived in this paper, compared to the lower bound $E_{\text{LB}}(R_1, R_2)$ of [9], as a function of $R_1$, for different values of $P_{X_1}$, $P_{X_2}$, and $R_2$. It can be seen that our exponents can be strictly better than those of [9].

- *Generalization to other ensemble*: As was mentioned before, in [9] only random codebooks comprised of i.i.d. codewords were considered. These ensembles are much simpler than the superposition codebooks of [8]. Unfortunately, it very tedious to analyze superposition codebooks using the methods of [9], and even if we do so, the tightness is questionable. However, the new tools that we have derived enable us to analyze more involved random coding ensembles. Due to space limitations, we do not present the error exponents achieved by the following schemes. All the details can be found in [22, Subsection III.C]. For example, we can derive the error exponents for the HK scheme, which gives the best known inner bound. The idea in this scheme is to split the message $M_1$ into "private" and "common" messages, $M_{11}$ and $M_{12}$ at rates $R_{11}$ and $R_{12}$, respectively, such that $R_1 = R_{11} + R_{12}$. Similarly $M_2$ is split into $M_{21}$ and $M_{22}$ at rates $R_{21}$ and $R_{22}$, respectively, such that $R_2 = R_{21} + R_{22}$. Then, receiver $k = 1, 2$, recovers its intended message $M_k$, and the common message from the other sender (although it is not required to) each decoder. Also, using the same techniques, we can analyze the error exponents resulting from the *hierarchical code ensemble* [20], in which the case has a tree structure with two levels, where the first serves for "cloud centers", and the second for the "satellites".

## Acknowledgment

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

## References

[1] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[2] E. C. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1. Berkeley, CA: Univ. California Press, 1961, pp. 611–64.

[3] R. Ahlswede, "The capacity region of a channel with two senders and two receivers," *Annals Probabil.*, vol. 2, no. 5, pp. 805–814, 1974.

[4] A. B. Carleial, "Interference channels," *IEEE Trans. on Inf. Theory*, vol. IT-24, pp. 60–70, Jan. 1978.

[5] H. Sato, "Two-user communication channels," *IEEE Trans. on Inf. Theory*, vol. IT-23, pp. 295–304, May. 1977.

[6] A. B. Carleial, "A case where intereference does not reduce capacity," *IEEE Trans. on Inf. Theory*, vol. IT-21, pp. 569–570, Sep. 1975.

[7] R. Benzel, "The capacity region of a class of discrete additive degraded interference channels," *IEEE Trans. on Inf. Theory*, vol. IT-25, pp. 228–231, Mar. 1979.

[8] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. on Inf. Theory*, vol. IT, no. 27, pp. 49–60, Jan. 1981.

[9] R. Etkin, N. Merhav, and E. Ordentlich, "Error exponents of optimum decoding for the interference channel," *IEEE Trans. on Inf. Theory*, vol. 56, no. 1, pp. 40–56, Jan. 2010.

[10] C. Chang, R. Etkin, and E. Ordentlich, "Interference channel capacity region for randomized fixed-composition codes," *HP Labs Technical Report*, 2009. [Online]. Available: http://www.hpl.hp.com/techreports/2008/HPL-2008-194R1.html

[11] J. Scarlett, A. Martinez, and A. G. Fábregas, "Multiuser coding techniques for mismatched decoding," *submitted to IEEE Trans. on Inf. Theory*, Nov. 2013. [Online]. Available: arxiv.org/pdf/1311.6635

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

[13] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 2011.

[14] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," *IEEE Trans. on Inf. Theory*, vol. 55, no. 1, pp. 83–92, Jan. 2009.

[15] J. Scarlett, A. Martinez, and A. G. Fábregas, "Mismatched multi-letter successive decoding for the multiple-access channel," in *Proc. ISIT 2014*, Jul. 2014, pp. 2539–2543.

[16] N. Merhav, "Exact random coding exponents of optimal bin index decoding," *IEEE Trans. on Inf. Theory*, vol. 60, no. 10, pp. 6024–6031, Oct. 2014.

[17] ——, "List decoding-random coding exponents and expurgated exponents," *IEEE Trans. on Inf. Theory*, vol. 60, no. 11, pp. 6749–6759, Nov. 2014.

[18] ——, "Exact correct-decoding exponent for the wiretap channel decoder," *IEEE Trans. on Inf. Theory*, vol. 60, no. 12, pp. 7606–7615, Dec. 2014.

[19] W. Huleihel, N. Weinberger, and N. Merhav, "Erasure/list random coding error exponents are not universally achievable," *submitted to IEEE Trans. on Inf. Theory*, Oct. 2014. [Online]. Available: http://arxiv.org/abs/1410.7005

[20] N. Merhav, "The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes," *IEEE Trans. on Inf. Theory*, vol. 55, no. 3, pp. 1250–1268, May 2009.

[21] N. Shulman, "*Communication over an unknown channel via common broadcasting*," Ph.D. dissertation, Tel-Aviv University, 2003, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.407.7542.

[22] W. Huleihel and N. Merhav, "Exact random coding error exponents for the two-user interference channel," Mar. 2015. [Online]. Available: http://webee.technion.ac.il/people/merhav/papers/p179arxiv.pdf

[23] B. Bandemer, A. El Gamal, and Y. K. Kim, "Optimal achievable rates for interference networks with random codes," *submitted to IEEE Trans. on Inf. Theory*, 2012. [Online]. Available: http://arxiv.org/abs/1210.4596

[24] L. Weng, S. S. Pradhan, and A. Anastasopoulos, "Error exponent regions for Gaussian broadcast and multiple-access channels," *IEEE Trans. on Inf. Theory*, vol. 54, no. 7, pp. 2919–2942, July 2008.

# A Necessary and Sufficient Condition for the Asymptotic Tightness of the Shannon Lower Bound

Tobias Koch
Universidad Carlos III de Madrid, Spain
& Gregorio Marañón Health Research Institute
Email: koch@tsc.uc3m.es

***Abstract*—The Shannon lower bound is one of the few lower bounds on the rate-distortion function that holds for a large class of sources. In this paper, it is demonstrated that its gap to the rate-distortion function vanishes as the allowed distortion tends to zero for all sources that have a finite differential entropy and whose integer parts have a finite entropy. Conversely, it is demonstrated that if the integer part of the source has an infinite entropy, then its rate-distortion function is infinite for any finite distortion. Consequently, the Shannon lower bound provides an asymptotically tight bound on the rate-distortion function if, and only if, the integer part of the source has a finite entropy.**

## I. Introduction

Suppose a source produces the sequence of independent and identically distributed (i.i.d.), real-valued, random variables $\{X_k,\, k \in \mathbb{Z}\}$ according to the distribution $P_X$, and suppose that we employ a vector quantizer that produces a sequence of reconstruction symbols $\{\hat{X}_k,\, k \in \mathbb{Z}\}$ satisfying

$$\overline{\lim_{n\to\infty}} \frac{1}{n}\sum_{k=1}^{n} \mathsf{E}\Big[\big(X_k - \hat{X}_k\big)^2\Big] \leq D. \tag{1}$$

(We use $\overline{\lim}$ to denote the *limit superior* and $\underline{\lim}$ to denote the *limit inferior*.) Rate-distortion theory tells us that if for every blocklength $n$ and distortion constraint $D$ we quantize the sequence of source symbols $X_1, \ldots, X_n$ to one of $e^{nR(D)}$ possible sequences of reconstruction symbols $\hat{X}_1, \ldots, \hat{X}_n$, then the smallest rate $R(D)$ (in nats per source symbol) for which there exists a vector quantizer satisfying (1) is given by [1], [2]

$$R(D) = \inf_{P_{\hat{X}|X}} I\big(X;\hat{X}\big) \tag{2}$$

where the infimum is over all conditional distributions of $\hat{X}$ given $X$ satisfying

$$\mathsf{E}\Big[\big(X - \hat{X}\big)^2\Big] \leq D \tag{3}$$

and where the expectation is computed with respect to the joint distribution $P_X P_{\hat{X}|X}$. Here and throughout the paper we omit the time indices where they are immaterial. The rate $R(D)$ as a function of $D$ is referred to as the *rate-distortion function*.

This work has been supported in part by a Marie Curie Career Integration Grant through the 7th European Union Framework Programme under Grant 333680, by the Ministerio de Economía y Competitividad of Spain under Grants TEC2013-41718-R, RYC-2014-16332, and TEC2015-69648-REDC, and by the Comunidad de Madrid under Grant S2013/ICE-2845.

Unfortunately, the rate-distortion function is unknown except in very few special cases. It therefore needs to be assessed by means of upper and lower bounds. Arguably, for sources with a finite differential entropy, the most important lower bound is the *Shannon lower bound* [1], [2]

$$R_{\mathrm{SLB}}(D) = h(X) - \frac{1}{2}\log(2\pi e D) \tag{4}$$

where $\log(\cdot)$ denotes the natural logarithm. While this lower bound is tight only for some special sources, it converges to the rate-distortion function as the allowed distortion $D$ tends to zero, provided that the source satisfies some conditions [3]–[6]. Thus, in this case the Shannon lower bound provides a good approximation of the rate-distortion function for small distortions. A finite-blocklength refinement of the Shannon lower bound has recently been given by Kostina [7].

To the best of our knowledge, the most general conditions for the asymptotic tightness of the Shannon lower bound are due to Linder and Zamir [6]. While Linder and Zamir considered more general distortion measures, specialized to quadratic distortion (3), they showed the following.

*Theorem 1 (Linder and Zamir [6, Cor. 1]):* Assume that $X$ has a probability density function (pdf) and $h(X) > -\infty$. Further assume that there exists an $\alpha > 0$ such that $\mathsf{E}[|X|^\alpha] < \infty$. Then the Shannon lower bound is asymptotically tight, i.e.,

$$\lim_{D\downarrow 0}\big\{R(D) - R_{\mathrm{SLB}}(D)\big\} = 0. \tag{5}$$

The theorem's conditions are very mild and satisfied by the most common source distributions. In fact, Theorem 1 demonstrates that the Shannon lower bound is asymptotically tight even if there exists no quantizer with a finite number of codevectors and of finite distortion, i.e., when $\mathsf{E}\big[X^2\big] = \infty$. However, the conditions are more stringent than the ones sometimes required for the analysis of the rate and distortion redundancies of high-resolution quantizers. For example, in [8] Gray et al. analyzed the asymptotic distortion of entropy-constrained vector quantization in the limit as the rate tends to infinity, thereby rigorously proving a theorem by Zador [9]. In their work, they considered source vectors $X$ that have a density, whose differential entropy is finite, and that satisfy

$$H(\lfloor X \rfloor) < \infty \tag{6}$$

where $\lfloor a \rfloor$ denotes the integer part of $a$, i.e., the largest integer not larger than $a$. Furthermore, Koch and Vazquez-Vilar [10]

demonstrated that these assumptions are also sufficient to recover the result by Gish and Pierce [11] that, among all scalar quantizers, uniform quantizers are asymptotically optimal as the allowed distortion tends to zero. In words, condition (6) demands that quantizing the source with a uniform quantizer of unit-length cells gives rise to a discrete random variable of finite entropy. This ensures that the quantizer output can be further compressed using a lossless variable-length code of finite expected length.

The quantity $H(\lfloor X \rfloor)$ is intimately related with the *Rényi information dimension* [12], defined as

$$d(X) \triangleq \lim_{m\to\infty} \frac{H\left(\lfloor mX \rfloor / m\right)}{\log m}, \quad \text{if the limit exists} \tag{7}$$

which in turn coincides with the *rate-distortion dimension* introduced by Kawabata and Dembo [13]; see also [14]. Indeed, the Rényi information dimension is finite if, and only if, condition (6) is satisfied [14, Prop. 1]. Furthermore, a sufficient condition for finite Rényi information dimension is $\mathsf{E}[\log(1+|X|)] < \infty$ [14, Prop. 1], which in turn holds for any source for which $\mathsf{E}[|X|^\alpha] < \infty$ for some $\alpha > 0$. Thus, (6) is weaker than the assumption that $\mathsf{E}[|X|^\alpha] < \infty$.

It is common to assume that the differential entropy of the source is finite, since otherwise the Shannon lower bound (4) is uninteresting. One may thus wonder how (6) and the assumption of a finite differential entropy are related. As demonstrated, for example, in the proof of Theorem 3 in [15], condition (6) implies that $h(X) < \infty$. In fact, one can show that if (6) holds and $X$ has a pdf, then $h(X) \leq H(\lfloor X \rfloor)$ [16, Cor. 1]. Conversely, one can find sources for which the differential entropy is finite but $H(\lfloor X \rfloor)$ is infinite. For example, consider a source with pdf

$$f_X(x) = \sum_{m=2}^{\infty} p_m m \, \mathbb{1}\left\{ m \leq x < m + \frac{1}{m} \right\}, \quad x \in \mathbb{R} \tag{8}$$

where

$$p_m = \frac{1}{\mathsf{K} m \log^2 m}, \qquad m = 2, 3, \ldots \tag{9a}$$

$$\mathsf{K} = \sum_{m=2}^{\infty} \frac{1}{m \log^2 m} \tag{9b}$$

and $\mathbb{1}\{\cdot\}$ denotes the indicator function. It is easy to check that for such a source

$$\begin{aligned} H(\lfloor X \rfloor) &= \sum_{m=2}^{\infty} p_m \log \frac{1}{p_m} \\ &= \sum_{m=2}^{\infty} \frac{\log \mathsf{K} + \log m + 2 \log\log m}{\mathsf{K} m \log^2 m} \\ &= \infty \end{aligned} \tag{10}$$

and

$$\begin{aligned} h(X) &= -\int_{\mathbb{R}} f_X(x) \log f_X(x) \, \mathrm{d}x \\ &= \sum_{m=2}^{\infty} \frac{\log \mathsf{K} + 2 \log\log m}{\mathsf{K} m \log^2 m} \\ &< \infty. \end{aligned} \tag{11}$$

(See remark after Theorem 1 in [12, pp. 197–198].) Thus, for sources satisfying $h(X) > -\infty$, a finite Rényi information dimension implies a finite differential entropy but not vice versa.

In this paper, we demonstrate that for sources that have a pdf and whose differential entropy is finite, the Shannon lower bound (4) is asymptotically tight if (6) is satisfied. This ensures the asymptotic tightness of the Shannon lower bound under the most general conditions imposed in the analysis of high-resolution quantizers. Conversely, we demonstrate that for sources that do not satisfy (6) the rate-distortion function is infinite for any finite distortion.

## II. Problem Setup and Main Result

We consider a one-dimensional, real-valued source $X$ with support $\mathcal{X} \subseteq \mathbb{R}$ whose distribution is absolutely continuous with respect to the Lebesgue measure, and we denote its pdf by $f_X$. We assume that $x \mapsto f_X(x) \log f_X(x)$ is integrable, ensuring that the differential entropy

$$h(X) \triangleq -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x \tag{12}$$

is well-defined and finite. We have the following result.

*Theorem 2 (Main Result):* Assume that the one-dimensional, real-valued source $X$ has a pdf and $h(X) > -\infty$. If $H(\lfloor X \rfloor) < \infty$, then the Shannon lower bound is asymptotically tight, i.e.,

$$\lim_{D\downarrow 0} \big\{ R(D) - R_{\text{SLB}}(D) \big\} = 0. \tag{13}$$

Conversely, if $H(\lfloor X \rfloor) = \infty$, then $R(D) = \infty$ for $D > 0$.

*Proof:* See Section III. ∎

Theorem 2 thus demonstrates that the Shannon lower bound is asymptotically tight if, and only if, $H(\lfloor X \rfloor)$ is finite.

## III. Proof of Theorem 2

The proof consists of two parts. In the first part, we show that if $H(\lfloor X \rfloor) < \infty$, then the Shannon lower bound is asymptotically tight (Section III-A). In the second part, we show that if $H(\lfloor X \rfloor) = \infty$, then $R(D) = \infty$ for every $D > 0$ (Section III-B).

### A. Asymptotic Tightness

In this section, we demonstrate the asymptotic tightness of the Shannon lower $R_{\text{SLB}}(D)$ for sources that satisfy $H(\lfloor X \rfloor) < \infty$ and $h(X) > -\infty$. The first steps in our proof are identical to the ones in the proof of Theorem 1 in [6]. To keep this paper self-contained, we reproduce the main steps.

To prove the asymptotic tightness of $R_{\text{SLB}}(D)$, we derive an upper bound on $R(D)$ whose gap to $R_{\text{SLB}}(D)$ vanishes as $D$ tends to zero. In view of (2), an upper bound on $R(D)$ follows by choosing $\hat{X} = X + Z_D$, where $Z_D$ is a zero-mean, variance-$D$, Gaussian random variable that is independent of $X$. It follows that

$$\begin{aligned} R(D) &\leq I(X; X + Z_D) \\ &= h(X + Z_D) - h(Z_D). \end{aligned} \tag{14}$$

Furthermore, using that $h(Z_D) = \frac{1}{2}\log(2\pi eD)$, the Shannon lower bound (4) can be written as

$$R_{\mathrm{SLB}}(D) = h(X) - h(Z_D). \tag{15}$$

Combining (14) and (15) gives

$$0 \le R(D) - R_{\mathrm{SLB}}(D) \le h(X + Z_D) - h(X). \tag{16}$$

Thus, the asymptotic tightness of $R_{\mathrm{SLB}}(D)$ follows by proving that

$$\overline{\lim_{D\downarrow 0}}\, h(X + Z_D) \le h(X). \tag{17}$$

To this end, we follow the steps (17)–(21) in [6] but with $Y_{\Delta(D)}$ and $Y_{\Delta(0)}$ replaced by the random variables $Y_D$ and $Y_0$ with respective pdfs

$$f_{Y_D}(y) = \sum_{i\in\mathbb{Z}} \Pr\big(\lfloor X + Z_D\rfloor = i\big)\mathbb{1}\{\lfloor y\rfloor = i\} \tag{18a}$$

$$f_{Y_0}(y) = \sum_{i\in\mathbb{Z}} \Pr\big(\lfloor X\rfloor = i\big)\mathbb{1}\{\lfloor y\rfloor = i\}. \tag{18b}$$

It follows that

$$D(f_{X+Z_D}\|f_{Y_D}) = H(\lfloor X + Z_D\rfloor) - h(X + Z_D) \tag{19}$$

and

$$D(f_X\|f_{Y_0}) = H(\lfloor X\rfloor) - h(X). \tag{20}$$

The random variable $Z_D$ converges to zero almost surely as $D$ tends to zero and, hence, also in distribution. Since $X$ and $Z_D$ are independent, it follows that $X + Z_D \to X$ in distribution as $D$ tends to zero. Furthermore, since by assumption the distribution of $X$ is absolutely continuous with respect to the Lebesgue measure, for every $i \in \mathbb{Z}$ the interval $[i, i+1)$ is a continuity set of $X$, so

$$\lim_{D\downarrow 0} \Pr\big(\lfloor X + Z_D\rfloor = i\big) = \Pr\big(\lfloor X\rfloor = i\big), \quad i \in \mathbb{Z}. \tag{21}$$

Thus, the pdf of $Y_D$ (18a) converges pointwise to the pdf of $Y_0$ (18b), which by Scheffe's lemma [17, Th. 16.12] implies that $Y_D \to Y_0$ in distribution as $D$ tends to zero.

By the lower semicontinuity of relative entropy (see, e.g., the proof of Lemma 4 in [18] and references therein),

$$\varliminf_{D\downarrow 0} D(f_{X+Z_D}\|f_{Y_D}) \ge D(f_X\|f_{Y_0}). \tag{22}$$

Combining (22) with (19) and (20) yields

$$\varliminf_{D\downarrow 0}\big\{H(\lfloor X+Z_D\rfloor) - h(X+Z_D)\big\} \ge H(\lfloor X\rfloor) - h(X). \tag{23}$$

Since $h(X) > -\infty$ and $H(\lfloor X\rfloor) < \infty$, the claim (17) (and hence the asymptotic tightness of $R_{\mathrm{SLB}}(D)$) follows by showing that $H(\lfloor X + Z_D\rfloor)$ tends to $H(\lfloor X\rfloor)$ as $D$ tends to zero. We present this result in the following lemma.

*Lemma 1:* Assume that $X$ has a pdf and $H(\lfloor X\rfloor) < \infty$. Let $Z_D$ be a zero-mean, variance-$D$, Gaussian random variable that is independent of $X$. Then

$$\lim_{D\downarrow 0} H(\lfloor X + Z_D\rfloor) = H(\lfloor X\rfloor). \tag{24}$$

*Proof:* Using basic properties of entropy, we obtain

$$\begin{aligned} H(\lfloor X + Z_D\rfloor) &\le H(\lfloor X\rfloor) + H\big(\lfloor X + Z_D\rfloor \,\big|\, \lfloor X\rfloor\big) \\ &\le H(\lfloor X\rfloor) + H(V_D) \end{aligned} \tag{25}$$

and

$$\begin{aligned} H(\lfloor X + Z_D\rfloor) &\ge H(\lfloor X\rfloor) - H\big(\lfloor X\rfloor \,\big|\, \lfloor X + Z_D\rfloor\big) \\ &\ge H(\lfloor X\rfloor) - H(V_D) \end{aligned} \tag{26}$$

where $V_D \triangleq \lfloor X + Z_D\rfloor - \lfloor X\rfloor$. Lemma 1 follows therefore by showing that $H(V_D)$ vanishes as $D$ tends to zero.

We first show that

$$\lim_{D\downarrow 0} \Pr(V_D = i) = \mathbb{1}\{i = 0\}. \tag{27}$$

Indeed, let $\bar{X} \triangleq X - \lfloor X\rfloor$, and recall that $Z_D \to 0$ in distribution as $D$ tends to zero. Noting that $V_D = \lfloor \bar{X} + Z_D\rfloor$, the probability mass function of $V_D$ can be written as

$$\Pr\big(V_D = i\big) = \Pr\big(\lfloor \bar{X} + Z_D\rfloor = i\big), \quad i \in \mathbb{Z}. \tag{28}$$

Furthermore, the independence of $X$ and $Z_D$ implies that $\bar{X} + Z_D \to \bar{X}$ in distribution as $D$ tends to zero. Since the distribution of $X$ is absolutely continuous with respect to the Lebesgue measure, so is the distribution of $\bar{X}$. Consequently, for every $i \in \mathbb{Z}$ the interval $[i, i+1)$ is a continuity set of $\bar{X}$ and

$$\lim_{D\downarrow 0} \Pr\big(\lfloor \bar{X} + Z_D\rfloor = i\big) = \Pr\big(\lfloor \bar{X}\rfloor = i\big) = \mathbb{1}\{i = 0\} \tag{29}$$

where the last step follows because the support of $\bar{X}$ is $[0, 1)$. This proves (27).

We continue by expressing the entropy of $V_D$ as

$$\begin{aligned} H(V_D) = &\sum_{i=-1}^{1} \Pr(V_D = i)\log\frac{1}{\Pr(V_D = i)} \\ &+ \sum_{i\in\mathbb{Z}:\ |i|>1} \Pr(V_D = i)\log\frac{1}{\Pr(V_D = i)}. \end{aligned} \tag{30}$$

The first sum on the right-hand side (RHS) of (30) consists of finitely many terms, so (27) and the continuity of $x \mapsto x\log(1/x)$ give[1]

$$\begin{aligned} &\lim_{D\downarrow 0} \sum_{i=-1}^{1} \Pr(V_D = i)\log\frac{1}{\Pr(V_D = i)} \\ &\quad= \sum_{i=-1}^{1} \lim_{D\downarrow 0} \Pr(V_D = i)\log\frac{1}{\Pr(V_D = i)} \\ &\quad= 0. \end{aligned} \tag{31}$$

To show that the second sum on the RHS of (30) vanishes as $D \to 0$, it suffices to show that

$$\overline{\lim_{D\downarrow 0}} \sum_{i\in\mathbb{Z}:\ |i|>1} \Pr(V_D = i)\log\frac{1}{\Pr(V_D = i)} \le 0 \tag{32}$$

since the summands are nonnegative. As observed above, the distribution of $\bar{X}$ is absolutely continuous with respect to the

[1] Here and throughout the paper we define $0\log(1/0) \triangleq 0$.

Lebesgue measure, so $\bar{X}$ has a pdf which we shall denote by $f_{\bar{X}}$. Since $Z_D$ and $\bar{X}$ are independent, the pdf of $\bar{X}+Z_D$ is given by [19, Th. 4.10, p. 29]

$$f_{\bar{X}+Z_D}(\xi) = \int_0^1 f_{\bar{X}}(\bar{x}) \frac{1}{\sqrt{2\pi D}} e^{-\frac{(\xi-\bar{x})^2}{2D}} \,\mathrm{d}\bar{x}, \quad \xi \in \mathbb{R}. \tag{33}$$

Combining (28) and (33), we obtain

$$\begin{aligned} \Pr\big(V_D = i\big) &= \Pr\big(\lfloor \bar{X} + Z_D \rfloor = i\big) \\ &= \int_i^{i+1} \int_0^1 f_{\bar{X}}(\bar{x}) \frac{1}{\sqrt{2\pi D}} e^{-\frac{(\xi-\bar{x})^2}{2D}} \,\mathrm{d}\bar{x}\,\mathrm{d}\xi \\ &\geq \frac{1}{\sqrt{2\pi D}} e^{-\frac{(|i|+1)^2}{2D}}, \quad i \in \mathbb{Z} \end{aligned} \tag{34}$$

where the inequality follows because for $\xi \in [i, i+1)$ and $\bar{x} \in [0,1)$ we have $|\xi - \bar{x}| \leq |i| + 1$. Applying (34) to the second sum on the RHS of (30) gives

$$\begin{aligned} &\sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) \log \frac{1}{\Pr(V_D = i)} \\ &\quad \leq \frac{1}{2} \log(2\pi D) \sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) \\ &\qquad + \sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) \frac{(|i|+1)^2}{2D}. \end{aligned} \tag{35}$$

To demonstrate that the first term on the RHS of (35) vanishes as $D \to 0$, we use that any variable $z$ satisfying $\big|\lfloor \bar{x} + z \rfloor\big| > 1$ must also satisfy $|z| > 1$, irrespective of $\bar{x} \in [0,1)$. Consequently,

$$\begin{aligned} \sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) &= \Pr\big(\big|\lfloor \bar{X} + Z_D \rfloor\big| > 1\big) \\ &\leq \Pr\big(|Z_D| > 1\big) \\ &\leq D \end{aligned} \tag{36}$$

where the last inequality follows from Chebyshev's inequality [20, Th. 4.10.7, p. 192]. Combining (36) with (35), we obtain

$$\left| \frac{1}{2} \log(2\pi D) \sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) \right| \leq \frac{D}{2} \big|\log(2\pi D)\big| \tag{37}$$

which tends to zero as $D \to 0$.

We next consider the second term on the RHS of (35). To this end, we write

$$\Pr(V_D = i)(|i|+1)^2 = \int_i^{i+1} f_{\bar{X}+Z_D}(\xi)(|i|+1)^2 \,\mathrm{d}\xi. \tag{38}$$

By Fubini's theorem [20, Th. 2.6.4, p. 105], we obtain from (38) and (33) that

$$\begin{aligned} &\sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i)(|i|+1)^2 \\ &= \sum_{i\in\mathbb{Z}\colon\, |i|>1} \int_i^{i+1} \int_0^1 f_{\bar{X}}(\bar{x}) \frac{1}{\sqrt{2\pi D}} e^{-\frac{(\xi-\bar{x})^2}{2D}} (|i|+1)^2 \,\mathrm{d}\bar{x}\,\mathrm{d}\xi \\ &= \int_0^1 f_{\bar{X}}(\bar{x}) \sum_{i\in\mathbb{Z}\colon\, |i|>1} \int_{i-\bar{x}}^{i+1-\bar{x}} \frac{(|i|+1)^2}{\sqrt{2\pi D}} e^{-\frac{z^2}{2D}} \,\mathrm{d}z\,\mathrm{d}\bar{x}. \end{aligned} \tag{39}$$

For every $|i| = 2, 3, \ldots$, $z \in [i-\bar{x}, i+1-\bar{x})$, and $\bar{x} \in [0,1)$ we have $|i|+1 \leq 3|z|$. Hence,

$$\begin{aligned} &\sum_{i\in\mathbb{Z}\colon\, |i|>1} \int_{i-\bar{x}}^{i+1-\bar{x}} \frac{(|i|+1)^2}{\sqrt{2\pi D}} e^{-\frac{z^2}{2D}} \,\mathrm{d}z \\ &\quad \leq \sum_{i\in\mathbb{Z}\colon\, |i|>1} \int_{i-\bar{x}}^{i+1-\bar{x}} \frac{9z^2}{\sqrt{2\pi D}} e^{-\frac{z^2}{2D}} \,\mathrm{d}z \\ &\quad \leq 9 \int_{\{|z|\geq 1\}} \frac{z^2}{\sqrt{2\pi D}} e^{-\frac{z^2}{2D}} \,\mathrm{d}z \end{aligned} \tag{40}$$

where the last inequality follows because, for every $\bar{x} \in [0,1)$,

$$\bigcup_{i\in\mathbb{Z}\colon\, |i|>1} [i-\bar{x}, i+1-\bar{x}) \subseteq \{z \in \mathbb{R}\colon |z| \geq 1\}.$$

The RHS of (40) does not depend on $\bar{x}$, so together with (39) this yields

$$\sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i)(|i|+1)^2 \leq 9\mathsf{E}\big[Z_D^2 \mathbb{1}\{|Z_D| \geq 1\}\big]. \tag{41}$$

Writing $Z_D$ as $Z_D = \sqrt{D} Z_1$, where $Z_1$ is a zero-mean, unit-variance, Gaussian random variable, the expected value on the RHS of (41) can be written as

$$\mathsf{E}\big[Z_D^2 \mathbb{1}\{|Z_D| \geq 1\}\big] = D\mathsf{E}\big[Z_1^2 \mathbb{1}\big\{Z_1^2 \geq 1/D\big\}\big]. \tag{42}$$

Combining (42) and (41), we obtain

$$\begin{aligned} &\sum_{i\in\mathbb{Z}\colon\, |i|>1} \Pr(V_D = i) \frac{(|i|+1)^2}{2D} \\ &\qquad \leq \frac{9}{2} \mathsf{E}\big[Z_1^2 \mathbb{1}\big\{Z_1^2 \geq 1/D\big\}\big]. \end{aligned} \tag{43}$$

Since the function $z \mapsto z^2 \mathbb{1}\big\{z^2 \geq 1/D\big\}$ is dominated by $z \mapsto z^2$, and since $\mathsf{E}\big[Z_1^2\big] = 1$, it follows from the Dominated Convergence Theorem [20, Th. 1.6.9, p. 50] that

$$\lim_{D\downarrow 0} \mathsf{E}\big[Z_1^2 \mathbb{1}\big\{Z_1^2 \geq 1/D\big\}\big] = 0. \tag{44}$$

Together with (43) this demonstrates that the second term on the RHS of (35) vanishes as $D$ tends to zero.

Thus, (35), (37), (43), and (44) prove (32), which together with (30) and (31) demonstrates that

$$\lim_{D\downarrow 0} H(V_D) = \lim_{D\downarrow 0} \sum_{i\in\mathbb{Z}} \Pr(V_D = i) \log \frac{1}{\Pr(V_D = i)} = 0. \tag{45}$$

This was the last step required to prove Lemma 1. ∎

Combining Lemma 1 with (23) implies (17), which in turn demonstrates that the Shannon lower bound is asymptotically tight if $H(\lfloor X \rfloor) < \infty$ and $h(X) > -\infty$. This proves the first part of Theorem 2.

### B. Infinite Rate-Distortion Function

To prove that $H(\lfloor X \rfloor) = \infty$ implies $R(D) = \infty$ for every $D > 0$, we show that $I(X; \hat{X}) = \infty$ for every pair of random variables $(X, \hat{X})$ satisfying (3) and $H(\lfloor X \rfloor) = \infty$. To this end, we follow along the lines of the proof of Theorem 6 in [16, App. A]. Indeed, by the Data Processing Inequality [21, Cor 7.16],

$$I(X; \hat{X}) \geq I\big(\lfloor X \rfloor; \lfloor \hat{X} \rfloor\big). \tag{46}$$

The mutual information on the RHS of (46) can be written as

$$I\big(\lfloor X \rfloor; \lfloor \hat{X} \rfloor\big) = H\big(\lfloor X \rfloor\big) - H\big(\lfloor X \rfloor \,\big|\, \lfloor \hat{X} \rfloor\big). \tag{47}$$

Since $H(\lfloor X \rfloor) = \infty$ by assumption, the claim follows by showing that the conditional entropy on the RHS of (47) is bounded for every pair of random variables $(X, \hat{X})$ satisfying (3). Indeed, we have

$$\begin{aligned} H\big(\lfloor X \rfloor \,\big|\, \lfloor \hat{X} \rfloor\big) &\leq H\big(\lfloor X - \hat{X} \rfloor\big) \\ &\quad + H\big(\lfloor X \rfloor \,\big|\, \lfloor \hat{X} \rfloor, \lfloor X - \hat{X} \rfloor\big). \end{aligned} \tag{48}$$

Since $\mathsf{E}[\log(1 + |X - \hat{X}|)] < \infty$ for $(X, \hat{X})$ satisfying (3), Proposition 1 in [14] yields that

$$H\big(\lfloor X - \hat{X} \rfloor\big) < \infty. \tag{49}$$

Furthermore, denoting $Y = X - \hat{X}$, we obtain

$$\begin{aligned} H\big(\lfloor X \rfloor \,\big|\, \lfloor \hat{X} \rfloor, \lfloor X - \hat{X} \rfloor\big) &= H\big(\lfloor \hat{X} + Y \rfloor \,\big|\, \lfloor \hat{X} \rfloor, \lfloor Y \rfloor\big) \\ &\leq \log 2 \end{aligned} \tag{50}$$

since, conditioned on $\lfloor \hat{X} \rfloor$ and $\lfloor Y \rfloor$, the random variable $\lfloor \hat{X} + Y \rfloor$ can only take on the values $\lfloor \hat{X} \rfloor + \lfloor Y \rfloor$ or $\lfloor \hat{X} \rfloor + \lfloor Y \rfloor + 1$. Combining (48)–(50) yields

$$H\big(\lfloor X \rfloor \,\big|\, \lfloor \hat{X} \rfloor\big) < \infty. \tag{51}$$

Summing up, (46)–(51) demonstrate that $I(X; \hat{X}) = \infty$ for every pair of random variables $(X, \hat{X})$ satisfying (3) and $H(\lfloor X \rfloor) = \infty$. Hence, the rate-distortion function $R(D)$ is infinite for every finite $D$. This proves the second part of Theorem 2.

## IV. Conclusions

The Shannon lower bound is one of the few lower bounds on the rate-distortion function that hold for a large class of sources. We have demonstrated that this lower bound is asymptotically tight as the allowed distortion vanishes for all sources having a finite differential entropy and a finite Rényi information dimension. Conversely, we have demonstrated that if the source has an infinite Rényi information dimension, then its rate-distortion function is infinite for any finite distortion.

Assuming a finite Rényi information dimension is tantamount to assuming that quantizing the source with a uniform scalar quantizer of unit-length cells gives rise to a discrete random variable of finite entropy. The latter assumption is natural in rate-distortion theory and often encountered. To this effect, we have demonstrated that this assumption is not only natural, but it is also a necessary and sufficient condition for the asymptotic tightness of the Shannon lower bound.

Finally, the presented results can be generalized to $d$-dimensional, real-valued sources and distortion measures of the form $\|\mathbf{x} - \hat{\mathbf{x}}\|^r$, where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$ and $r > 0$. For details, see our paper [22] on arXiv.

## Acknowledgment

The author wishes to thank Helmut Bölcskei, David Stotz, and Gonzalo Vazquez-Vilar for helpful discussions. The author further wishes to thank Giuseppe Durisi and Tamás Linder for calling his attention to references [14] and [15], respectively.

## References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE International Convention Record*, vol. 7, pp. 142–163, 1959.

[2] T. Berger, *Rate Distortion Theory: Mathematical Basis for Data Compression*, ser. Electrical Engineering Series. Prentice Hall, 1971.

[3] Y. N. Linkov, "Evaluation of epsilon entropy of random variables for small epsilon," *Problemy Peredachi Informatsii (Problems of Inform. Transm.)*, vol. 1, pp. 12–18, 1965.

[4] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. 10, pp. 265–271, Oct. 1964.

[5] J. Binia, M. Zakai, and J. Ziv, "On the $\epsilon$-entropy and the rate-distortion function of certain non-Gaussian process," *IEEE Trans. Inform. Theory*, vol. 20, pp. 514–524, July 1974.

[6] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

[7] V. Kostina, "Data compression with low distortion and finite block-length," in *Proc. 53rd Allerton Conf. Comm., Contr. and Comp.*, Allerton H., Monticello, Il, Sep. 30 – Oct. 2, 2015.

[8] R. M. Gray, T. Linder, and J. Li, "A Langrangian formulation of Zador's entropy-constrained quantization theorem," *IEEE Trans. Inform. Theory*, vol. 28, no. 3, pp. 695–707, Mar. 2002.

[9] P. L. Zador, "Topics in the asymptotic quantization of continuous random variables," Bell Laboratories, Tech. Rep., 1966.

[10] T. Koch and G. Vazquez-Vilar, "Rate-distortion bounds for high-resolution vector quantization via Gibbs's inequality," July 2015. [Online]. Available: http://arxiv.org/abs/1507.08349

[11] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. 14, no. 5, pp. 676–683, Sept. 1968.

[12] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1–2, Mar. 1959.

[13] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inform. Theory*, vol. 40, no. 5, pp. 1564–1572, Sept. 1994.

[14] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inform. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[15] I. Csiszár, "Some remarks on the dimension and entropy of random variables," *Acta Mathematica Hungarica*, vol. 12, no. 3–4, pp. 399–408, Sept. 1961.

[16] D. Stotz and H. Bölcskei, "Degrees of freedom in vector interference channels," Sept. 26, 2014, subm. to *IEEE Trans. Inform. Theory*. [Online]. Available: www.nari.ee.ethz.ch/commth/pubs/p/dof_transit

[17] P. Billingsley, *Probability and Measure*, 3rd ed., ser. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, 1995.

[18] I. Csiszár, "Arbitrarily varying channel with general alphabets and states," *IEEE Trans. Inform. Theory*, vol. 38, no. 6, pp. 1725–1742, Nov. 1992.

[19] R. Durrett, *Probability: Theory and Examples*, 3rd ed. Brooks/Cole, 2005.

[20] R. B. Ash and C. A. Doléans-Dade, *Probability and Measure Theory*, 2nd ed. Elsevier/Academic Press, 2000.

[21] R. M. Gray, *Entropy and Information Theory*, 2nd ed. Springer Verlag, 2011.

[22] T. Koch, "The Shannon lower bound is asymptotically tight for sources with a finite Rényi information dimension," Apr. 2015. [Online]. Available: http://arxiv.org/abs/1504.08245

# Rate-Distortion of a Heegard-Berger Problem with Common Reconstruction Constraint

Meryem Benammar
Mathematics and Algorithmic Sciences Lab
Huawei Technologies France
Boulogne-Billancourt, 92100, France
Email: meryem.benammar@huawei.com

Abdellatif Zaidi
Mathematics and Algorithmic Sciences Lab
Huawei Technologies France
Boulogne-Billancourt, 92100, France
Email: abdellatif.zaidi@huawei.com

***Abstract*—In this work, we establish the rate-distortion function $R(D_1, D_2)$ of a Heegard-Berger Problem with two sources, $S_1$ and $S_2$, under the assumption of *degraded* reconstruction sets and *common encoder-receivers* reconstruction. Specifically, the source $S_1$ needs to be recovered at only Decoder 1, to within some prescribed average distortion level $D_1$; and the source $S_2$ needs to be recovered at both decoders, to within some same prescribed average distortion level $D_2$. In addition, the encoder and decoders must agree on a *common* reconstruction version of $S_2$.**

## I. Introduction

The Heegard Berger problem [1] is one of the most crucial extensions of Wyner and Ziv's result, on lossy source coding with side information, to multi-terminal scenarios. In such a setting, a memoryless source $S^n$ has to be reconstructed at two decoders 1 and 2, respectively to within prescribed distortion levels $D_1$ and $D_2$ – Decoder 1 has access to side information $Y_1^n$ and Decoder 2 has access to side information $Y_2^n$. Heegard and Berger derived in [1] an inner bound on the rate-distortion function $R(D_1, D_2)$ of the model that remains the best bound known to date for the two-user case. This inner bound proves to be optimal for many a setting, among which, *degraded* side information and the wider class of *conditionally less-noisy* side information as summarized by Timo *et al.* in [2], as well as for some settings where the side information sequences are not *ordered*, e.g. the *complementary delivery* investigated by Kimura *et al.* in [3], and the *unmatched product of two degraded sources* investigated by Watanabe in [4].

In this work, we investigate a wider class of Heegard-Berger problems in which we assume that the source is composed of two sources $S_1^n$ and $S_2^n$ arbitrarily correlated between them, and to the side information sequences $Y_1^n$ and $Y_2^n$. The source sequence $S_1^n$ has to be reconstructed at only receiver 1 to with some prescribed distortion $D_1$, while the source $S_2^n$ has to be reconstructed at both receivers. As such, we impose a *degraded reconstruction set* but we do, *by no means*, impose any sort of hierarchy or ordering on the side information sequences $Y_1^n$ and $Y_2^n$. In previous works [5] and [6], the authors derived the optimal rate-distortion function $R(D_1)$ where $S_2^n$ had to be recovered losslessly at both decoders. In this work, we investigate the more general case where we allow the source $S_2^n$ to be reconstructed to within a certain distortion $D_2$, but impose however that the reconstructions of $S_2^n$ at all terminals, including the source, be almost equal. This model, as depicted in Figure 1, is useful for applications in which the source component $S_2^n$ represents some critical information, such as sensitive medical information, and each of the sender and the receivers need to share a common compressed version of it. Such a *common reconstruction* constraint was first investigated by Steinberg in [7] for a Wyner-Ziv setting and it was shown that, under a *common source-receiver* reconstruction constraint, the use of side information was prevented in the *estimation phase* and allowed only in the *binning phase*, as elaborated on in Section II-A. Yet, the utility of side information is less easy to understand in multi-terminal settings, such as the Heegard-Berger problem with common reconstruction constraints. Indeed, as an intuitive result for the Heegard-Berger problem with a common *source-receivers* reconstruction constraint as in Figure 2, [8] shows that side information is only useful for binning since it is not known to the source and can be of no use for estimation. However, for the Heegard-Berger problem with *common receivers reconstruction only* shown in Figure 3, Vellambi and Timo [8] observed that the *common* part of the two side information sequences can still be used at the estimation phase. ( The reader may also refer to the related work in [9] where Ahmadi *et al.* investigate a Heegard-Berger problem with degraded side information in which the encoder is constrained to be able to produce each of the receivers reconstructions, without imposing that these

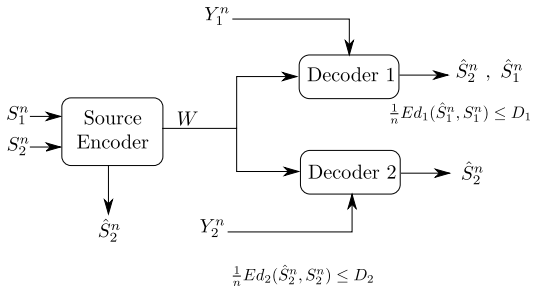



Fig. 1. Heegard-Berger problem with two sources, side information and degraded reconstruction sets.

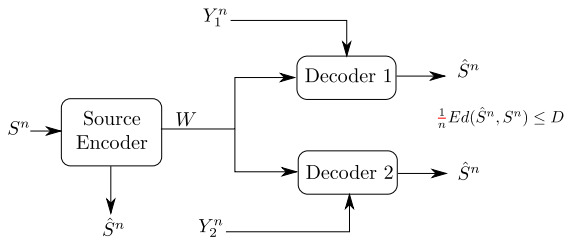



Fig. 2. Heegard-Berger problem with common source-receiver reconstruction

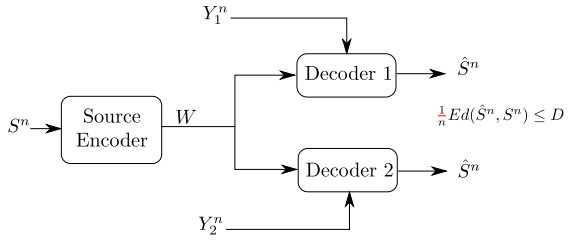



Fig. 3. Heegard-Berger problem with common receiver reconstruction only

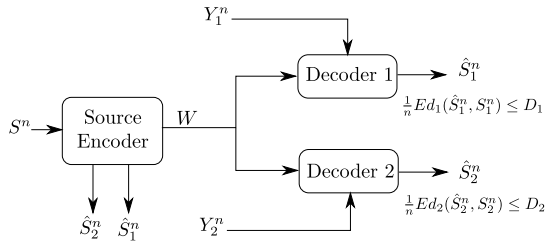



Fig. 4. Heegard-Berger problem with distinct common source-receiver reconstruction

decoders' reconstructions be identical to each other with high probability – see Figure 4, or to [10] where Timo *et al.* investigate the Heegard-Berger problem with complimentary delivery and common receivers' reconstructions.)

In this work, we assume unlike previous works that we are in the presence of *two sources*, and impose a common *source-receiver* reconstruction constraints for source $S_2^n$. Our contribution is outlined as follows. In section II, we give necessary definitions and briefly review and comment on some related results on the role of side information for binning and/or estimation. The main result, as stated in Section III, is a full characterization of the rate-distortion function $R(D_1, D_2)$ of the model of Figure 1. Finally, Section IV is dedicated to the proof of the main result.

*Notations*

We use the following notations. Upper case letters are used to denote random variables, e.g., $S$; lower case letters are used to denote realizations of random variables, e.g., $s$; and calligraphic letters designate alphabets, i.e., $\mathcal{S}$. $\mathcal{S}$ denotes the cardinality of a set $\mathcal{X}$. For random variables $X$, $Y$ and $Z$, the notation $X \multimap Y \multimap Z$ indicates that $X$, $Y$ and $Z$, in this order, form a Markov Chain. For integers $i \leq j$, we define $[i:j] := \{i, i+1, \ldots, j\}$. p.m.f stands for probability mass function while a.r.v. stands for auxiliary random variable.

## II. Problem Setup and Definitions

Let $(\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{Y}_1 \times \mathcal{Y}_2, P_{S_1,S_2,Y_1,Y_2})$ be a discrete memoryless four-source with generic variables $S_1$, $S_2$, $Y_1$ and $Y_2$. Also, let $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ be two reconstruction alphabets and, for $i \in \{1,2\}$, $d_i$ a distortion measure defined as

$$\begin{array}{rccl} d_i & : & \mathcal{S}_i \times \hat{\mathcal{S}}_i & \to & \mathbb{R}_+ \\ & & (s_i, \hat{s}_i) & \to & d_i(s_i, \hat{s}_i) \ . \end{array} \tag{1}$$

An $(n, M_n, D_1, D_2)$ code for the lossy Heegard-Berger problem with degraded reconstruction sets and common reconstruction consists of:

- A set of messages $\mathcal{W} \triangleq [1 : M_n]$.
- An encoding function $f$ such that:

$$\begin{array}{rccl} f & : & \mathcal{S}_1^n \times \mathcal{S}_2^n & \to & \mathcal{W} \\ & & (S_1^n, S_2^n) & \to & W = f(S_1^n, S_2^n) \ . \end{array} \tag{2}$$

- Two decoding functions $g_1$ and $g_2$, one at each user:

$$\begin{array}{rccl} g_1 & : & \mathcal{W} \times \mathcal{Y}_1^n & \to & \hat{\mathcal{S}}_2^n \times \hat{\mathcal{S}}_1^n \\ & & (W, Y_1^n) & \to & (\hat{S}_{2,1}^n, \hat{S}_1^n) = g_1(W, Y_1^n) \ , \end{array} \tag{3}$$

and

$$\begin{array}{rccl} g_2 & : & \mathcal{W} \times \mathcal{Y}_2^n & \to & \hat{\mathcal{S}}_2^n \\ & & (W, Y_2^n) & \to & \hat{S}_{2,2}^n = g_2(W, Y_2^n) \ . \end{array} \tag{4}$$

- An additional encoder reconstruction function $g_s$ defined by

$$\begin{array}{rccl} g_s & : & \mathcal{W} & \to & \hat{\mathcal{S}}_2^n \\ & & W & \to & \hat{S}_{2,s}^n = g_s(W). \end{array} \tag{5}$$

The expected distortions of this code are given by

$$\mathbb{E}\left(d_1(S_1^n, \hat{S}_1^n)\right) \triangleq \mathbb{E}\frac{1}{n}\sum_{i=1}^n d_1(S_{1,i}, \hat{S}_{1,i}) \tag{6}$$

$$\mathbb{E}\left(d_2(S_2^n, \hat{S}_{2,j}^n)\right) \triangleq \mathbb{E}\frac{1}{n}\sum_{i=1}^n d_2(S_{2,i}, \hat{S}_{2,ji}), \text{ for } j = 1, 2 \tag{7}$$

The probability of error of this code is given by

$$P_e^{(n)} \triangleq \mathbb{P}\big(\hat{S}_{2,1}^n \neq \hat{S}_{2,s}^n \text{ or } \hat{S}_{2,2}^n \neq \hat{S}_{2,s}^n\big) \ . \tag{8}$$

*N.B: Imposing that the probability of error of the source $S_2$'s reconstructions be arbitrarily small implies that the three terminals, i.e. encoder and receivers, can not have distinct distortion levels and thus, we only impose one common distortion level for the source $S_2$.*

**Definition 1.** *A rate $R$ is said to be $(D_1, D_2)$-achievable for the lossy HB problem with degraded reconstruction sets and common source-receivers reconstruction if there exists a sequence of codes $(n, M_n, D_1, D_2)$ such that:*

$$\limsup_{n\to\infty} P_e^{(n)} \ = \ 0 \ , \tag{9}$$

$$\limsup_{n\to\infty} \mathbb{E}\left(d_1(S_1^n, \hat{S}_1^n)\right) \ \leq \ D_1 \ , \tag{10}$$

$$\limsup_{n\to\infty} \mathbb{E}\left(d_2(S_2^n, \hat{S}_{2,s}^n)\right) \ \leq \ D_2 \ , \tag{11}$$

$$\liminf_{n\to\infty} \log_2(M_n) \quad \geq \quad R \ . \tag{12}$$

*The rate-distortion $R(D_1, D_2)$ of this problem is defined by*

$$R(D_1, D_2) \triangleq \inf\{R : R \text{ is } (D_1, D_2)\text{-achievable}\} \ . \tag{13}$$

### A. Role of Side Information, Binning and/or Estimation

In source coding problems with side information at the decoder, side information may be utilized for binning and/or estimation, depending on the configuration. For example, in the standard Wyner-Ziv setup [11] with source $S$ and arbitrarily correlated side information $Y$ available non-causally only at the decoder, the side information is utilized both for binning and for the estimation of the reconstruction. This is reflected through the following rate-distortion function,

$$R_{\text{WZ}} = \min_{P_{V|SY}} I(V; S|Y) \tag{14}$$

where $P_{VSY}$ is such that:

$$1) \quad V \multimap S \multimap Y \tag{15}$$

$$2) \quad \exists \quad \hat{S} \triangleq \phi(V; Y) \quad \text{s.t} \quad \mathbb{E}d(S, \hat{S}) \leq D. \tag{16}$$

in which it is clear that the rate $I(V; S|Y) = I(V; S) - I(V; Y)$ takes benefit from the binning against side information, while the reconstructed sequence at the estimation is allowed to depend on the side information sequence through the function $\phi(\cdot)$.

The side information plays a similar role, but with a generally stronger binning leading to a better rate, if it is also given to the encoder (i.e., the standard conditional rate-distortion problem [12]); and it plays a less important role (only estimation) if it is given only causally to the decoder but not to the encoder [13]. If the encoder is constrained to produce an exact copy of the decoder's reconstruction, referred to as *common source-receiver reconstruction constraint* in [7], side information can be used for binning but not for estimation – as otherwise, the encoder, which does not know the side information, can not estimate the decoder's reconstruction. This is reflected through the associated rate-distortion function, which in this case reduces to [7]

$$R_{\text{CR}} = \min_{P_{\hat{S}|SY}} I(\hat{S}; S|Y) \tag{17}$$

where $P_{\hat{S}SY}$ is such that:

$$1) \quad \hat{S} \multimap S \multimap Y \tag{18}$$

$$2) \quad \mathbb{E}d(S, \hat{S}) \leq D \tag{19}$$

It is clear from the above equation that the reconstruction $\hat{S}$ can not depend on the side information sequence $Y$ due to the Markov Chain $\hat{S} \multimap S \multimap Y$ while the transmission rate still from the binning the binning.

## III. Rate-Distortion Function

Recall the definitions of section II. The following theorem characterizes the rate-distortion function of the Heegard-Berger model with degraded reconstruction sets and common reconstruction shown in Figure 1.

**Theorem 1** (The rate-distortion function)**.** *The rate-distortion function $R(D_1, D_2)$ of the Heegard-Berger model with degraded reconstruction sets and common reconstruction shown in Figure 1 is given by*

$$R(D_1, D_2) = \min_{\mathcal{P}} \max\big\{I(U_0\hat{S}_2; S_1S_2|Y_1),\ I(U_0\hat{S}_2; S_1S_2|Y_2)\big\} + I(U_1; S_1S_2|Y_1\hat{S}_2U_0) \tag{20}$$

*where the minimization is over the set $\mathcal{P}$ of joint conditional p.m.fs $P_{U_0U_1\hat{S}_2|S_1S_2}$ that satisfy i), ii), and iii) where:*
*i) $(U_0, U_1, \hat{S}_2) \multimap (S_1, S_2) \multimap (Y_1, Y_2)$ forms a Markov chain,*
*ii) there exists a function $\Phi$ such that:*

$$\begin{aligned} \Phi \quad : \quad & \mathcal{U}_0 \times \mathcal{U}_1 \times \hat{\mathcal{S}}_2 \times \mathcal{Y}_1 \to \hat{\mathcal{S}}_1 \\ & (U_0, U_1, \hat{S}_2, Y_1) \to \hat{S}_1 = \Phi(U_0, U_1, \hat{S}_2, Y_1) \end{aligned}$$

*and $\mathbb{E}(d_1(S_1, \hat{S}_1)) \leq D_1$ .*
*iii) The distortion constraint of source $S_2$ is such that*

$$\mathbb{E}(d_2(S_2, \hat{S}_2)) \leq D_2. \tag{21}$$

**Proof:** The proof of Theorem 1 is given in Appendix IV.

In the following remarks, we elaborate more on Theorem 1 and its connection to related results.

**Remark 1.** *The result of Theorem 1 can be seen as a generalization of that of [5, Theorem 1], in the sense that setting $D_2 = 0$ in Theorem 1, one recovers [5, Theorem 1].*

*Also, a characterization of the rate-distortion function for the model of Figure 2 can be readily obtained from the result of Theorem 1 by setting $S_1 = \emptyset$. In this sense, Theorem 1 can also be seen as a generalization of [8, Theorem 1] to the case in which one of the decoders also recovers an individual description.*

**Remark 2.** *The coding scheme that we use for the proof of achievability of Theorem 1 requires that for the encoding and decoding of the source component $s_2^n$, the side information sequences $y_1^n$ and $y_2^n$ are used for the binning stage, but not for the estimation stage. However, they are used for both binning and estimation for the encoding/decoding of the source component $s_1^n$.*

## IV. Proof of Theorem 1

### A. Proof of converse

Let $R$ be a $(D_1, D_2)$-achievable rate for our Heegard-Berger problem with degraded reconstruction sets and common reconstruction of Figure 1. Let $\Phi$ be the associated encoding function, and $g_1$, $g_2$, and $g_s$ the associated reconstruction functions. That is, $W = \Phi(S_1^n, S_2^n)$, $\hat{S}_1^n = g(W, Y_1^n)$, $\hat{S}_{2,1}^n = g_1(W, Y_1^n)$ and $\hat{S}_{2,2}^n = g_2(W, Y_2^n)$ with

$$P_e^{(n)} \triangleq \mathbb{P}\big(\hat{S}_{2,1}^n \neq \hat{S}_{2,s}^n \text{ or } \hat{S}_{2,2}^n \neq \hat{S}_{2,s}^n\big) \leq \epsilon_n. \tag{22}$$

First, note that the imposed common encoder-decoders reconstruction constraint for the source component $s_2^n$ implies the following Fano's inequalities,

$$\frac{1}{n} H(\hat{S}_{2,s}^n|\hat{S}_{2,1}^n) \leq \frac{1}{n} + \log_2(||\hat{\mathcal{S}}_2||) P_e^{(n)} \tag{23}$$

$$\frac{1}{n}H(\hat{S}_{2,s}^n|\hat{S}_{2,2}^n) \le \frac{1}{n} + \log_2(||\hat{\mathcal{S}}_2||)P_e^{(n)}. \quad (24)$$

Combining (22) and (24), one gets that

$$\frac{1}{n}H(\hat{S}_{2,s}^n|\hat{S}_{2,1}^n) \le \epsilon_n' \quad \text{and} \quad \frac{1}{n}H(\hat{S}_{2,s}^n|\hat{S}_{2,2}^n) \le \epsilon_n' \ , \quad (25)$$

where $\lim_{n\Rightarrow\infty}\epsilon_n' = 0$.
Next, for the first constraint on the rate $R$, we have

$$\begin{aligned} nR &\ge H(W|Y_1^n) && (26)\\ &\ge I(W; S_1^n S_2^n|Y_1^n) && (27)\\ &= I(W\hat{S}_{2,1}^n; S_1^n S_2^n|Y_1^n) && (28)\\ &\overset{(a)}{\ge} I(W\hat{S}_{2,1}^n\hat{S}_{2,s}^n; S_1^n S_2^n|Y_1^n) - 2n\epsilon_n && (29)\\ &\ge I(W\hat{S}_{2,s}^n; S_1^n S_2^n|Y_1^n) - 2n\epsilon_n && (30)\end{aligned}$$

Note then that:

$$\begin{aligned} & I(W\hat{S}_{2,s}^n; S_1^n S_2^n|Y_1^n)\\ &= \sum_{i=1}^n I(W\hat{S}_{2,s}^n; S_{1,i}S_{2,i}|S_1^{i-1}S_2^{i-1}Y_{1,i}Y_1^{i-1}Y_{1,i-1}^n) && (31)\\ &= \sum_{i=1}^n I(W\hat{S}_{2,s}^n Y_1^{i-1}Y_{1,i+1}^n S_1^{i-1}S_2^{i-1}; S_{1,i}S_{2,i}|Y_{1,i}) && (32)\\ &\overset{(b)}{=} \sum_{i=1}^n I(W\hat{S}_{2,s}^n Y_1^{i-1}Y_{1,i+1}^n S_1^{i-1}S_2^{i-1}Y_2^{i-1}; S_{1,i}S_{2,i}|Y_{1,i})\\ &\ge \sum_{i=1}^n I(W\hat{S}_{2,s}^{i-1}\hat{S}_{2,s,i+1}^n Y_1^{i-1}Y_{1,i+1}^n Y_2^{i-1}\hat{S}_{2,s,i}; S_{1,i}S_{2,i}|Y_{1,i})\end{aligned}$$

where $(a)$ holds using (25) and $(b)$ holds using the following Markov chain the justification of which will follow,

$$Y_2^{i-1} \multimap (W, \hat{S}_{2,s}^n, Y_1^{i-1}, Y_{1,i+1}^n, S_1^{i-1}, S_2^{i-1}, Y_{1,i}) \multimap (S_{1,i}, S_{2,i}). \quad (33)$$

At this stage, we pause to justify (33). We have the following list of Markov chains and implications,

$$\begin{aligned} &(a)\,(Y_2^{i-1}, Y_1^{i-1}, S_1^{i-1}, S_2^{i-1})\\ &\qquad \multimap Y_{1,i} \multimap (Y_{1,i+1}^n, S_{1,i+1}^n, S_{2,i+1}^n, S_{1,i}, S_{2,i})\\ &\Rightarrow Y_2^{i-1} \multimap (Y_{1,i}, Y_1^{i-1}, S_1^{i-1}, S_2^{i-1})\\ &\qquad \multimap (Y_{1,i+1}^n, S_{1,i+1}^n, S_{2,i+1}^n, S_{1,i}, S_{2,i}) && (34)\\ &\overset{(b)}{\Rightarrow} Y_2^{i-1} \multimap (Y_{1,i}, Y_1^{i-1}, S_1^{i-1}, S_2^{i-1})\\ &\qquad \multimap (W, \hat{S}_{2,s}^n, Y_{1,i+1}^n, S_{1,i+1}^n, S_{2,i+1}^n, S_{1,i}, S_{2,i}) && (35)\\ &\Rightarrow Y_2^{i-1} \multimap (Y_{1,i}, Y_1^{i-1}, S_1^{i-1}, S_2^{i-1})\\ &\qquad \multimap (W, \hat{S}_{2,s}^n, Y_{1,i+1}^n, S_{1,i}, S_{2,i}) && (36)\\ &\Rightarrow Y_2^{i-1} \multimap (Y_{1,i}, W, \hat{S}_{2,s}^n, Y_1^{i-1}, Y_{1,i+1}^n, S_1^{i-1}, S_2^{i-1})\\ &\qquad \multimap (S_{1,i}, S_{2,i})\ . && (37)\end{aligned}$$

where $(a)$ holds since the source is memoryless and $(b)$ holds since $W$, and so $\hat{S}_{2,s}^n$, are deterministic functions of $(S_1^n, S_2^n)$. Defining, for $i \in [1:n]$, the auxiliary random variables $U_{0,i} = W\hat{S}_{2,s}^{i-1}\hat{S}_{2,s,i+1}^n Y_2^{i-1}Y_{1,i+1}^n$ and $U_{1,i} = (U_{0,i}Y_1^{i-1})$, the inequality (33) given

$$nR \ge \sum_{i=1}^n I(U_{0,i}U_{1,i}\hat{S}_{2,s,i}; S_{1,i}S_{2,i}|Y_{1,i}) - 2n\epsilon_n. \quad (38)$$

For the second constraint on the rate $R$, we have

$$\begin{aligned} nR &\ge H(W|Y_2^n) && (39)\\ &\ge I(W; S_1^n S_2^n|Y_2^n) && (40)\\ &= I(W\hat{S}_{2,2}^n; S_1^n S_2^n|Y_2^n) && (41)\\ &\ge I(W\hat{S}_{2,2}^n\hat{S}_{2,s}^n; S_1^n S_2^n|Y_2^n) - 2n\epsilon_n && (42)\\ &\ge I(W\hat{S}_{2,s}^n; S_1^n S_2^n|Y_2^n) - 2n\epsilon_n && (43)\\ &= H(S_1^n S_2^n|Y_2^n) - H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_2^n) - 2n\epsilon_n && (44)\\ &= H(S_1^n S_2^n|Y_2^n) - H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_2^n) - 2n\epsilon_n\\ &\quad + H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_1^n) - H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_2^n). && (45)\end{aligned}$$

The term $[H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_1^n) - H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_1^n)]$ on the RHS of (45) can be written as

$$\begin{aligned} & H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_1^n) - H(S_1^n S_2^n|W\hat{S}_{2,s}^n Y_2^n)\\ &= I(S_1^n S_2^n; Y_2^n|W\hat{S}_{2,s}^n) - I(S_1^n S_2^n; Y_1^n|W\hat{S}_{2,s}^n) && (46)\\ &\overset{(a)}{=} \sum_{i=1}^n \big[I(S_1^n S_2^n; Y_{2,i}|W\hat{S}_{2,s}^n Y_2^{i-1}Y_{1,i+1}^n)\\ &\qquad - I(S_1^n S_2^n; Y_{1,i}|W\hat{S}_{2,s}^n Y_2^{i-1}Y_{1,i+1}^n)\big] && (47)\\ &\overset{(b)}{=} \sum_{i=1}^n \big[I(S_{1,i}S_{2,i}; Y_{2,i}|W\hat{S}_{2,s}^n Y_2^{i-1}Y_{1,i+1}^n)\\ &\qquad - I(S_{1,i}S_{2,i}; Y_{1,i}|W\hat{S}_{2,s}^n Y_2^{i-1}Y_{1,i+1}^n)\big] && (48)\\ &= \sum_{i=1}^n \big[I(S_{1,i}S_{2,i}; Y_{2,i}|U_{0,i}\hat{S}_{2,s,i})\\ &\qquad - I(S_{1,i}S_{2,i}; Y_{1,i}|U_{0,i}\hat{S}_{2,s,i})\big] && (49)\\ &= \sum_{i=1}^n \big[H(S_{1,i}S_{2,i}|U_{0,i}\hat{S}_{2,s,i}Y_{1,i})\\ &\qquad - H(S_{1,i}S_{2,i}|U_{0,i}\hat{S}_{2,s,i}Y_{2,i})\big] && (50)\end{aligned}$$

where $(a)$ follows using Csiszár-Körner sum identity, applied twice, and $(b)$ holds since the following is a Markov chain, the justification of which will follow,

$$\begin{aligned} &(S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1}, S_{2,i+1}^n) \multimap (U_{0,i}, \hat{S}_{2,s,i}, S_{1,i}, S_{2,i})\\ &\qquad \multimap (Y_{1,i}, Y_{2,i}). && (51)\end{aligned}$$

We pause to justify (51). This is obtained using the following easy Markov chains and implications,

$$\begin{aligned} &(Y_2^{i-1}, Y_{1,i+1}^n, S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1}, S_{2,i+1}^n)\\ &\qquad \multimap (S_{1,i}, S_{2,i}) \multimap (Y_{1,i}, Y_{2,i}) && (52)\\ &\overset{(c)}{\Rightarrow} (W, \hat{S}_{2,s}^n, Y_2^{i-1}, Y_{1,i+1}^n, S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1}S_{2,i+1}^n)\\ &\qquad \multimap (S_{1,i}, S_{2,i}) \multimap (Y_{1,i}, Y_{2,i}) && (53)\\ &\Rightarrow (U_{0,i}, \hat{S}_{2,s,i}, S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1}, S_{2,i+1})\\ &\qquad \multimap (S_{1,i}, S_{2,i}) \multimap (Y_{1,i}, Y_{2,i}) && (54)\\ &\Rightarrow (S_1^{i-1}, S_{1,i+1}^n, S_2^{i-1}, S_{2,i+1})\\ &\qquad \multimap (U_{0,i}, \hat{S}_{2,s,i}, S_{1,i}, S_{2,i}) \multimap (Y_{1,i}, Y_{2,i}) && (55)\end{aligned}$$

and where $(c)$ follows since $W$, and so $\hat{S}_{2,s}^n$, are deterministic functions of $(S_1^n, S_2^n)$.

Finally, we terminate the proof of converse of Theorem 1 by noticing that the reconstruction $\hat{S}_{1,i} = g(W, Y_1^n)$ clearly satisfies $\hat{S}_{1,i} = g'(U_{0,i}, U_{1,i}, Y_{1,i})$ for some function $g'$. □

*B. Proof of achievability*

The proof of achievability of Theorem 1 is as follows. First, we show the achievability of the following rate distortion function,

$$R(D_1, D_2) = \min \max\{I(U; S_1S_2|Y_1),\ I(U; S_1S_2|Y_2)\} + I(U_1; S_1S_2|Y_1U) \tag{56}$$

where the minimization is over all conditionals $P_{UU_1|S_1S_2}$ satisfying that $(U, U_1) \multimap (S_1, S_2) \multimap (Y_1, Y_2)$ is a Markov chain and there exist functions $\phi$ and $\psi$ such that:

$$\mathbb{E}d_1(S_1, \hat{S}_1) \leq D_1 \mathbb{E}d_2(S_2, \hat{S}_2) \leq D_2 \tag{57}$$

where

$$\hat{S}_2 = \psi(U) \text{ and} \hat{S}_1 = \phi(Y_1, U, U_1). \tag{58}$$

Next, we evaluate the above region with the choice $U = (U_0, \hat{S}_2)$ to recover the result of Theorem 1.

*Codebook generation:*

Generate $2^{nR_0}$ sequences $u^n(w_0)$ following $\prod_{i=1}^n P_U$ where $w_0 \in [1 : 2^{nR_0}]$ and set them in $2^{nR_0'}$ bins $\mathcal{B}^n(w_0')$ with $w_0' \in [1 : 2^{nR_0'}]$. Then, for each $w_0$, generate $2^{nR_1}$ sequences $u_1^n(w_0, w_1)$ following $\prod_{i=1}^n P_{U_1|U}$ with $w_1 \in [1 : 2^{nR_1}]$ and set them in $2^{nR_1'}$ bins $\mathcal{B}^n(w_0', w_1')$ with $w_1' \in [1 : 2^{nR_1'}]$. Also, fix a reconstruction function $\psi$ such that: $\mathbb{E}d_2(S_2, \psi(U)) \leq D_2$ and denote $\psi^n$ its n-letter extension such that $\hat{s}_{2,s}^n = \psi^n(u^n(w_0))$.

*Encoding:*

Upon observing $S_1^n$ and $S_2^n$, find an index $w_0 \in [1 : 2^{nR_0}]$ such that:

$$(u^n(w_0), s_2^n, s_1^n) \in \mathcal{T}_{[US_2S_1]}^{(n)}\ , \tag{59}$$

and an index $w_1 \in [1 : 2^{nR_1}]$ such that:

$$(u_1^n(w_0, w_1), u^n(w_0), s_2^n, s_1^n) \in \mathcal{T}_{[UU_1S_2S_1]}^{(n)}\ . \tag{60}$$

The encoder transmits $w_0'$ and $w_1'$,i.e., the indices of the bins in which $u^n$ and $u_1^n$ lie. This encoding step has small error as long as $n$ is large and

$$R_0 \geq I(U; S_1S_2)\ , \tag{61}$$

$$R_1 \geq I(U_1; S_1S_2|U) \tag{62}$$

*Decoding:*

Decoder 2 reconstructs the sequences $\hat{s}_2^n$. To this end, it first looks for the unique sequence $u^n(w_0) \in \mathcal{B}^n(w_0')$ such that

$$(u^n(w_0), y_2^n) \in \mathcal{T}_{[UY_2]}^{(n)}\ . \tag{63}$$

Then, it sets $\hat{s}_{2,2}^n = \psi^n(u^n(w_0))$ as the final reconstruction. (Note here that the reconstruction $\hat{S}_2$ can not depend on the available side information sequence).

The error in this decoding step can be made arbitrarily small as long as $n$ is large and

$$R_0 - R_0' \leq I(U; Y_2)\ , \tag{64}$$

Decoder 1 looks for a sequence $u^n \in \mathcal{B}^n(w_0')$ such that:

$$(u^n(w_0), y_1^n) \in \mathcal{T}_{[UY_1]}^{(n)}\ , \tag{65}$$

and then looks for a sequence $u_1^n \in \mathcal{B}^n(w_0', w_1')$ verifying:

$$(u^n(w_0), u_1^n(w_0, w_1), y_1^n) \in \mathcal{T}_{[UU_1Y_1]}^{(n)}\ , \tag{66}$$

Then, from $(u^n(w_0), u_1^n(w_0, w_1), y_1^n)$ the decoder can recover the reconstruction sequences $\hat{s}_{2,1}^n = \psi^n(u^n(w_0))$ and $\hat{s}_1^n = \phi(u^n(w_0), u_1^n(w_0, w_1), y_1^n)$.

Similarly, the error in this decoding step can be made arbitrarily small as long as $n$ is large and

$$R_0 - R_0' \leq I(U; Y_1)\ , \tag{67}$$

$$R_1 - R_1' \leq I(U_1; Y_1|U)\ . \tag{68}$$

Note that, in this scheme, if the encoding and decoding steps are performed correctly are successful, then all terminals can reconstruct $\hat{s}_2^n$ with essentially no error.

The rest of the proof follows by a standard application of FME to eliminate $R_0$, $R_1$ and $R_2$ from inequalities (61) – (68) and substituting $R = R_2' + R_0' + R_1'$ to get the desired result. □


## ACKNOWLEDGEMENT

The authors would like to thank Roy Timo for the helpful discussions about the model studied in this paper.


## REFERENCES


[1] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *Information Theory, IEEE Transactions on*, vol. 31, no. 6, pp. 727–734, 1985.

[2] R. Timo, T. Oechtering, and M. Wigger, "Source coding problems with conditionally less noisy side information," *Information Theory, IEEE Transactions on*, vol. 60, no. 9, pp. 5516–5532, 2014.

[3] A. Kimura and T. Uyematsu, "Multiterminal source coding with complementary delivery," *arXiv preprint arXiv:0804.1602*, 2008.

[4] S. Watanabe, "The rate-distortion function for product of two sources with side-information at decoders," *Information Theory, IEEE Transactions on*, vol. 59, no. 9, pp. 5678–5691, 2013.

[5] M. Benammar, A. Zaidi, "The rate distortion function of a heegard-berger problem with two sources and degraded reconstruction sets," in *In Proceedings, IEEE ITW Jeju Island,*, 2015.

[6] ——, "Lossless source coding for a heegard-berger problem with two sources and degraded reconstruction sets," in *In Proceedings, ISWCS Brussels,*, 2015.

[7] Y. Steinberg, "Coding and common reconstruction," *Information Theory, IEEE Transactions on*, vol. 55, no. 11, pp. 4995–5010, 2009.

[8] B. N. Vellambi and R. Timo, "The Heegard-Berger problem with common receiver reconstructions," in *Information Theory Workshop (ITW), 2013 IEEE*. IEEE, 2013, pp. 1–5.

[9] B. Ahmadi, R. Tandon, O. Simeone, and H. V. Poor, "Heegard-Berger and cascade source coding problems with common reconstruction constraints," *Information Theory, IEEE Transactions on*, vol. 59, no. 3, pp. 1458–1474, 2013.

[10] R. Timo, A. Grant, and G. Kramer, "Rate-distortion functions for source coding with complementary side information," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 2934–2938.

[11] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 22, pp. 1–10, 1976.

[12] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971. [Online]. Available: https://books.google.co.uk/books?id=-HV1QgAACAAJ

[13] T. Weissman and A. El Gamal, "Source coding with limited-look-ahead side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5218–5239, 2006.

# On the Burst Erasure Correctability of Spatially Coupled LDPC Ensembles

Narayanan Rengaswamy
Department of Electrical and Computer Engineering
Texas A&M University
College Station, Texas 77840, USA
Email: r_narayanan_92@tamu.edu

Laurent Schmalen and Vahid Aref
Alcatel-Lucent Bell Labs
Stuttgart 70435, Germany
Email: {laurent.schmalen,vahid.aref}@alcatel-lucent.com

***Abstract*—Spatially-Coupled LDPC (SC-LDPC) ensembles achieve the capacity of binary memoryless channels, asymptotically, under belief-propagation decoding. In this work, we are interested in the finite-length performance of these ensembles on binary channels with memory. We study the average performance of random regular SC-LDPC ensembles on single-burst-erasure channels and provide tight bounds for the block erasure probability. Further, we show the effect of expurgation on the performance by analyzing the minimal stopping sets.**

## I. Introduction

Low-density parity-check (LDPC) codes are widely used due to their outstanding performance under low-complexity belief propagation (BP) decoding. However, an error probability exceeding that of maximum-a-posteriori (MAP) decoding has to be tolerated with (sub-optimal) BP decoding. Recently, it has been empirically observed for spatially coupled LDPC (SC-LDPC) codes—first introduced by Jiminez Felström and Zigangirov as convolutional LDPC codes [1]—that the BP performance of these codes can improve dramatically towards the MAP performance of the underlying LDPC code under many different settings and conditions, e.g. [2]. This phenomenon, termed *threshold saturation*, has been proven rigorously in [3], [4]. In particular, the BP threshold of a coupled LDPC ensemble tends to its MAP threshold on any binary memoryless symmetric channel (BMS).

Besides their excellent performance on the BEC and AWGN channels, much less is known about the burst error correctability of SC-LDPC codes. In [5], the authors consider SC-LDPC ensembles over a block erasure channel (BLEC) where the channel erases complete spatial positions instead of individual bits. This block erasure model mimics block-fading channels frequently occurring in wireless communications. The authors give asymptotic lower and upper bounds for the bit and block erasure probabilities obtained from density evolution. In [6], the authors construct protograph-based codes that maximize the correctable burst lengths, while the authors in [7] apply interleaving (therein denoted band splitting) to a protograph-based SC-LDPC code to increase the correctable burst length. If windowed decoding is used, this approach results in an increased required window length and thus complexity. Recently, it has been shown that protograph-based LDPC codes can increase the diversity order of block fading channels and are thus good candidates for block erasure channels [8], [9]; however, they require large syndrome former memories if the burst length becomes large.

In this paper, we consider the $(d_v, d_c, w, L, M)$ code ensemble introduced in [3] and derive tight lower bounds on the correctability of a long burst of erasures. First, we consider the case when a complete spatial position is erased and then generalize the expression to the case where the burst can occur at any position within a codeword. We show that estimating the capability of correcting long burst erasures reduces to the problem of finding small stopping sets in the code structure. Also, we demonstrate that if we properly expurgate the ensemble, then a random code from the ensemble has very good average burst erasure capabilities. We focus on the general $(d_v, d_c, w, L, M)$ code ensemble as the common protograph-based approach contains unavoidable small stopping sets in each spatial position, which are not recoverable if erased [10].

This work was conducted while N. Rengaswamy was visiting Bell Labs as a research intern funded by a scholarship of the DAAD-RisePro programme. The work of L. Schmalen was funded by the German Government in the frame of the CELTIC+/BMBF project SASER-SaveNet.

## II. Preliminaries

### A. The Regular $(d_v, d_c, w, L, M)$ SC-LDPC Ensemble

We now briefly review how to sample a code from a random regular $(d_v, d_c, w, L, M)$ SC-LDPC ensemble [3]. We first lay out a set of positions indexed from $z = 1$ to $L$ on a *spatial dimension*. At each spatial position (SP), $z$, there are $M$ variable nodes (VNs) and $M\frac{d_v}{d_c}$ check nodes (CNs), where $M\frac{d_v}{d_c} \in \mathbb{N}$ and $d_v$ and $d_c$ denote the variable and check node degrees, respectively. Let $w > 1$ denote the smoothing (coupling) parameter. Then, we additionally consider $w-1$ sets of $M\frac{d_v}{d_c}$ CNs in SPs $L+1, \ldots, L+w-1$. Every CN is assigned with $d_c$ "sockets" and made to impose an even parity constraint on its $d_c$ neighboring VNs. Each VN in SP $z$ is connected to $d_v$ CNs in SPs $z, \ldots, z+w-1$ as follows: each of the $d_v$ edges of this VN is allowed to randomly and uniformly connect to any of the $wMd_v$ sockets arising from the CNs in SPs $z, \ldots, z+w-1$, such that parallel edges are avoided in the resultant bipartite graph. This graph represents the code so that we have $n = LM$ code bits, over $L$ SPs.

Because of additional check nodes in SPs $z > L$, the code rate $r = 1 - \frac{d_v}{d_c} - \delta$, where $\delta = O(\frac{w}{L})$. Throughout this paper, we assume that $d_v \geq 3$ and $wM > 2(d_v + 1)d_c$.

### B. Single-Burst-Erasure Channel Models

We introduce two channel models for computing the burst erasure recoverability. First, the *Single Position Burst Channel* (SPBC) erases all $M$ VNs of exactly one SP in the transmitted codeword and leaves all other bits undisturbed.

The second model is the more general *Random Burst Channel* (RBC) whose burst pattern is denoted by RBC($\ell$,$s$,$b$) where $s \in \{1, \ldots, M\}$ is the starting bit index of the burst in SP $\ell \in \{1, \ldots, L\}$, indicating the offset from the first VN of the SP $\ell$, and $b$ is the length of the burst. Note that in general $0 < b \leq (L-\ell)M - s$. As for the SPBC, all VNs in the random burst are erased while all other VNs are received correctly. We sometimes omit the SP $\ell$ when referring to the RBC for the following reason: neglecting boundary effects in the limit of large enough $L$, all SPs are structured identically. With some abuse of terminology, we will use the same notation to refer to the channel itself, rather than the burst introduced by it.

While multiple models exist for a correlated erasure channel, like the Gilbert-Elliott model used in [6], we use this model because it is sufficient to describe the scenarios that we consider: for instance, the SPBC can be used to model a slotted-ALOHA multiple access scheme where each user transmits an SC-LDPC codeword over $L$ time slots, but one SP might be erased in the case of a collision. Additionally, long burst erasures might occur in block fading scenarios, or in optical communications with, e.g., polarization dependent loss, where long burst erasures are common. Another scenario is optical storage, where long erasure bursts may occur as well.

## III. Error Analysis on the SPBC

Let $P_{\mathrm{B}}^{\mathrm{SPBC}}(d_v, d_c, w, L, M)$ denote the average block erasure (decoding error) probability of the $(d_v, d_c, w, L, M)$ ensemble on the SPBC under BP decoding, i.e., the probability that the iterative decoder fails to recover the codeword. For large enough $M$, size-2 stopping sets (each of which also form a codeword) are the dominant structures in the graph that cause the BP decoder to fail [10]. *Stopping sets* are subsets $\mathcal{A}$ of the VNs such that every neighbor of the VNs in $\mathcal{A}$ connects to $\mathcal{A}$ at least twice [11, Def. 3.138]. A *minimal stopping set* is one which does not contain a smaller size non-empty stopping set within itself. Hence, the number of size-2 stopping sets per SP, denoted $\mathbb{N}_2^{\mathrm{SP}}$, is a good starting point for analyzing the performance of the ensemble. We have

$$\begin{aligned} P_{\mathrm{B}}^{\mathrm{SPBC}} &= \text{Prob [At least one stopping set in a SP]} \\ &\geq \text{Prob } [\mathbb{N}_2^{\mathrm{SP}} \geq 1] \\ &\overset{(a)}{\geq} \frac{\mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}]^2}{\mathbb{E}[\mathbb{N}_2^{\mathrm{SP}^2}]} \overset{(b)}{\geq} \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] \left(1 - \frac{M^2}{(\frac{w}{d_c}M - 3)^{d_v}}\right) \\ &= \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] \left(1 - O\left(\frac{1}{M^{d_v - 2}}\right)\right) \approx \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] \doteq \lambda_{\mathrm{SP}}, \quad (1) \end{aligned}$$

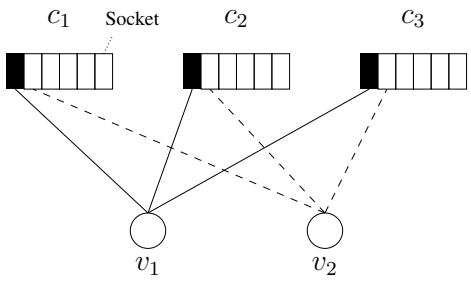



Fig. 1. A size-2 stopping set from a $(3, 6)$ random ensemble. CNs $\{c_1, c_2, c_3\}$ and VNs $\{v_1, v_2\}$ have been labeled for convenience. CNs have been expanded to show all their $d_c = 6$ sockets. The solid edges indicate definite connections and the dashed edges complete one configuration to form a stopping set. Parallel edges are not allowed in the ensemble.

where $(a)$ is the application of the second moment method and $(b)$ can be shown as follows: Define $U_{ij} = 1$ if VNs $i$ and $j$ form a stopping set, otherwise $U_{ij} = 0$. Then $\mathbb{N}_2^{\mathrm{SP}} = \sum_{1 \leq i < j \leq M} U_{ij}$ where the summation is over all $\binom{M}{2}$ pairs of VNs from a SP. We can see that $\lambda_{\mathrm{SP}} = \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] = \binom{M}{2} p$ where $p = \mathbb{E}[U_{ij}]$ is the probability of forming a size-2 stopping set. Furthermore,

$$\begin{aligned} \mathbb{E}\left[\mathbb{N}_2^{\mathrm{SP}^2}\right] &= \mathbb{E}\left[\left(\sum_{1 \leq i < j \leq M} U_{ij}\right)^2\right] \\ &= \sum_{1 \leq i < j \leq M} \mathbb{E}[U_{ij}^2] \quad + \sum_{\substack{(i,j) \neq (k,l) \\ i<j, k<l}} \mathbb{E}[U_{ij} U_{kl}], \end{aligned}$$

where in the last step, $\sum_{1 \leq i < j \leq M} \mathbb{E}[U_{ij}^2] = \binom{M}{2} p$ as $U_{ij} \in \{0, 1\}$ and the second term is over the remaining $\binom{M}{2}(\binom{M}{2} - 1)$ combinations. Using some combinatorial arguments, we can show that $\mathbb{E}[U_{ij} U_{kl}] = \mathbb{P}(U_{ij} = 1)\mathbb{P}(U_{kl} = 1 | U_{ij} = 1) \leq 2p / \binom{wM \frac{d_v}{d_c} - 2d_v}{d_v}$. As a result, we have

$$\begin{aligned} \mathbb{E}\left[\mathbb{N}_2^{\mathrm{SP}^2}\right] &< \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] \left(1 + \frac{2\binom{M}{2}}{\binom{wM \frac{d_v}{d_c} - 2d_v}{d_v}}\right) \\ &< \mathbb{E}[\mathbb{N}_2^{\mathrm{SP}}] \left(1 + \frac{M^2}{(\frac{w}{d_c}M - 3)^{d_v}}\right), \end{aligned}$$

which eventually implies (1). Note that following standard arguments [10], [11, Appendix C], we can also approximate the bound on $P_{\mathrm{B}}^{\mathrm{SPBC}}$ by a Poisson distribution with mean $\lambda_{\mathrm{SP}}$, for a large $M$, so that

$$P_{\mathrm{B}}^{\mathrm{SPBC}} \approx 1 - e^{-\lambda_{\mathrm{SP}}} \approx \lambda_{\mathrm{SP}}. \quad (2)$$

Both (1) and (2) are very tight when $w \geq d_v$ (which is a prerequisite for constructing capacity-achieving codes [3]) as otherwise, we have observed that the contribution of larger stopping sets becomes non-negligible.

### A. Calculation of p

We now calculate the probability $p$ of finding a size-2 stopping set within an SP of a code uniformly sampled from

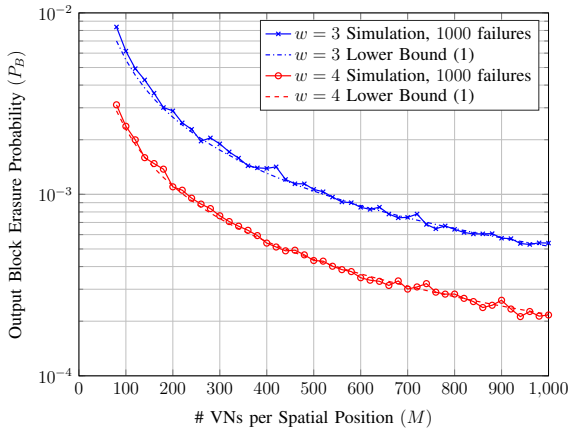



Fig. 2. Monte Carlo simulations on the SPBC with a $(3, 6)$ random ensemble for $w = 3$ and $w = 4$, along with their respective theoretical lower bound (1). The bound becomes tight very quickly with $M$.

an ensemble. As example, we randomly choose two VNs $v_1$ and $v_2$ from an SP of the $(d_v = 3, d_v = 6, w, L, M)$ ensemble. First, we connect the $d_v = 3$ edges of $v_1$ to randomly chosen empty sockets of $d_v$ distinct CNs as described in Section II-A. Let $c_1, c_2, c_3$ denote the CNs adjacent to $v_1$. A stopping set (and in this case, also a low-weight codeword) is formed if and only if the edges of $v_2$ are connected to the same CNs, i.e., $c_1, c_2, c_3$. This situation is shown in Fig. 1: once we have assigned $d_v$ CNs to $v_1$, we have $d_c - 1 = 5$ free distinct sockets each for CNs $c_1, c_2, c_3$. Thus, the first edge of $v_2$ has $d_v(d_c - 1) = 15$ ways to attach to these sockets, the second edge has $(d_v - 1)(d_c - 1) = 10$ ways and the last edge has $(d_v - 2)(d_c - 1) = 5$ ways. In general, the edges of $v_2$ can be connected to any of the $wMd_v - d_v$ possible sockets.

By a counting argument, we can compute $p = \frac{T_{ss}}{T}$ where $T_{ss}$ is the total number of combinations by which the edges of $v_2$ can form a stopping set with $v_1$ and $T$ is the total number of combinations by which the edges of $v_2$ can be fit to the possible CN sockets without forming parallel edges.

Hence, for a general $(d_v, d_c, w, M)$ ensemble we can calculate $p = \frac{T_{ss}}{T}$ with

$$T_{ss} = \prod_{i=0}^{d_v-1} (d_v - i)(d_c - 1) = d_v!(d_c - 1)^{d_v},$$

$$T = \sum_{i=0}^{d_v} \frac{(d_c - 1)^i d_v!}{(d_v - i)!} \binom{d_v}{i} \left[ \prod_{k=0}^{d_v-1-i} (wMd_v - (d_v + k)d_c) \right].$$

For large $M$, $T$ can be well approximated by the dominating summand ($i = 0$) leading to

$$p \approx \prod_{i=0}^{d_v-1} \frac{(d_v - i)(d_c - 1)}{(wMd_v - (d_v + i)d_c)} \approx \frac{d_v!(d_c - 1)^{d_v}}{((wM - d_c)d_v)^{d_v}}. \quad (3)$$

We observe that $\lambda_{SP} = \binom{M}{2} p \sim O(M^{2-d_v})$.

### B. Simulations

We performed Monte-Carlo simulations where we randomly selected a spatial position from the middle of the graph (to avoid boundary effects) to be erased, for each transmitted codeword. At the receiver we performed BP decoding and averaged over the ensemble. We counted 1000 decoding failures for each $M$ to assess the average block erasure probability $P_{\mathrm{B}}^{\mathrm{SPBC}}$. The simulation results for a $(3, 6)$ random ensemble with $w = 3$ and $w = 4$ are shown in Fig. 2 along with their respective lower bounds calculated using (1) and (3). We observe that the bound indeed becomes a good approximation for large $M$, since large-size stopping sets (larger than 2) vanish. The simulation curve is slightly unstable because counting 1000 failures is not enough to keep the sample variance small as $P_{\mathrm{B}}^{\mathrm{SPBC}}$ decreases by $O(M^{2-d_v})$.

## IV. Error Analysis on the RBC

We now generalize our results to the RBC, where a burst can span multiple spatial positions and can be of arbitrary length. Besides the stopping sets within a single spatial position, we first have to derive an expression for stopping sets that span multiple SPs.

### A. Size-2 Stopping Sets across Coupled SPs

The results from Sec. III can be extended when the channel is a RBC, i.e., the burst occurs at arbitrary location and is of arbitrary length. This means that size-2 stopping sets formed across coupled SPs will also contribute to decoding failures. Hence, we will now calculate the probability that two VNs chosen each from two coupled spatial positions form a stopping set.

Let us first consider two VNs chosen from two adjacent SPs: w.l.o.g, call them $v_1$ and $v_2$ chosen from SPs 1 and 2, respectively. We immediately notice that the check positions adjacent to $v_1$ are $1, 2, \ldots, w$ and to $v_2$ are $2, 3, \ldots, w + 1$. Hence, to form a stopping set, $v_1$ should not have any edge connected to check position 1. This restricts the number of favorable constellations [3] for $v_1$ to be $(w - 1)^{d_v}$. Using the same ideas as in Section III-A and restricting the constellations for $v_1$, we have

$$p_{(1,2)} = \frac{(w - 1)^{d_v}}{w^{d_v}} p,$$

where $p$ can be approximated by (3). This idea can now be extended to VNs chosen from positions $(1, 3), (1, 4), \ldots, (1, w)$ by restricting the number of favorable constellations for $v_1$. Hereafter, we will refer to these as size-2 $(1, i)$-stopping sets. Hence, a $(d_v, d_c, w, L, M)$ ensemble can be completely characterized, for large enough $M$, by the vector

$$\underline{p}(d_v, d_c, w, L, M) = (p_{(1,1)}, p_{(1,2)}, \ldots, p_{(1,w)}) \quad (4)$$

$$\text{with } p_{(1,i)} = \left( \frac{w - (i - 1)}{w} \right)^{d_v} p.$$

The average number of size-2 stopping sets of each type, $\lambda_{(1,i)}$, can be calculated as

$$\lambda_{(1,1)} = \binom{M}{2} p_{(1,1)} = \lambda_{SP} \quad ; \quad \lambda_{(1,i)} = M^2 p_{(1,i)}, \quad (5)$$

where $i = 2, 3, \ldots, w$. Again, we see that $\lambda_{(1,i)} \sim O(M^{2-d_v})$.

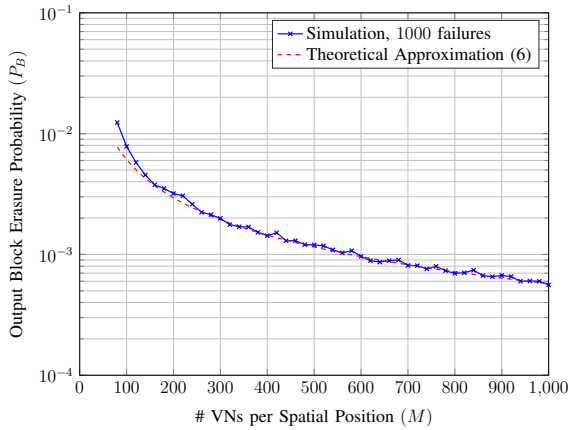



Fig. 3. Monte Carlo simulations for a $(3, 6, 3, 20, M)$ random ensemble on the RBC with burst length $b = 1.25M$, along with the theoretical approximation (6).

### B. Performance on the RBC

Now let us see the effect of RBC$(s, b)$ on the ensemble in terms of the average block erasure probability, $P_{\mathrm{B}}^{\mathrm{RBC}}$. For keeping the expressions simple, let us assume in the example that $w = 3$ and $0 < b \leq 2M$. This means that the burst can span a maximum of 3 SPs. Applying the same argument as in Section III and assuming all values for $s$ are equally likely,

$$P_{\mathrm{B}}^{\mathrm{RBC}} \approx \sum_{s=1}^{M} \frac{1 - P_{(1,1)}P_{(2,2)}P_{(3,3)}P_{(1,2)}P_{(2,3)}P_{(1,3)}}{M}; \quad (6)$$

$$P_{(k,k)} = 1 - \binom{m_k}{2} p_{(1,1)} \quad \text{for } k = 1, 2, 3,$$

$$P_{(k,k+1)} = 1 - m_k m_{k+1} p_{(1,2)} \quad \text{for } k = 1, 2,$$

$$P_{(k,k+2)} = 1 - m_k m_{k+2} p_{(1,3)} \quad \text{for } k = 1,$$

where $m_1 = (M - s), m_2 = \min(b - m_1, M), m_3 = (b - m_1 - m_2)$ are the lengths of the burst in each SP that it affects, progressing from left to right. If any of these lengths is zero, all probabilities involving that length are 1, i.e., the probability of forming no size-2 stopping sets involving the SP corresponding to this (zero) length is 1. For general $w$ and longer bursts, this strategy can be extended for finding a very good approximation for the average block erasure probability for the ensemble.

To verify the tightness of (6), we again performed Monte-Carlo simulations and counted 1000 decoding failures for each $M$ to assess the average block erasure probability $P_{\mathrm{B}}^{\mathrm{RBC}}$. For the sake of example, we fixed the burst length to be $b = 1.25M$. We selected a value for $s$, uniformly from $\{1, \ldots, M\}$, for each codeword. The simulation results for the $(3, 6, 3, 20, M)$ ensemble are shown in Fig. 3 along with (6). We see that (6) is indeed a tight approximation.

## V. Effects of Expurgation

### A. Minimal Stopping Set Size

As the performance is mainly dominated by size-2 stopping sets, we can improve the burst erasure correction capability by expurgating the ensemble and thereby removing all small stopping sets. Observing that a size-2 stopping set, as shown in Fig. 1, is built around 4-cycles, we can reduce the size of the minimal stopping sets by removing small cycles from the graph. For example, increasing the girth of the graph to 6 leads to minimal stopping sets (i.e., of smallest size) of size $d_v + 1$ [12].

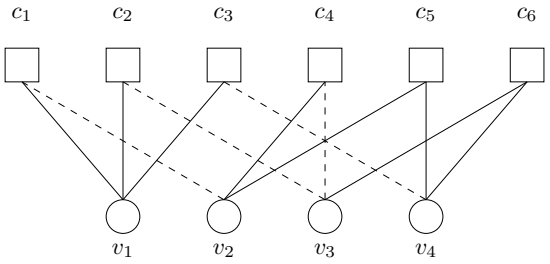



Fig. 4. A size-4 stopping set from an expurgated $(3, 6, w, L, M)$ random ensemble. CNs $\{c_1, c_2, c_3, c_4, c_5, c_6\}$ and VNs $\{v_1, v_2, v_3, v_4\}$ have been labeled for convenience. The solid edges indicate definite connections and the dashed edges complete one configuration to form a stopping set. Parallel edges are not allowed in the ensemble.

### B. Performance on the SPBC

We can use the same approach as in Section III-A to calculate the probability of occurrence of the stopping set shown in Fig. 4 within a spatial position of a code sampled uniformly from the ensemble. Once again we have $p = \frac{T_{ss}}{T}$, where $T_{ss}$ is the total number of combinations of the edges of $v_1, v_2, v_3, v_4$ that form a stopping set and $T$ is the total number of combinations by which these edges can fit to the available CN sockets. Since $T$ is the total number of combinations in which the edges of $(d_v + 1)$ VNs can be assigned to sockets ensuring no 4-cycles, we can again approximate it by its dominant term as

$$T \approx \prod_{j=0}^{d_v(d_v+1)-1} (wMd_v - jd_c).$$

For a general $(d_v, d_c, w, M)$ random ensemble, the expression for $T_{ss}$ can be calculated as

$$T_{ss} = \prod_{i=0}^{d_v} \left[ \prod_{j=1}^{i} j(d_c - 1)(d_v - i + 1) \right] \times \left[ \prod_{k=\sum_{m=0}^{i-1}(d_v-m)}^{\sum_{m=0}^{i-1}(d_v-m)+(d_v-i-1)} (wMd_v - kd_c) \right] \binom{d_v}{i}.$$

It can be verified that the last value for $k$ in the above expression is $k = \frac{d_v(d_v+1)}{2} - 1$. Then, we can simplify and rearrange the expression as

$$T_{ss} = T_{1/2} \prod_{i=1}^{d_v} [(d_c - 1)(d_v - i + 1)]^i \frac{d_v!}{(d_v - i)!},$$

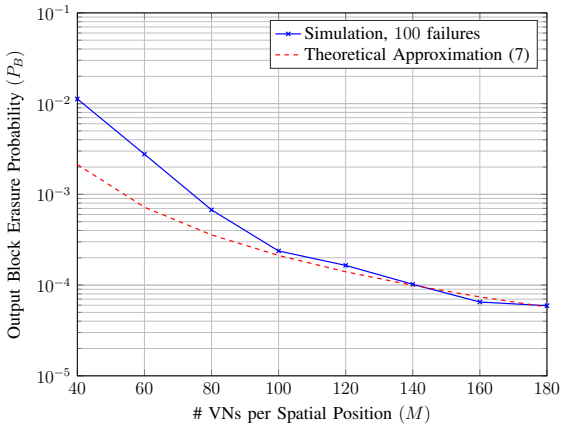



Fig. 5. Monte Carlo simulations on the SPBC with an expurgated $(3, 6)$ random ensemble for $w = 3$ along with the theoretical approximation. The approximation becomes tight very quickly with $M$.

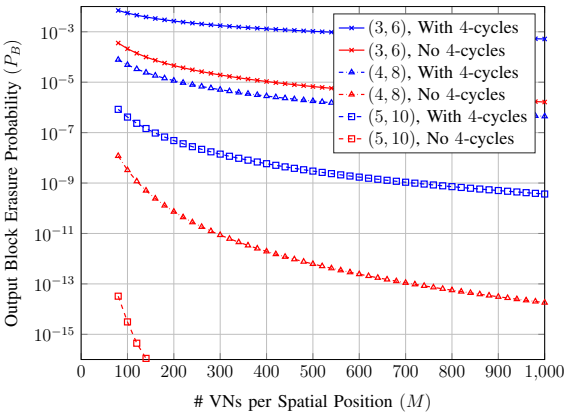



Fig. 6. The theoretical approximation (1) and (7) on $P_B^{\text{SPBC}}$ for various ensembles in both the unexpurgated and expurgated scenarios.

where, $T_{1/2} = \prod_{k=0}^{\frac{d_v(d_v+1)}{2}-1}(wMd_v - kd_c)$ is the first half of the products in $T$ which can be canceled while calculating $p$, so that

$$\frac{T}{T_{1/2}} \approx \prod_{j=\frac{d_v(d_v+1)}{2}}^{d_v(d_v+1)-1}(wMd_v - jd_c).$$

Hence, for a general $(d_v, d_c, w, L, M)$ ensemble, the probability of forming such a minimal stopping set of size $(d_v + 1)$ can be shown to be

$$p = \frac{T_{ss}}{T} \approx \frac{\prod_{i=1}^{d_v}\left[(d_c-1)(d_v-i+1)\right]^i \frac{d_v!}{(d_v-i)!}}{\prod_{j=\frac{d_v(d_v+1)}{2}}^{d_v(d_v+1)-1}(wMd_v - jd_c)} \quad (7)$$

which means the expected number of such stopping sets within a SP of the code is $\lambda_{SP} = \binom{M}{d_v+1}p$. Using similar arguments as in Section III, we have $P_{\text{B,exp}}^{\text{SPBC}} \approx \lambda_{SP}$.

### C. Comparison of Ensembles

We now compare the average performance of different SC-LDPC ensembles on the SPBC. We fix the asymptotic design code rate as $r = \frac{1}{2}$, the smoothing parameter as $w = d_v$ and plot the (tight) approximations on $P_B^{\text{SPBC}}$ of three ensembles, namely $(3, 6)$, $(4, 8)$ and $(5, 10)$, for both the unexpurgated and the expurgated cases in Fig. 6.

For the unexpurgated case, the average block erasure probability varies as $P_{\text{B}}^{\text{SPBC}} \sim O(M^{2-d_v})$. When the ensemble is expurgated, the improvement is by an order of $\frac{d_v+1}{2}$ in $M$ and we have $P_{\text{B,exp}}^{\text{SPBC}} \sim O(M^{(d_v+1)(2-d_v)/2})$. Therefore, for a fixed (asymptotic design) rate of $\frac{1}{2}$, a unit increase in $d_v$ improves the performance by a factor of about $M^{-d_v}$.

## VI. Conclusion

We have analyzed random regular SC-LDPC ensembles on the burst erasure channel and provided insights into improving the block erasure probability by increasing VN degree and expurgating the code. We have shown, through these results, that the vector in (4) completely characterizes the average ensemble performance on the erasure channel.

Future work will focus on, among others, finding good approximations for the expurgated ensembles in the case of the RBC model and to extend the considerations to the case where we have independent random erasures besides the burst erasures. We note that the vector in (4) will play a significant role in the analysis of random erasures too.

## References

[1] A. Jiminez Felström and K. Zigangirov, "Time-varying periodic convolutional codes with low-density parity-check matrix," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2181–2191, Sep. 1999.

[2] M. Lentmaier, G. P. Fettweis, K. Zigangirov, and D. J. Costello, Jr., "Approaching capacity with asymptotically regular LDPC codes," in *Proc. ITA*, 2009.

[3] S. Kudekar, T. Richardson, and R. Urbanke, "Threshold saturation via spatial coupling: Why convolutional LDPC ensembles perform so well over the BEC," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 803–834, Feb 2011.

[4] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7761–7813, 2013.

[5] A. Jule and I. Andriyanova, "Performance bounds for spatially-coupled LDPC codes over the block erasure channel," in *Proc. IEEE ISIT*, July 2013, pp. 1879–1883.

[6] A. Iyengar, M. Papaleo, G. Liva, P. Siegel, J. Wolf, and G. Corazza, "Protograph-based LDPC convolutional codes for correlated erasure channels," in *Proc. IEEE ICC*, May 2010, pp. 1–6.

[7] H. Mori and T. Wadayama, "Band splitting permutations for spatially coupled LDPC codes enhancing burst erasure immunity," *arXiV*, 2015. [Online]. Available: http://arxiv.org/abs/1501.04394

[8] N. ul Hassan, M. Lentmaier, I. Andriyanova, and G. P. Fettweis, "Improving code diversity on block-fading channels by spatial coupling," in *Proc. IEEE ISIT*, Honolulu, HI, USA, Jun. 2014.

[9] N. ul Hassan, I. Andriyanova, M. Lentmaier, and G. P. Fettweis, "Protograph design for spatially-coupled codes to attain an arbitrary diversity order," in *Proc. ITW*, Jeju City, South Korea, Oct. 2015.

[10] P. Olmos and R. Urbanke, "Scaling behavior of convolutional LDPC ensembles over the BEC," in *Proc. IEEE ISIT*, July 2011, pp. 1816–1820.

[11] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[12] A. Orlitsky, R. Urbanke, K. Viswanathan, and J. Zhang, "Stopping sets and the girth of Tanner graphs," in *Proc. IEEE ISIT*, 2002.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# The Fractality of Polar Codes

Bernhard C. Geiger
Institute for Communications Engineering, Technische Universität München, Germany
geiger@ieee.org

***Abstract*—The generator matrix of a polar code is obtained by selecting rows from the Kronecker product of a lower-triangular binary square matrix. The selection is based on the Bhattacharyya parameter of the row, which is closely related to the error probability of the corresponding input bit under sequential decoding. This work investigates the properties of the index set pointing to those rows in the infinite blocklength limit. In particular, the Lebesgue measure, the Hausdorff dimension, and the self-similarity of this set will be discussed. It is shown that these index sets fulfill several properties that are common to fractals.**

## I. Introduction

Applying the polarization transform proposed by Arıkan [1] to sufficiently many instances of a binary-input memoryless channel, causes a portion of the resulting channels to have a capacity close to one, while the remaining portion has a capacity close to zero. These *polarized channels* can thus be split into two sets: The set of "good" channels, and the set of "bad" channels. Despite their importance for code construction, very little is known about their structure. A recent exception is the work by Renes, Sutter, and Hassani, stating conditions under which polarized sets are aligned, i.e., under which the good (bad) channels derived from one binary-input memoryless channel are a subset of the good (bad) channels derived from another [2].

Polar codes are Kronecker product-based codes. Such a code of block-length $2^n$ is based on the $n$-fold Kronecker product $G(n) := F^{\otimes n}$, where

$$F := \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{1}$$

Following the terminology of [3], a rate-$K/2^n$ Kronecker product-based code is uniquely defined by a set $\mathcal{F}$ of $K$ indices: Its generator matrix is the submatrix of $G(n)$ consisting of the rows indexed by $\mathcal{F}$. For polar codes, in which each row of $G(n)$ can be interpreted as a (partially polarized) channel, $\mathcal{F}$ consists of rows corresponding to the $K$ channels with the lowest Bhattacharyya parameter [4] (see Section II).

That Kronecker product-based codes, such as polar codes [1] or Reed-Muller codes, possess a fractal nature has been observed in [3], where it was noted that $G(n)$ resembles a Sierpinski triangle. Much earlier, Abbe suspected that the set of "good" channels has fractal nature [5]. Nevertheless, to the best of the author's knowledge, no definite statement regarding this fractal nature has been made yet. In this paper, we try to fill this gap and present results about the set of "good" channels (Sections III). Specifically, we study the properties of the set $\mathcal{F}$ for infinite blocklengths, i.e., for $n \to \infty$.

To simplify analysis, we represent every infinite binary sequence indexed in $\mathcal{F}$ by a point in the unit interval $[0,1]$. Let $\Omega = \{0,1\}^\infty$ be the set of infinite binary sequences, and let $b := (b_1 b_2 \cdots) \in \Omega$ be an arbitrary such sequence. We abbreviate $b^n := (b_1 b_2 \cdots b_n)$. Let $(\Omega, \mathfrak{B}, \mathbb{P})$ be a probability space with $\mathfrak{B}$ the Borel field generated by the cylinder sets $S(b^n) := \{w \in \Omega\colon w_1 = b_1, \ldots, w_n = b_2\}$ and $\mathbb{P}$ a probability measure satisfying $\mathbb{P}(S(b^n)) = 1/2^n$. The following function $f\colon \Omega \to [0,1]$ permits us to convert these sequences to real numbers:

$$f(b) := \sum_{n=1}^{\infty} \frac{b_n}{2^n} \tag{2}$$

Letting $\mathbb{D} := [0,1] \cap \{p/2^n\colon p \in \mathbb{Z}, n \in \mathbb{N}\}$ denote the set of dyadic rationals in the unit interval, we recognize that $f$ is not injective:

**Example 1.** $f$ maps both $b = (01111111\cdots)$ and $b = (10000000\cdots)$ to 0.5. We call the latter binary expansion *terminating*.

However, as the following lemma shows, $f$ is bijective if we exclude the dyadic rationals:

**Lemma 1** ([6, Exercises 7-10, p. 80])**.** *Let $\mathfrak{B}_{[0,1]}$ be the Borel $\sigma$-algebra on $[0,1]$ and let $\lambda$ be the Lebesgue measure. Then, the function $f$ in* (2) *satisfies the following properties:*

1) *$f$ is measurable w.r.t. $\mathfrak{B}_{[0,1]}$*
2) *$f$ is bijective on $\Omega \setminus f^{-1}(\mathbb{D})$*
3) *for all $I \in \mathfrak{B}_{[0,1]}$, $\mathbb{P}(f^{-1}(I)) = \lambda(I)$*

We believe that the results we prove in the following not only improve our understanding of polar codes: Since its introduction in 2009, the polarization technique proposed by Arıkan has found its way into areas different from polar coding. Haghighatshoar and Abbe showed in the context of compression of analog sources that Rényi information dimension can be polarized [7], and Abbe and Wigderson used polarization for the construction of high-girth matrices [8]. Recently, Nasser proved that a binary operation is polarizing if and only if it is uniformity preserving and its inverse is strongly ergodic [9], [10]. We believe that our results might carry over to these areas as well and point to possible extensions in Section IV.

## II. Preliminaries for Polar Codes

We adopt the notation of [1]: Let $W\colon \{0,1\} \to \mathcal{Y}$ be a binary-input memoryless channel with output alphabet $\mathcal{Y}$, capacity $0 < I(W) < 1$, and with Bhattacharyya parameter

$$Z(W) := \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \tag{3}$$

That $Z(W) = 0 \Leftrightarrow I(W) = 1$ and $Z(W) = 1 \Leftrightarrow I(W) = 0$ is a direct consequence of [1, Prop. 1].

The heart of Arıkan's polarization technique is that two channel uses of $W$ can be *combined and split* into one use of a "worse" channel

$$W_2^0(y_1^2|u_1) := \frac{1}{2}\sum_{u_2} W(y_1|u_1 \oplus u_2)W(y_2|u_2) \tag{4a}$$

and one use of a better channel

$$W_2^1(y_1^2, u_1|u_2) := \frac{1}{2}W(y_1|u_1 \oplus u_2)W(y_2|u_2) \tag{4b}$$

where $u_1, u_2 \in \{0, 1\}$ and $y_1, y_2 \in \mathcal{Y}$. In essence, the combining operation codes two input bits by $F$ in (1) and transmits the coded bits over $W$ via two channel uses, creating a vector channel. The splitting operation splits this vector channel into the two binary-input memoryless channels indicated in (4). Of these, the better (worse) channel has a strictly larger (smaller) capacity than the original channel $W$, i.e., $I(W_2^0) < I(W) < I(W_2^1)$, while the sum capacity equals twice the capacity of the original channel, i.e., $I(W_2^0) + I(W_2^1) = 2I(W)$ [1, Prop. 4].

The effect of combining and splitting on the channel capacities $I(W_2^0)$ and $I(W_2^1)$ admits no closed-form expression; the effect on the Bhattacharyya parameter at least admits bounds:

**Lemma 2** ([1, Prop. 5 & 7])**.**

$$Z(W_2^1) = g_1(Z(W)) = Z^2(W) < Z(W) \tag{5a}$$

$$Z(W) < Z(W_2^0) \leq g_0(Z(W)) = 2Z(W) - Z^2(W) \tag{5b}$$

*with equality if $W$ is a binary erasure channel.*

Channels with larger blocklengths $2^n$, $n > 1$, can either be obtained by direct $n$-fold combining (using the matrix $G(n)$) and $n$-fold splitting, or by recursive pairwise combining and splitting. For $b^n \in \{0, 1\}^n$, we obtain

$$\left(W_{2^n}^{b^n}, W_{2^n}^{b^n}\right) \to \left(W_{2^{n+1}}^{b^n 0}, W_{2^{n+1}}^{b^n 1}\right) \tag{6}$$

where $b^n 0$ and $b^n 1$ denote the sequences of zeros and ones obtained by appending 0 and 1 to $b^n$, respectively. Note that $g_1$ and $g_0$ from Lemma 2 are non-negative and non-decreasing functions mapping the unit interval onto itself, hence the inequality in (5b) is preserved under composition:

$$Z(W_{2^n}^{b^n}) \leq p_{b^n}(Z(W)) := g_{b_n}(g_{b_{n-1}}(\cdots g_{b_1}(Z(W))\cdots)) \tag{7}$$

The channel polarization theorem shows that, with probability one, after infinitely many combinations and splits, only perfect or useless channels remain, i.e., either $I(W_\infty^b) = 1$ or $I(W_\infty^b) = 0$ for $b \in \{0, 1\}^\infty$. This is made precise in:

**Proposition 1** ([1, Prop. 10])**.** *With probability one, the limit RV $I_\infty(b) := I(W_\infty^b)$ takes values in the set $\{0, 1\}$: $\mathbb{P}(I_\infty = 1) = I(W)$ and $\mathbb{P}(I_\infty = 0) = 1 - I(W)$.*

If the polarization procedure is stopped at a finite blocklength $2^n$ for $n$ large enough, it can still be shown that the vast majority of the resulting $2^n$ channels are either almost perfect or almost useless, in the sense that the channel capacities are close to one or to zero (or, that the corresponding Bhattacharyya parameters are close to zero or to one). The idea of polar coding is to transmit data only on those channels that are almost perfect: $n$-fold combining, which employs the matrix $G(n)$, leads to $2^n$ virtual channels, each corresponding to a row of $G(n)$. The channels with high capacity are indicated by the set $\mathcal{F}$, and the generator matrix of the corresponding polar code is precisely the submatrix of $G(n)$ consisting of those indicated rows.

The difficulty of polar coding lies in code construction, i.e., in determining *which* channels/row indices are in the set $\mathcal{F}$. This immediately translates to the question which sequences $b \in \{0, 1\}^\infty$ correspond to combinations and splits leading to a perfect channel (or which finite-length sequences $b^n$ lead to channels with capacity sufficiently close to one). Determining the capacity of the virtual channels is an inherently difficult operation, since, whenever $W$ is not a binary erasure channel (BEC), the cardinality of the output alphabet increases exponentially in $2^n$ [11, Ch. 3.3], [12, p. 36]. To circumvent this problem, Tal and Vardy presented an approximate construction method in [13], that relies on working with reduced output alphabet channels that are either upgraded or degraded w.r.t. the real channel. As these upgrading/degrading properties – mentioned earlier in Korada's PhD thesis [12] – will play a fundamental role in this work, we present

**Definition 1** (Channel Up- and Degrading)**.** A channel $W^-$: $\{0, 1\} \to \mathcal{Z}$ is *degraded* w.r.t. the channel $W$ (short: $W^- \preccurlyeq W$) if there exists a channel $P$: $\mathcal{Y} \to \mathcal{Z}$ such that

$$W^-(z|u) = \sum_{y\in\mathcal{Y}} W(y|u)P(z|y). \tag{8}$$

A channel $W^+$: $\{0, 1\} \to \mathcal{Z}$ is *upgraded* w.r.t. the channel $W$ (short: $W^+ \succcurlyeq W$) if there exists a channel $P$: $\mathcal{Z} \to \mathcal{Y}$ such that

$$W(y|u) = \sum_{z\in\mathcal{Z}} W^+(z|u)P(y|z). \tag{9}$$

Moreover, $W^+ \succcurlyeq W$ if and only if $W \preccurlyeq W^+$.

The upgraded (degraded) approximation remains upgraded (degraded) during combining and splitting:

**Lemma 3** ([12, Lem. 4.7] & [13, Lem. 3])**.** *Assume that $W^- \preccurlyeq W \preccurlyeq W^+$. Then,*

$$I(W^-) \leq I(W) \leq I(W^+) \tag{10a}$$

$$Z(W^-) \geq Z(W) \geq Z(W^+) \tag{10b}$$

$$(W^-)_2^1 \preccurlyeq W_2^1 \preccurlyeq (W^+)_2^1 \tag{10c}$$

$$(W^-)_2^0 \preccurlyeq W_2^0 \preccurlyeq (W^+)_2^0. \tag{10d}$$

It can be shown that the better channel (4b) obtained from combining and splitting is upgraded w.r.t. the original channel (as already mentioned in [11, p. 9]). That the worse channel (4a) is degraded holds at least for the BEC:

**Lemma 4.** *$W \preccurlyeq W_2^1$. If $W$ is a BEC, then $W_2^0 \preccurlyeq W \preccurlyeq W_2^1$.*

*Proof.* The proof of the first part follows by choosing

$$P(y|y_1^2, u_1) = \begin{cases} 1, & \text{if } y = y_2 \\ 0, & \text{else.} \end{cases} \tag{11}$$

For the BEC, note that if $W$ has erasure probability $\epsilon$, then $W_2^1$ is a BEC with erasure probability $\epsilon^2$ and $W_2^0$ is a BEC with erasure probability $2\epsilon - \epsilon^2$ [1, Prop. 6]. The channel $W_2^1$ is an upgrade of $W$, because it can be degraded to $W$ by appending a BEC with erasure probability $\epsilon/(1+\epsilon)$. The channel $W_2^0$ is degraded w.r.t. $W$ by appending a BEC with erasure probability $\epsilon$. □

## III. Properties of the Sets $\mathcal{G}$ and $\mathcal{B}$

In this section we develop the properties of the sets of good and bad channels. For the sake of brevity, we only sketch the proofs here; complete proofs are given in [14].

**Definition 2** (The Good and the Bad Channels)**.** Let $\mathcal{G}$ denote the set of good channels, i.e.,

$$x \in \mathcal{G} \Leftrightarrow \exists b \in f^{-1}(x)\colon I(W_\infty^b) = 1; \tag{12}$$

let $\mathcal{B}$ denote the set of bad channels, i.e.,

$$x \in \mathcal{B} \Leftrightarrow \exists b \in f^{-1}(x)\colon I(W_\infty^b) = 0. \tag{13}$$

**Proposition 2.** *For almost all $x$, there exists a value $0 \leq \vartheta(x) \leq 1$ such that $Z(W) < \vartheta(x)$ implies $x \in \mathcal{G}$. If $W$ is a BEC, then additionally $Z(W) > \vartheta(x)$ implies $x \in \mathcal{B}$.*

*Sketch of Proof:* This proposition is an adaption of [15, Lem. 11] to our setting: The lemma states that, for $\mathbb{P}$-almost every sequence $b$, there is a threshold $\theta(b)$ such that $\lim_{n\to\infty} p_{b^n}(z)$ converges to zero (one) if $z$ is smaller (larger) than $\theta(b)$. The rest follows from Lemma 2. ■

Note that if $W$ is not a BEC, it may occur that $Z(W) > \vartheta(f(b))$ while still $I(W_\infty^b) = 1$. This in turn opens the question whether the set of good channels is (almost surely) increasing with decreasing Bhattacharyya parameter: Are there channels $W$ and $W'$ (from the same family) with good channel sets $\mathcal{G}$ and $\mathcal{G}'$, respectively, such that $Z(W) > Z(W') > \vartheta(f(b))$, but $I(W_\infty^b) = 1$ and $I({W'}_\infty^b) = 0$? We leave this question for future research but mention that Proposition 2 answers it negatively for BECs: The set of good channels for a BEC is also good for any binary-input memoryless channel with a smaller Bhattacharyya parameter [16].

**Example 2.** For $x \in \mathbb{D}$, $\vartheta(x) = 1$: If $Z(W) < 1$, i.e., if the channel is not completely useless a priori, the non-terminating expansion of $x$ will make it a perfect channel (cf. Proposition 3).

**Example 3.** Let $x = 2/3$, hence $f^{-1}(x) = 101010101\cdots$. The binary expansion is recurring. It thus suffices to consider exactly one period of the recurring sequence and determine its fixed points. In this case we get $p_{10}(z) = 2z^2 - z^4$. Its fixed point lies at the intersection of $p_{10}(z)$ and $z$; removing the trivial intersections at $z = 0$ and $z = 1$ leaves two further roots at $(\pm\sqrt{5} - 1)/2$. One of these roots lies outside $[0, 1]$ and is hence irrelevant. The remaining root determines the threshold: $\vartheta(2/3) = (\sqrt{5} - 1)/2$. Now let $W$ be a BEC with erasure probability $\epsilon = Z(W) = \vartheta(2/3)$. Since $\epsilon = \vartheta(2/3)$ is a fixed point of the iterated function system corresponding to the recurring binary expansion, one gets $Z(W_\infty^{f^{-1}(2/3)}) = \epsilon \notin \{0, 1\}$. This example illustrates why Proposition 1 holds only almost surely.

**Proposition 3.** $\mathcal{G} \cap \mathcal{B} = \mathbb{D}$.

*Sketch of Proof:* The proof is based on the fact that dyadic rationals admit two possible binary expansions (see Example 1): The Bhattacharyya parameter of the non-terminating expansion $a^k 111\cdots$, for $a^k \in \{0, 1\}^k$ an appropriate prefix, is driven down to zero by squaring $Z(W_{2^k}^{a^k})$ infinitely often.

The terminating expansion has the same prefix $a^k$ with the last bit inverted. All binary sequences starting with this prefix lead to a channel that is upgraded w.r.t. the one corresponding to the terminating expansion (Lemmas 3 and 4). By Proposition 1, some sequences with this prefix lead to bad channels, hence the terminating expansion must lead to a bad channel as well. ■

That the intersection of the sets of good and bad channels is non-empty is a direct consequence of the non-injectivity of $f$. Note further that this intersection cannot be larger, since $\mathbb{D}$ is the only set to which $f$ maps non-injectively. Since $\mathbb{D}$, a common subset of $\mathcal{G}$ and $\mathcal{B}$, is dense in $[0, 1]$, both the set of good channels and the set of bad channels are dense in the unit interval. But even if dyadic rationals are excluded, results about denseness can be proved:

**Proposition 4.** *$\mathcal{G} \setminus \mathbb{D}$ is dense in $[0, 1]$. If $W$ is a BEC, then also $\mathcal{B} \setminus \mathbb{D}$ is dense in $[0, 1]$.*

*Sketch of Proof:* We sketch only the first part of the proof, the second part involving BECs follows along the same lines. The proof is based on the polynomial $p_b(z)$. Let $b^n$ be an arbitrary prefix (corresponding to a dyadic rational), leading to a Bhattacharyya parameter $Z(W_{2^n}^{b^n})$. There exists a sequence $a^k$ with one zero and sufficiently many ones such that $p_{a^k}(z) < z$ for all $z$ below a certain threshold $z^*(a^k) > Z(W_{2^n}^{b^n})$. It follows by Lemma 2 that $Z(W_\infty^{b^n a^k a^k \cdots}) \leq p_{a^k a^k \ldots}(Z(W_{2^n}^{b^n})) \to 0$, hence $f(b^n a^k a^k \cdots) \in \mathcal{G}$. Finally, between any two dyadic rationals a rational can be found with binary expansion $b^n a^k a^k \cdots$ that satisfies these properties. This proves that the good channels are dense even excluding the dyadic rationals. The inequality in Lemma 2 is the reason why denseness of bad channels can only be proved for BECs. ■

The proposition states that, at least for the BEC, there is no interval which contains only good channels. Hence, given a specific channel $W_{2^n}^{b^n}$, it is not possible to assume that a well-specified subset of channels (e.g., all $W_\infty^{b^n a}$ for $a$ starting with 1) generated from this channel by combining and splitting will be perfect. The construction algorithm for an infinite-blocklength, *vanishing-error* polar code hence cannot stop at a finite blocklength, as it can be done for a finite-blocklength polar code, cf. [17].

**Proposition 5.** *$\mathcal{G}$ is Lebesgue measurable and has Lebesgue measure $\lambda(\mathcal{G}) = I(W)$. $\mathcal{B}$ is Lebesgue measurable and has Lebesgue measure $\lambda(\mathcal{B}) = 1 - I(W)$. The Hausdorff dimensions of $\mathcal{G}$ and $\mathcal{B}$ satisfy $d(\mathcal{G}) = 1$ and $d(\mathcal{B}) = 1$.*

*Sketch of Proof:* The proof for the good channels follows

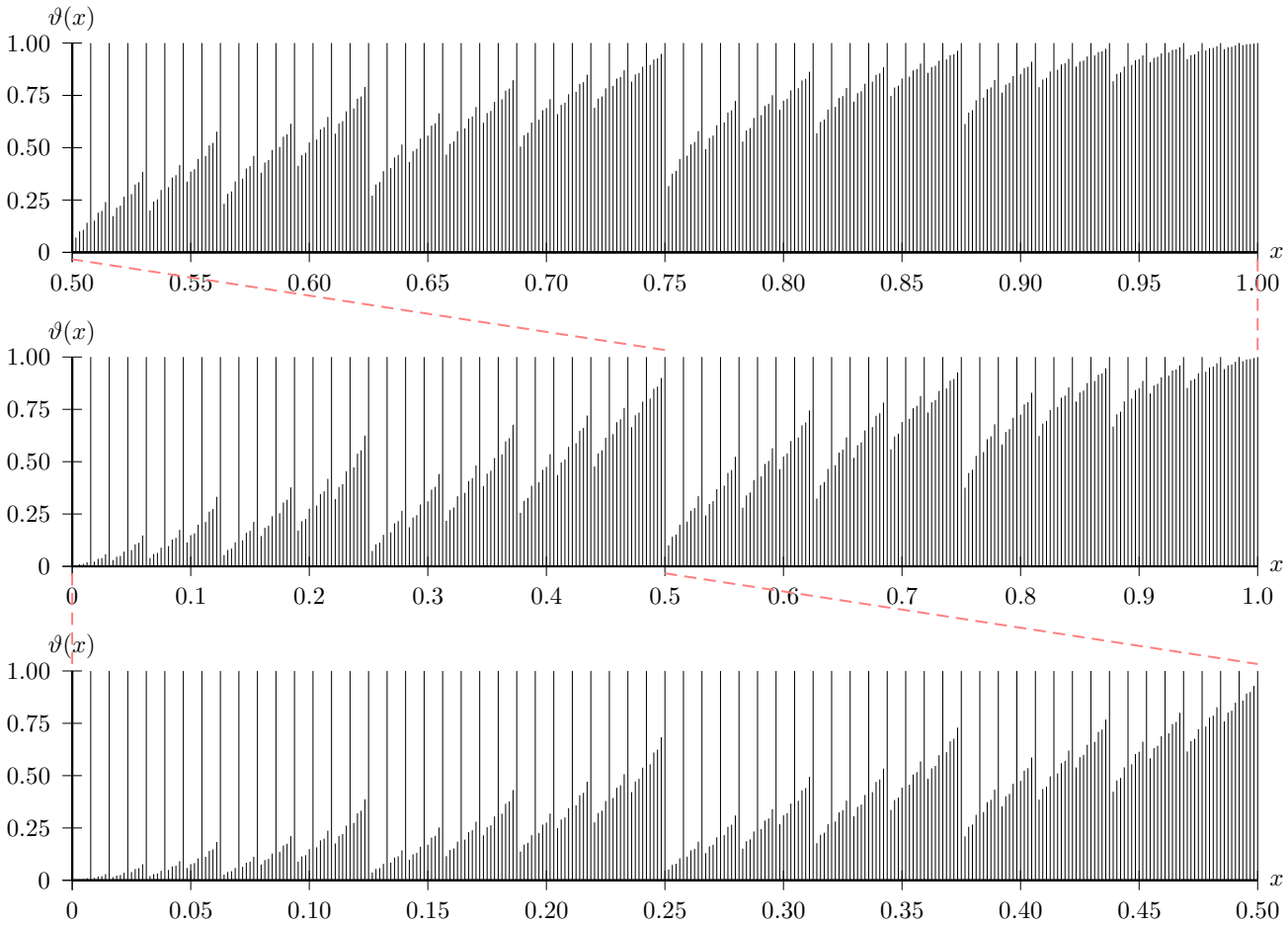



Fig. 1. The polar fractal for a BEC. The center plot shows the thresholds $\vartheta(x)$ for $x \in [0,1]$, while the bottom and the top plots show these thresholds for the scaled and shifted sets $[0, 0.5]$ and $[0.5, 1]$, respectively. Hence, the thresholds in the top plot are larger than the thresholds in the center plot, which are larger than those in the bottom plot. The set $\mathcal{G}$ is obtained by setting each value in the plot to one (zero) if the erasure probability $\epsilon$ is smaller (larger) than the threshold.

from the fact that $\lambda(\mathcal{G}) = \lambda(\mathcal{G} \setminus \mathbb{D})$, from Definition 2 stating

$$x \notin \mathbb{D}\text{: } x \in \mathcal{G} \Leftrightarrow I(W_\infty^{f^{-1}(x)}) = 1, \tag{14}$$

and from Proposition 1; the proof for the bad channels follows along the same lines. That the Hausdorff dimension of both sets is unity follows from the fact that the one-dimensional Hausdorff measure of a set equals its Lebesgue measure up to a constant [18, eq. (3.4), p. 45]. ■

Note that despite the fact that $\lambda(\mathcal{G} \cup \mathcal{B}) = 1$, $\mathcal{G} \cup \mathcal{B} \subset [0,1]$. The reason is that convergence to good or bad channels is only almost sure, and that there may be channels $W_\infty^b$ which are neither good nor bad (see Example 3).

We finally come to the claim that polar codes are fractal. Following Falconer's definition [18, p. xxviii], a set is fractal if it is (at least approximately) self-similar and has detail on arbitrarily small scales, or if its fractal dimension (e.g., its Hausdorff dimension) is larger than its topological dimension. Whether or not the result shown below will convince the reader of this property is a mere question of definition; strictly speaking, we can show only *quasi self-similarity* of $\mathcal{G}$:

**Proposition 6.** *Let* $\mathcal{G}_n(k) := \mathcal{G} \cap [(k-1)2^{-n}, k2^{-n}]$ *for* $k = 1, \ldots, 2^n$. $\mathcal{G} = \mathcal{G}_0(1)$ *is* quasi self-similar *in the sense that, for all* $n$ *and all* $k$, $\mathcal{G}_n(k) = \mathcal{G}_{n+1}(2k-1) \cup \mathcal{G}_{n+1}(2k)$ *is quasi self-similar to its right half:*

$$\mathcal{G}_n(k) \subset 2\mathcal{G}_{n+1}(2k) - k2^{-n} \tag{15}$$

*If* $W$ *is a BEC,* $\mathcal{G}_n(k)$ *is quasi self-similar:*

$$2\mathcal{G}_{n+1}(2k-1) - (k-1)2^{-n} \subset \mathcal{G}_n(k) \subset 2\mathcal{G}_{n+1}(2k) - k2^{-n} \tag{16}$$

*Sketch of Proof:* We only prove the result for $x \notin \mathbb{D}$, since the dyadic rationals are self-similar and since $\mathbb{D} \subset \mathcal{G}$. If $b_k^n = b_1 b_2 \cdots b_n$ is the terminating binary expansion of $(k-1)2^{-n}$, every value in $[(k-1)2^{-n}, k2^{-n}]$ has a binary expansion $b_k^n a$ for some $a \in \{0,1\}^\infty$, where $b_n = 1$ if and only if $(k-1)$ is odd. Similarly, and since $(2k-1)$ is always odd, every value in $[(2k-1)2^{-n-1}, k2^{-n}]$ has a binary expansion $b_k^n 1 a'$ for some $a' \in \{0,1\}^\infty$. Assume that $a' = a$. Then, by Lemmas 3 and 4, $W_\infty^{b_k^n a} \preccurlyeq W_\infty^{b_k^n 1a}$ for all $a$. Hence, if $f(b_k^n a) \in \mathcal{G}_n(k)$, then $f(b_k^n 1a) \in \mathcal{G}_{n+1}(2k)$. The proof follows by showing that $2f(b_k^n 1a) - f(b_{k+1}^n) = f(b_k^n a)$. For the BEC, the proof follows from the fact that by Lemmas 3 and 4, $W_\infty^{b_k^n 0a} \preccurlyeq W_\infty^{b_k^n a}$ for all $a$. ■

In other words, at least for the BEC, $\mathcal{G}$ is composed of two

similar copies of itself (see Fig. 1). Along the same lines, the quasi self-similarity of $\mathcal{B}$ can be shown.

**Example 4.** By careful computations we obtain $\vartheta(1/6) \approx 0.214$, $\vartheta(1/3) \approx 0.382$, and $\vartheta(2/3) \approx 0.618$. Indeed, if we consider $1/3$ in $\mathcal{G}$, then $1/6$ and $2/3$ are the corresponding values in $\mathcal{G}_1(1)$ and $\mathcal{G}_1(2)$. Since $\vartheta(1/6) < \vartheta(1/3) < \vartheta(2/3)$, for the BEC we have the inclusion indicated in Proposition 6.

## IV. Discussion & Outlook

That polar codes satisfy fractal properties has long been suspected: Every nontrivial, partly polarized channel $W_{2^n}^{b^n}$ gives rise, by further polarization, to both perfect and useless channels, regardless how close $I(W_{2^n}^{b^n})$ is to zero or one. This fact is reflected in our Propositions 3 and 4, which state that the good channels are dense in the unit interval (and so are the bad channels for BECs): A partial polarization with sequence $b^n$ corresponds to an interval with dyadic endpoints, and denseness implies that in this interval there will be both perfect and useless channels. Proposition 6, claiming the self-similarity of the sets of good and bad channels, goes one step further and gives these sets structure: If a channel polarized according to the sequence $b^n a$ is good, then so is the channel polarized according to $b^n 1 a$. Proposition 2 is also of interest in this context: In [14, Prop. 3], we prove that the thresholds $\vartheta(x)$ are symmetric, in the sense that $\vartheta(1-x) = 1 - \vartheta(x)$, a fact that is also visible in Fig. 1.

An obvious extension of our work should deal with the fractal properties of non-binary polar codes. If $q$ is a prime number, then every invertible $\ell \times \ell$ matrix with entries from $\{0, \dots, q-1\}$ is polarizing, unless it is upper-triangular [11, Thm. 5.2]. The $n$-fold Kronecker product of one of these matrices generates $\ell^n$ channels. It should be easily possible to design a function mapping $\{0, \dots, \ell-1\}^\infty$ to $[0, 1]$ (cf. (2)), admitting an analysis similar to the one presented in this paper. Since choosing appropriate polarization matrices for non-binary alphabets is not trivial, we propose to evaluate choices based on the properties of the corresponding *polar fractal* (see Fig. 1). This would, in addition to error probabilities or polarization rates, present another objective for the design of non-binary polar codes.

Whether binary or not, it is presently not clear if our infinite-blocklength results can be carried over to practically relevant finite-length codes. If this was the case, a possible application of our results would be code construction, which requires knowledge about the structure of the set of good channels. Future work shall investigate this issue.

## Acknowledgments

The author thanks Emmanuel Abbe, Princeton University, and Hamed Hassani, ETH Zurich, for fruitful discussions and suggesting material.

## References

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.

[2] J. M. Renes, D. Sutter, and S. H. Hassani, "Alignment of polarized sets," in *Proc. IEEE Int. Sym. on Information Theory Proceedings (ISIT)*, Jul. 2015, pp. 2446–2450, extended version available: arXiv:1411.7925 [cs.IT].

[3] S. Kahraman, E. Viterbo, and M. E. Çelebi, "Folded tree maximum-likelihood decoder for Kronecker product-based codes," in *Proc. Allerton Conf.*, Oct. 2013, pp. 629–636.

[4] E. Arıkan, "A performance comparison of polar codes and Reed-Muller codes," *IEEE Commun. Lett.*, vol. 12, no. 6, pp. 447–449, Jul. 2008.

[5] E. Abbe, "personal communication," Nov. 2011.

[6] M. Taylor, *Measure Theory and Integration*, ser. Graduate studies in mathematics. American Mathematical Soc., 2006.

[7] S. Haghighatshoar and E. Abbe, "Polarization of the Rényi information dimension for single and multi terminal analog compression," in *Proc. IEEE Int. Sym. on Information Theory Proceedings (ISIT)*, Jul. 2013, pp. 779–783.

[8] E. Abbe and Y. Wigderson, "High-girth matrices and polarization," Jan. 2015, arXiv:1501.06528 [cs.IT].

[9] R. Nasser, "Ergodic theory meets polarization. I: An ergodic theory for binary operations," Feb. 2015, arXiv:1406.2943v4 [cs.IT].

[10] ——, "Ergodic theory meets polarization. II: A foundation of polarization theory," Feb. 2015, arXiv:1406.2949v4 [cs.IT].

[11] E. Şaşoğlu, "Polarization and polar codes," *Foundations and Trends® in Communications and Information Theory*, vol. 8, no. 4, pp. 259–381, 2011.

[12] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2009.

[13] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.

[14] B. C. Geiger, "The fractality of polar and Reed-Muller codes," Jun. 2015, in preparation: arXiv:1506.05231 [cs:IT].

[15] S. H. Hassani, K. Alishahi, and R. L. Urbanke, "Finite-length scaling for polar codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5875–5898, Oct. 2014.

[16] S. H. Hassani, S. Korada, and R. Urbanke, "The compound capacity of polar codes," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Sep. 2009, pp. 16–21.

[17] M. El-Khamy, H. Mahdavifar, G. Feygin, J. Lee, and I. Kang, "Relaxed polar codes," Jan. 2015, arXiv:1501.06091 [cs.IT].

[18] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 3rd ed. Chichester: John Wiley & Sons, 2014.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Sampling Algorithms for Lattice Gaussian Codes

Antonio Campello
and Jean-Claude Belfiore
Communications & Electronics Department
Télécom ParisTech
Paris, France
Email: {campello,belfiore}@telecom-paristech.fr

***Abstract*—We consider the problem of sampling a discrete Gaussian distribution whose support is an $n$-dimensional lattice. Fast sampling algorithms for lattices decomposed as a finite union of cosets are proposed. This includes the low dimensional lattices with the best coding gains, their duals, and the 24 dimensional Leech lattice. Our methods are then applied to assess the performance of recent sampling-based codes for the AWGN channel, illustrating the gains of the discrete Gaussian distribution.**

**In the derivation of our algorithms, a number of results concerning the theta series of notable lattices will be discussed, including relations between the theta series and its derivatives to the power and rate of a lattice Gaussian code.**

## I. Introduction

In recent capacity achieving lattice coding schemes for the AWGN channel [7] and for the semantically secure wiretap channel [8], the sent vector is drawn from a discrete Gaussian distribution whose support is a multidimensional Euclidean lattice, as in Figure 1. Other applications of such a distribution include some of the state-of-the-art lattice-based cryptographic models (e.g. [11], [2]), and the generation of information theoretic secure secret keys [9].

A worth element towards practical implementations of all aforementioned schemes is the ability of sampling from the *lattice Gaussian distribution*. This is not a trivial task, and even unidimensional samplers for the lattice $\mathbb{Z}$ have been object of research (see [5], [6] and [2, Sec. 5.1]) - in fact, most multidimensional samplers use them as sub-routines. An obstacle for these algorithms is sampling over Gaussians which are not sufficiently flat, i.e., when the variance parameter $\sigma$ is small or moderate.

In this work we focus on specialized algorithms for lattices commonly used for coding. We propose algorithms for sampling on lattices obtained from constructions A, B, their complex versions, and the density doubling construction. This includes the low dimensional lattices with best coding gain, their duals, and the 24 dimensional Leech lattice. In the derivation of our algorithms, a number of results concerning the theta series of these lattices and their relations to coding parameters will be discussed. Particularly interesting are closed form expressions for the power and rate of a lattice Gaussian code (Prop. 5) in terms of the theta series and its derivative.

Our algorithms output the correct distribution for the specific lattices and any $\sigma$, comparing favorably to universal Markov Chain based algorithms like [14]. For a concrete example, sampling within statistical distance $10^{-3}$ from the centered discrete Gaussian over the Leech lattice $\Lambda_{24}$ in the worst case $\sigma = 1/\sqrt{2\pi}$ requires 13434 iterations (cf. [14, Eq. 26 and Lem. 3]), or $24 \times 13434 = 322416$ calls of an unidimensional $\mathbb{Z}$-sampler. In a huge contrast, the number of calls of the $\mathbb{Z}$-sampler for our tailor-made Leech lattice sampler is 24. This is the same contrast between universal decoders (e.g. the sphere decoder) and specialized decoders for particular lattices (e.g., root lattices, etc.).

The rest of this paper is organized as follows. In Section III we review unidimensional samplers for cosets of the lattice $\mathbb{Z} + c$. In Section IV, we describe a general principle for lattices decomposed as the union of cosets, which is then applied in Sections V-VII to several lattices obtained from codes. In Section VIII, we apply our algorithm to assessing the codeword error probability performance of lattice Gaussian codes [7] for a code based on the Leech lattice $\Lambda_{24}$.

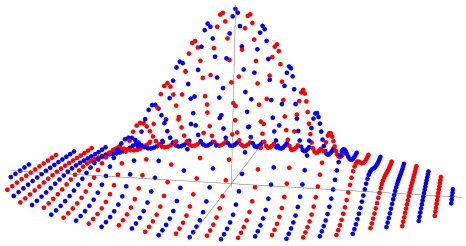


Fig. 1. Distribution obtained from the $A_2$ sampler (Sec V). Blue and red dots correspond to points in $\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}$ and $\mathbb{Z} \oplus \sqrt{3}\mathbb{Z} + (1/2, \sqrt{3}/2)$, respecitively.

## II. Preliminaries and Notation

### A. Lattices

We consider real and complex lattices. A *real lattice* $\Lambda$ is a discrete additive subgroup of $\mathbb{R}^n$, whereas a *complex lattice* is a discrete additive subgroup of $\mathbb{C}^n$. A complex lattice can be always identified with a real lattice in $\mathbb{R}^{2n}$ in a straightforward way by considering the real and imaginary parts. For example, when $\omega = -1/2 + \sqrt{3}/2i$, the lattice of Eisenstein integers $\mathbb{Z}[\omega] = \{a + b\omega : a, b \in \mathbb{Z}\}$ is identified with the real hexagonal plane lattice denote by $A_2$ [4].

There are classical ways of constructing lattices from error correcting codes. Let $\mathcal{C} \subset \mathbb{F}_2^n$ be a linear code and $P_n \subset \mathbb{F}_2^n$ be the parity-check code. By identifying the elements of $\mathbb{F}_2$

as elements of $\mathbb{Z}$ we can write the constructions A and B of $\mathcal{C}$ via the"code-formulas":

$$\Lambda_A(\mathcal{C}) = 2\mathbb{Z}^n + \mathcal{C} \text{ and } \Lambda_B(\mathcal{C}) = 4\mathbb{Z}^n + 2P_n + \mathcal{C}. \quad (1)$$

Alternatively $\Lambda_B(\mathcal{C}) = 2D_n + \mathcal{C}$, where $D_n = \Lambda_A(P_n)$. We can write analogous formulas for complex lattices:

$$\Lambda_A(\mathcal{C}) = \theta\mathbb{Z}[\omega]^n + \mathcal{C} \text{ and } \Lambda_B(\mathcal{C}) = \theta^2\mathbb{Z}[\omega]^n + \theta P_n + \mathcal{C}, \quad (2)$$

where $\theta \in \mathbb{Z}[\omega]$ is a prime with norm $|\theta|^2 = p$ and $\mathcal{C} \subset \mathbb{F}_p$ is a linear code. Again, $\Lambda_B(\mathcal{C}) = \theta\Lambda_n + \mathcal{C}$, where $\Lambda_n = \Lambda_A(P_n)$.

### B. Jacobi Theta Functions

Let $\Lambda$ be a lattice and $\boldsymbol{c}$ a vector in $\mathbb{R}^n$. The *theta series* of $\Lambda + \boldsymbol{c}$ is defined as:

$$\Theta_{\Lambda+\boldsymbol{c}}(\tau) := \sum_{\boldsymbol{y}\in\Lambda+\boldsymbol{c}} e^{-\pi\tau\|\boldsymbol{y}\|^2} = \sum_{\boldsymbol{x}\in\Lambda} e^{-\pi\tau\|\boldsymbol{x}+\boldsymbol{c}\|^2}. \quad (3)$$

The theta series provides useful information about a lattice, such as its minimal norm, kissing number and determinant (for undefined terms, see [4]). In Communications, the theta series bounds the probability of error of a lattice code used for the Gaussian channel (see, e.g. [4, Ch. 3, Eq. (35)]) and for the Gaussian wiretap channel [10] among other applications.

The theta series of all lattices discussed in this paper can be written in terms of the Jacobi theta functions (in what follows, let $q = e^{-\pi\tau}$, $\tau > 0$ and $z = i\tau$):

$$\theta_3(\xi|z) := \sum_{m=-\infty}^{\infty} e^{2im\xi+\pi izm^2} = \sum_{m=-\infty}^{\infty} \cos(2m\xi)e^{\pi izm^2}$$

$$\theta_2(\tau) := \sum_{m=-\infty}^{\infty} q^{(m+1/2)^2}, \theta_3(\tau) := \sum_{m=-\infty}^{\infty} q^{m^2}.$$

Notice that $\theta_2(\tau)$, and $\theta_3(\tau)$ are the theta series of the unidimensional lattice $\mathbb{Z}$ and its shift $\mathbb{Z} + 1/2$, respectively. More generally (see Eq. (2.2.5) of [1]):

$$\Theta_{\mathbb{Z}+c}(\tau) = \sum_{m=-\infty}^{\infty} e^{-\pi\tau(m+c)^2} = \tau^{-2} \sum_{m=-\infty}^{\infty} e^{2\pi imc-\pi m^2/\tau} = \tau^{-2}\theta_3(\pi c|i\tau^{-1}). \quad (4)$$

For numerical aspects and efficient evaluations of the Jacobi series the reader is referred to [1, Ch. 2-3].

### C. Discrete Gaussian Distributions

Define the Gaussian function $\rho_\sigma(\boldsymbol{x}) = e^{\frac{-\|\boldsymbol{x}\|^2}{2\sigma^2}}$ and, for a discrete set $S \subset \mathbb{R}^n$, let

$$\rho_\sigma(S) := \sum_{\boldsymbol{x}\in S} \rho_\sigma(\boldsymbol{x}).$$

The *discrete Gaussian distribution* over $\Lambda+\boldsymbol{c}$ is defined as the distribution with support in $\Lambda + \boldsymbol{c}$, such that the probability of choosing a vector $\boldsymbol{y} \in \Lambda + \boldsymbol{c}$ is proportional to $\rho_\sigma(\boldsymbol{y})$. We denote the probability that a random vector drawn according to the discrete Gaussian distribution is equal to $\boldsymbol{y} \in \Lambda + \boldsymbol{c}$ by

$$D_{\Lambda+\boldsymbol{c},\sigma}(\boldsymbol{y}) := \frac{\rho_\sigma(\boldsymbol{y})}{\rho_\sigma(\Lambda+\boldsymbol{c})} = \frac{\rho_\sigma(\boldsymbol{x}+\boldsymbol{c})}{\Theta_{\Lambda+\boldsymbol{c}}\left(\frac{1}{2\pi\sigma^2}\right)} \quad (5)$$

Some simple but useful properties of $D_{\Lambda+\boldsymbol{c},\sigma}(\boldsymbol{y})$ are stated next:

**Proposition 1.** *The lattice Gaussian distribution satisfies:*

(i) $D_{\alpha(\Lambda+\boldsymbol{c}),\sigma}(\alpha\boldsymbol{y}) = D_{\Lambda+\boldsymbol{c},\sigma/\alpha}(\boldsymbol{y})$.

(ii) $D_{(\Lambda_1+\boldsymbol{c}_1)\oplus(\Lambda_2+\boldsymbol{c}_2)}(\boldsymbol{y}_1,\boldsymbol{y}_2) = D_{\Lambda_1+\boldsymbol{c}_1}(\boldsymbol{y}_1)D_{\Lambda_2+\boldsymbol{c}_2}(\boldsymbol{y}_2)$.

A *lattice Gaussian sampler* is an algorithm that outputs a point $\boldsymbol{y} \in \Lambda + \boldsymbol{c}$ with probability $D_{\Lambda+\boldsymbol{c},\sigma}(\boldsymbol{y})$.

## III. Building Blocks: Gaussians over $\mathbb{Z} + c$

Unidimensional discrete Gaussians are the building blocks for the main multi-dimensional samplers, including the ones described in this paper. Efficient practical samplers over $\mathbb{Z}$ can be found e.g., in [5], [6] and all these methods can be used as subroutines our algorithms. A theoretical method described in [2] shows that it is possible to output the *exact* distribution $D_{\mathbb{Z}+\boldsymbol{c}}$ by calling a continuous Gaussian sampler and using a rejection principle. The expected number of iterations is [2, Sec. 5.1]) (consider $0 < c < 1$ for simplicity):

$$\frac{\rho_\sigma(c) + \rho_\sigma(1-c) + \int_c^\infty \rho_\sigma(x)\mathrm{d}x + \int_{-\infty}^{1-c} \rho_\sigma(x)\mathrm{d}x}{\Theta_{\mathbb{Z}+c}(\frac{1}{2\pi\sigma^2})}. \quad (6)$$

Using Equation (4) we can prove that the expected number of iterations tend to 1 as $\sigma \to 0$ or $\sigma \to 1$. Numericaly evaluations for the probability of acceptance (inverse of expected iterations) are shown in Fig. 2.

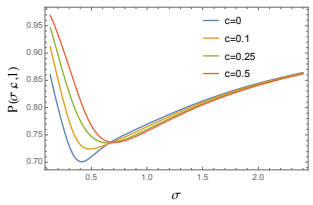



Fig. 2. Acceptance probability as a function of $\sigma$ for fixed $l = 1$ (left) and fixed $c = 0$ (right). In the worst case, $c = 0$ and $\sigma \approx 0.412680$, the average number of iterations is no bigger than 1.426764

## IV. Coset Decompositions

Suppose that the lattice $\Lambda$ can be decomposed as the disjoint union of cosets $\Lambda = \bigcup_{\boldsymbol{c}\in\mathcal{C}} \Lambda' + \boldsymbol{c}$. Let $p_c = D_{\Lambda,\sigma}(\boldsymbol{c} + \Lambda')$ be the probability that a point drawn from a discrete Gaussian in $\Lambda$ lies in the coset $\Lambda' + \boldsymbol{c}$. A general principle for sampling $\Lambda$ is the following:

1) Pick a vector $\boldsymbol{c}$ at random, with probability $p_c$.
2) Pick a vector from $D_{\Lambda'+c,\sigma}$ and output it.

The procedure outputs a point $\boldsymbol{c}+\boldsymbol{x} \in \boldsymbol{c}+\Lambda'$ with the correct probability $p_{\boldsymbol{c}}D_{\Lambda'+\boldsymbol{c},\sigma}(\boldsymbol{c}+\boldsymbol{x}) = D_{\Lambda,\sigma}(\boldsymbol{c}+\boldsymbol{x})$. To apply the general principle we to calculate the probabilities $p_c$, samplers for shifts of the superlattice $\Lambda'$ and a systematic description of the cosets.

The following table is a collection of results on the theta series of some lattices constructed from codes. To facilitate the statements, let

$$\phi_0(\tau) := \theta_3(\tau)\theta_3(3\tau) + \theta_2(\tau)\theta_2(3\tau),$$

$\phi_1(\tau) := \theta_2(\tau)\theta_3(3\tau)+\theta_3(\tau)\theta_2(3\tau) = \frac{1}{2}\theta_2(\tau/4)\theta_2(3\tau/4)$, and

$$\phi_2(\tau) := \frac{\phi_0(\tau/3) - \phi_0(\tau)}{2}$$

($\phi_0$ and $\phi_1$ differ from [4] by a change of variables).

**Proposition 2.** *Let $\Lambda = \Lambda' + \mathcal{C}$ be a lattice obtained from Construction A, B, $A_c$ ($\theta = \sqrt{-3}$ or $\theta = 2$) or $B_c$ ($\theta = \sqrt{-3}$). The probability of each coset, $D_{\Lambda,\sigma}(\Lambda' + \boldsymbol{c})$, depends only on the Hamming weight of $\boldsymbol{c}$.*

*Proof.* See Table A. The result for Constructions A, B, and $A_c$ followsfrom [4, p. 184], [4, Thm. 15, p. 191] and [12]. The result for Construction $B_c$ seems to be unpublished. For Construction $B_c$, $\theta = 2$, a result depending on the *complete weight* of $\boldsymbol{c}$ is proved in [3]. □

| Construction | Theta Series of a Coset | |
|---|---|---|
| A | $\theta_2(4\tau)^w\theta_3(4\tau)^{n-w}$ | |
| B | $(1/2)\theta_2(4\tau)^w\theta_3(4\tau)^{n-w}$ | $w \geq 1$ |
| | $(1/2)\theta_3(4\tau)^n + (1/2)\theta_4(4\tau)^n$ | $w = 0$ |
| $A_c, \theta = 2$ | $\phi_1(4\tau)^w\phi_0(4\tau)^{n-w}$ | |
| $A_c, \theta = \sqrt{-3}$ | $\phi_2(3\tau)^w\phi_0(3\tau)^{n-w}$ | |
| $B_c, \theta = \sqrt{-3}$ | $(1/3)\phi_2(3\tau)^w\phi_0(3\tau)^{n-w}$ | $w \geq 1$ |
| | $(1/3)(\phi_0(3\tau)^n + 2(\phi_0(9\tau) - \phi_2(9\tau))^n)$ | $w = 0$ |

TABLE I
THETA SERIES OF A COSET $\Lambda' + \boldsymbol{c}$, $\mathrm{wt}(\boldsymbol{c}) = w$, FOR SEVERAL CONSTRUCTIONS

In the constructions listed above, item 1) can be split in two parts. Let $A_w$ be the number of codewords of weight $w$ in $\mathcal{C}$. Take any codeword $\boldsymbol{c}_w$ and define $p_w = A_w D_\Lambda(\Lambda' + \boldsymbol{c}_w)$.

1') Choose $w \in \{0, 1, \ldots, n\}$ with probability $p_w$.
1") Choose a codeword in $\mathcal{C}$ uniformly at random over all codewords of weight $w$.

If the code has small cardinality, item 1") can be performed by listing the codewords and organizing them by weight (this process can be done ad-hoc, only once for any number of samples). For some structured higher dimensional codes, we can explore their symmetries to get around the listing (see Section VII for the case of the binary Golay code).

## V. BASE-LATTICES

Of course one can sample over $\mathbb{Z}^n + \boldsymbol{c}$ with complexity $n$ times the complexity of sampling over $\mathbb{Z} + c$. This is enough to sample binary Construction A lattices. As it happens, for sampling Construction B and Construction $A_c$ lattices, we need to know how to sample over shifts of the base lattices $A_2$ and $D_n$.

### A. The lattice $A_2$

An efficient sampler for the $A_2$ lattice is based on the decomposition of $A_2$ as the disjoint union of two translates of a rectangular lattice. This follows from noting that the rectangular lattice $\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}$ is a sublattice of $A_2$ of index 2. We can write

$$A_2 = \left(\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}\right) \bigcup \left(\mathbb{Z} \oplus \sqrt{3}\mathbb{Z} + \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)\right)$$

From this:

$$D_{A_2,\sigma}(\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}) = \frac{\theta_3(\frac{1}{2\pi\sigma^2})\theta_3(\frac{3}{2\pi\sigma^2})}{\theta_3(\frac{1}{2\pi\sigma^2})\theta_3(\frac{3}{2\pi\sigma^2}) + \theta_2(\frac{1}{2\pi\sigma^2})\theta_2(\frac{3}{2\pi\sigma^2})}.$$

Note that if $\sigma$ is small, all mass is concentrated in the origin, hence $D_{A_2,\sigma}(\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}) \approx 1$, and if $\sigma$ is large, the distribution is flat, therefore $D_{A_2,\sigma}(\mathbb{Z} \oplus \sqrt{3}\mathbb{Z}) \approx 1/2$.

### B. The lattice $D_n$ and the shift $D_n + (\alpha, \beta, \ldots, \beta)$

Using Construction A, write $D_n = 2\mathbb{Z}^n + \mathcal{C}$, where $\mathcal{C}$ is a parity-check code $(n, n-1)_2$. For sampling Construction $B$ we need samplers over a shift of $D_n$ lattice, by a vector of the form $(\alpha, \beta, \ldots, \beta)$. We begin by calculating the theta series of the shift via Construction A. Some of these calculations can be found in [4] for $(\alpha, \beta) = (0, 0), (1/2, 1/2)$, and $(1, 0)$.

**Proposition 3.** *Let*

$$\begin{aligned} W_e(X,Y) &= \sum_{i=0}^{\lfloor (n-1)/2 \rfloor} \binom{n-1}{2l} Y^{2l} X^{n-1-2l} \text{ and} \\ W_o(X,Y) &= \sum_{i=1}^{\lfloor (n+1)/2 \rfloor} \binom{n-1}{2l-1} Y^{2l-1} X^{n-2l} \end{aligned} \tag{7}$$

*be the weight enumerators of the vectors in $\mathbb{F}_2^{n-1}$ with even and odd weights, respectively. We have*

$$\begin{aligned} \Theta_{D_n + (\alpha^1, \beta^{n-1})}(q) = \Theta_{\mathbb{Z}+\frac{\alpha}{2}}(q^4) W_e(\Theta_{\mathbb{Z}+\frac{\beta}{2}}(q^4), \Theta_{\mathbb{Z}+\frac{\beta+1}{2}}(q^4)) + \\ + \Theta_{\mathbb{Z}+\frac{\alpha+1}{2}}(q^4) W_o(\Theta_{\mathbb{Z}+\frac{\beta}{2}}(q^4), \Theta_{\mathbb{Z}+\frac{\beta+1}{2}}(q^4)). \end{aligned}$$

The case $\beta = \alpha = 1/2$ is very particular, since the theta series of all cosets $\boldsymbol{c} + (1/2, \ldots, 1/2) + 2\mathbb{Z}^n$ are equal. Hence, a simple sampler in this case is obtained by sampling a word $\boldsymbol{c}$ uniformly at random over all codewords in $\mathcal{C}$ and then sampling over the lattice $\boldsymbol{c} + (1/2, \ldots, 1/2) + 2\mathbb{Z}^n$. This procedure was described in [7]. For $\alpha = \beta \neq 1/2$, let

$$p_{2l} := \binom{n}{2l} \frac{\Theta_{\mathbb{Z}+\frac{\beta+1}{2}}(q^4)^w \Theta_{\mathbb{Z}+\frac{\beta}{2}}(q^4)^{n-w}}{\Theta_{D_n + \beta\boldsymbol{e}}(q)} \tag{8}$$

be the probability that a vector drawn from distribution $D_{D_n,\beta\boldsymbol{e}}$ lies in the coset of a codeword of weight $w$. Algorithm 7 provides a sampling procedure for $D_{D_n,\beta\boldsymbol{e}}$. In the algorithm let $\mathcal{I}_n = \{1, \ldots, n\}$.

The case $\beta \neq \alpha$ is very similar to the previous one, except that we have to distinguish codewords with $c_1 = 0$ and $c_1 = 1$. Given that a point lies in a coset of a codeword of weight $w$, the probability that this codeword has $c_1 = 0$ is given by:

$$p_c = \frac{\Theta_{\mathbb{Z}+\frac{\alpha}{2}}(q^4)\Theta_{\mathbb{Z}+\frac{\beta+1}{2}}(q^4)}{\Theta_{\mathbb{Z}+\frac{\alpha}{2}}(q^4)\Theta_{\mathbb{Z}+\frac{\beta+1}{2}}(q^4) + \frac{2l}{n-2l}\Theta_{\mathbb{Z}+\frac{\alpha+1}{2}}(q^4)\Theta_{\mathbb{Z}+\frac{\beta}{2}}(q^4)}.$$

Sampler$D_n(\beta,\sigma)$

1: Choose a number $l$ from 1 to $\lfloor n/2 \rfloor$ with probability $p_{2l}$
2: Choose uniformly at random a set $\mathcal{J} \subset \mathcal{I}_n$, with size $2l$
3: **For** $j \in \mathcal{J}$
4: $x_j \leftarrow 2\text{Sampler}\mathbb{Z}((\beta+1)/2, \sigma/2)$
5: **For** $j \in \mathcal{I}_n \backslash \mathcal{J}$
6: $x_j \leftarrow 2\text{Sampler}\mathbb{Z}(\beta, \sigma/2)$
7: **Output:** $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

Thus a modification of Algorithm 7 can be done as follows. After picking a weight $2l$ according to the right probability, we throw a biased coin with probability $p_c$ of heads. If the result is heads, then choose $c_1 = 0$, choose a set $\mathcal{J} \subset \mathcal{I}_{n-1}$ with $|\mathcal{I}| = 2l$ to be the support of $(c_2, \ldots, c_n)$ and sample over $2\mathbb{Z}^n + \boldsymbol{c}$, as in steps 3-6 of Algorithm 7. If the result is tails, set $c_1 = 1$ and choose $\mathcal{J} \subset \mathcal{I}_{n-1}$, with $|\mathcal{J}| = 2l - 1$.

## VI. Construction B

As in Construction A, the probability of a coset $2D_n + \boldsymbol{c}$ only depends on the Hamming weight of $\boldsymbol{c}$, since any permutation of coordinates is an automorphism of $D_n$. Let $\overline{D}_n = D_n + (1, 0^{n-1})$. To calculate the probability of a coset in a suitable way to sampling, consider the decomposition, for $1 \le w \le n$:

$$D_n = (D_w \oplus D_{n-w}) \bigcup (\overline{D}_w \oplus \overline{D}_{n-w}), \tag{9}$$

which we refer to as the even-even/odd-odd decomposition. The theta series of $2(D_n + \boldsymbol{c}/2)$ is

$$\begin{aligned}\Theta_{D_n+\frac{\boldsymbol{c}}{2}}(q^4) &= \Theta_{D_w+(\frac{1}{2}^w)}(q^4)\Theta_{D_{n-w}}(q^4) \\ &+ \Theta_{D_w+(\frac{3}{2}^1,\frac{1}{2}^{w-1})}(q^4)\Theta_{D_{n-w}+(1^1,0^{n-w-1})}(q^4)\end{aligned} \tag{10}$$

But all the terms in the rhs of (9) are in the form of Section V-B. Let $p_w = A_w \Theta_{2D_n+(1^w 0^{n-w})}(q)/\Theta\Lambda(q)$ and define $p_{\text{even},w} = D_{2D_n+\boldsymbol{c}}((2D_w + 1^w) \oplus 2D_{n-w})$.

A sampling procedure for Construction B lattices works as follows. Draw a weight $w$ according to the probabilities $p_w$. Draw a codeword $\boldsymbol{c}$ uniformly at random over all codewords with weight $w$ in $\mathcal{C}$. Throw a biased coin with probability $p_{\text{even},w}$ of heads, and if the result is heads sample over the even-even part of decomposition (9) (formally, if $\mathcal{J}$ be the support of $\boldsymbol{c}$, draw $x_{\mathcal{J}}$ as from the distribution over $2D_w + (1^w)$ and $x_{\mathcal{I}^n \backslash \mathcal{J}}$ from the distribution over $2D_{n-w}$). Otherwise, sample over the odd-odd part of $D_n$.

## VII. The Leech Lattice

We provide a sampling algorithm based on the density doubling construction. Let $\mathcal{G}_{24}$ be the $(24, 12, 8)_2$ Golay code. We write half of the Leech lattice as $H_{24} = 2D_{24} + \mathcal{G}_{24}$. Let $\boldsymbol{a} = ((-3/2)^1, (1/2)^{23})$. The (scaled) Leech lattice is:

$$\Lambda_{24} = H_{24} \cup (H_{24} + \boldsymbol{a}). \tag{11}$$

The weight enumerator of $\mathcal{G}_{24}$ is given by

$$W(X,Y) = X^{24} + 759X^{16}Y^8 + 2576X^{12}Y^{12} + 759X^8Y^{16} + Y^{24}.$$

The first half of the Leech is a construction $B$, hence its theta series is

$$\Theta_{H_{24}}(q) = (1/2)W(\theta_2(q^4), \theta_3(q^4)) + (1/2)\theta_4(q^4)^{24}.$$

and a sampler is provided by the techniques in Section V-B. For the other half, an application of Prop. 3 gives us closed forms. A simpler closed form can be derived from the auxiliary series:

$$\alpha(q) = \sum_{m=-\infty}^{\infty} (-1)^m q^{(m+1/4)^2} \text{ and} \tag{12}$$

$$\beta(q) = \Theta_{\mathbb{Z}+1/4}(q) = \frac{\theta_3(q^{1/16}) - \theta_3(q) - \theta_2(q)}{2}$$

**Proposition 4.** *All cosets $2D_{24} + \boldsymbol{c} + \boldsymbol{a}$ have the same theta series, given by:*

$$\Theta_{2D_{24}+\boldsymbol{c}+\boldsymbol{a}}(q) = \frac{\beta(q^4)^{24} - \alpha(q^4)^{24}}{2}. \tag{13}$$

*The other half of the Leech has theta series*

$$2^{11}(\beta(q^4)^{24} - \alpha(q^4)^{24}) = 98304q^8 + 8388608q^{12} + \ldots$$

.

*Proof.* First note that any permutation of coordinates is an automorphism of $2D_{24} + \boldsymbol{a}$. Thus, by permuting coordinates, we obtain a set isometric to $2D_{24} + (1^w, 0^{24-w}) + \boldsymbol{a} \simeq 2D_{24} + \boldsymbol{a}$.For calculating the theta series of $2D_{24} + \boldsymbol{a}$, notice that $\alpha(q)^n$ takes negative sign in the terms associated to the vectors $\boldsymbol{x} + (1/4^{24}) \in \mathbb{Z}_{24} + (1/4^{24})$, where $\boldsymbol{x}$ has odd weight. Hence $\Theta_{D_{24}+(1/4)^{24}}(q) = (\beta(q)^{24} + \alpha(q)^{24})/2$. This, together with the equality

$$(D_{24} + (-3/4)^{24}) \cup (D_{24} + (1/4)^{24}) = \mathbb{Z}^{24} + \boldsymbol{a}/2,$$

gives us the desired form. □

Let $D_{\Lambda_{24},\sigma}(H_{24}) = \Theta_{H_{24}}(q)/\Theta_{\Lambda_{24}}(q)$ be the probability that a point sampled from the distribution in $\Lambda_{24}$ lies in $H_{24}$. The following procedure outputs a point $\boldsymbol{x} \in \Lambda_{24}$ distributed according to $D_{\Lambda_{24},\boldsymbol{\sigma}}$.

Sampler $\Lambda_{24}$

1: Throw a biased coin with prob. $D_{\Lambda_{24},\sigma}(H_{24})$ of heads.
2: **if** the output is heads **then**
3: Sample $\boldsymbol{x} \in H_{24}$ from the Construction B sampler
4: **else**
5: Choose $\boldsymbol{c} \in \mathcal{G}_{24}$ uniformly at random
6: Draw $\boldsymbol{x} \in 2D_{24} + \boldsymbol{a} + \boldsymbol{c}$ using sampler in Sec. V-B.
7: **end if**
8: Output $\boldsymbol{x}$.

Step 3 involves sampling the codewords of $\mathcal{C}$ according to its weight and to probabilities $p_w$ (as in Sec.IV item (1')). In what follows we provide an alternative and more efficient procedure than listing all the $2^12$ codewords.

First, pick a number $w \in \{0, 8, 12, 16, 24\}$ according to probabilities $p_w$. For $w = 0$ and 24 there is nothing to do. For $w = 12$, a simple rejection algorithm works, since

$2576/4096 = 62\%$ of the codewords in $\mathcal{G}_{24}$ have weight 12. Thus a rejection algorithms performs an average of $\approx 1.590062$ iterations. However, the same procedure applied to $w = 8$ yields over 5 iterations with variance 23.813002. A "non-rejection" method uses the well-known fact that minimum weight codewords form a $S(5, 8, 24)$ Steiner system [13]. This means that every word of weight 5 in $\mathbb{F}_2^{24}$ decodes to one and only one codeword of weight 8 in $\mathcal{G}_{24}$. But generating a word of weight 5 in $\mathbb{F}_2^{24}$ is a simple "$k$-out-of-$n$" procedure (pick a subset of size 5 in $\{1, \ldots, 24\}$), and $\mathcal{G}_{24}$ can be decoded very efficiently. For generating codewords of weight $w = 16$, just generate a codeword of weight 8 and add the vector $(1, \ldots, 1) \in \mathcal{G}_{24}$.

## VIII. The Lattice Gaussian Coding Scheme

In the scheme of [7], a lattice point $\boldsymbol{x}$, chosen according to the distribution $D_{\Lambda+\boldsymbol{c},\sigma}$, is sent over a Gaussian channel. It is proven that if $\Lambda$ is an AWGN-good lattice and the dimension $n \to \infty$, transmission rates up to the capacity of the channel can be achieved, provided that snr $> e$. The transmission rate and the power are the entropy and variance per dimension of $D_{\Lambda+\boldsymbol{c},\sigma}$, and asymptotic formulas for these quantities are given in [8, Lem. 6-7]. The following proposition shows closed forms based on the theta series of $\Lambda$ and its derivative.

**Proposition 5.** *The rate (in nats per channel use) and power of a Lattice Gaussian Code are given by*

$$P = \frac{-1}{n\pi}\frac{\Theta'_{\Lambda+\boldsymbol{c}}(\tau)}{\Theta_{\Lambda+\boldsymbol{c}}(\tau)} \text{ and } R = -\frac{\tau}{n}\frac{\Theta'_{\Lambda+\boldsymbol{c}}(\tau)}{\Theta_{\Lambda+\boldsymbol{c}}(\tau)} + \frac{1}{n}\ln\Theta_{\Lambda+\boldsymbol{c}}(\tau), \tag{14}$$

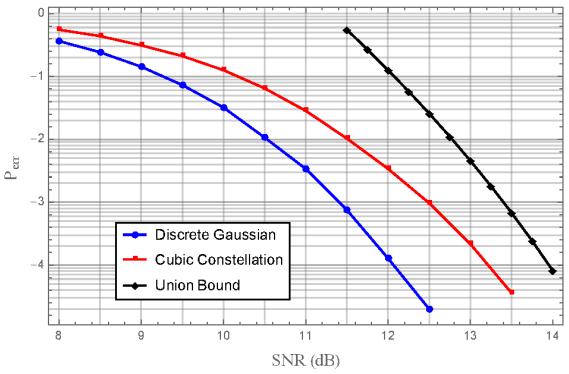
*where* $\tau = 1/2\pi\sigma^2$.

Alternatively, we can write

$$R = \frac{P}{2\sigma^2} + \ln(\Theta_{\Lambda+\boldsymbol{c}}(\tau)^{1/n}) \tag{15}$$

Notice that if we scale the lattice $\Lambda$ *and* $\sigma$ by $\alpha > 0$ we increase the power and keep the rate constant. A corollary of Prop. 5 is that the rate is maximized in the centered distribution, i.e., when $\boldsymbol{c} = 0$.

As an application of the Leech Sampler in Sec VII, we simulate the probability of error of a lattice Gaussian code scheme, in comparison with a cubic shaped Leech codebook. The choice $\sigma = 0.936797$ in this case leads to $R = 1.5$ bits (or 1.039720 nats) and $P = 0.936797$. Simulations were based in $10^6$ samples. Notice that the maximum possible shaping gain of a Voronoi codebook is $\sim 1.53$dB, when the dimension $n \to \infty$.




## Acknowledgment

The work of AC was supported by FAPESP under grant 2014/20602-8. The authors would like to thank Patrick Solé for helpful discussions on theta series.

## References

[1] J. M. Borwein. *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*. Wiley-Interscience, 1986.

[2] Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 575–584. Full Version: arxiv.org/pdf/1306.0281, 2013.

[3] Kok Seng Chua and Patrick Solé. Eisenstein lattices, galois rings, and theta series. *European Journal of Combinatorics*, 25(2):179 – 185, 2004. In memory of Jaap Seidel.

[4] J. H. Conway and N. J. A. Sloane. *Sphere-packings, lattices, and groups*. Springer-Verlag, New York, NY, USA, 1998.

[5] Léo Ducas, Alain Durmus, Tancrède Lepoint, and Vadim Lyubashevsky. Lattice signatures and bimodal gaussians. In Ran Canetti and JuanA. Garay, editors, *Advances in Cryptology – CRYPTO 2013*, volume 8042 of *Lecture Notes in Computer Science*, pages 40–56. Springer Berlin Heidelberg, 2013.

[6] Nagarjun C. Dwarakanath and Steven D. Galbraith. Sampling from discrete gaussians for lattice-based cryptography on a constrained device. *Appl. Algebra Eng., Commun. Comput.*, 25(3):159–180, June 2014.

[7] Cong Ling and J.-C. Belfiore. Achieving awgn channel capacity with lattice gaussian coding. *IEEE Transactions on Information Theory*, 60(10):5918–5929, Oct 2014.

[8] Cong Ling, L. Luzzi, J.-C. Belfiore, and D. Stehle. Semantically secure lattice codes for the gaussian wiretap channel. *Information Theory, IEEE Transactions on*, 60(10):6399–6416, Oct 2014.

[9] Cong Ling, L. Luzzi, and M.R. Bloch. Secret key generation from gaussian sources using lattice hashing. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2621–2625, July 2013.

[10] Frédérique E. Oggier, Patrick Solé, and Jean-Claude Belfiore. Lattice codes for the wiretap gaussian channel: Construction and analysis. *CoRR*, abs/1103.4086, 2011.

[11] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 84–93, New York, NY, USA, 2005. ACM.

[12] N.J.A Sloane. Codes over gf(4) and complex lattices. *Journal of Algebra*, 52(1):168 – 181, 1978.

[13] N.J.A. Sloane and F.J. MacWilliams. *The Theory of Error-Correcting Codes*. North Holland, 1977.

[14] Zheng Wang, Cong Ling, and G. Hanrot. Markov chain monte carlo algorithms for lattice gaussian sampling. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 1489–1493, June 2014.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Fixed-Energy Random Coding with Rescaled Codewords at the Transmitter

Daniel Fehr*, Jonathan Scarlett† and Alfonso Martinez*
* Universitat Pompeu Fabra, Barcelona
† École Polytechnique Fédérale de Lausanne
e-mail: daniel.co@bluewin.ch, jmscarlett@gmail.com, alfonso.martinez@ieee.org

***Abstract*—This paper proposes a new method to reduce the error rate of channel codes over an AWGN channel by renormalizing the codewords to a constant energy before transmission and decoding with the original codebook. Evaluation of the random-coding error exponent reveals that this normalization technique approaches the constant-composition error exponent for certain pairs of rate and signal-to-noise ratio.**

## I. Introduction

Given a general coded modulation scheme over an AWGN channel, we investigate the effect on the error rate of rescaling the transmitted codewords so that their energy is constant. We compare the proposed technique with the standard coded modulation scheme and with constant-composition codes [1], [2] by means of the respective random-coding error exponents.

The comparison reveals that, for low to moderate SNRs, codeword rescaling significantly improves the error exponent of random codes and nearly matches the performance of constant-composition codes. While constant-composition codes, which inherently have fixed codeword energy and are known to achieve the optimum exponent, are difficult to design in practice, codeword rescaling can be applied to existing practical codes without increasing their complexity.

In Sect. II we present the general channel coding model. In Sect. III we add codeword rescaling to the model and derive an achievable random-coding exponent. In Sect. IV we introduce two other well known random-coding exponents and compare them with the new one by numerical evaluation.

## II. Model

The channel input sequence $\boldsymbol{x} = (x_1, \dots, x_n)$ consists of $n$ symbols $x \in \mathcal{X}$, where $\mathcal{X}$ is the symbol constellation. We denote the channel output sequence by $\boldsymbol{y}$ and the channel law, that is, the conditional probability density of receiving sequence $\boldsymbol{y}$ when the sequence $\boldsymbol{x}$ has been sent, by $W^n(\boldsymbol{y} \,|\, \boldsymbol{x})$. We represent random variables by capital letters and their realizations by lowercase letters, e.g. $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ denote random input and output vectors. The channel is memoryless and $W^n(\boldsymbol{y} \,|\, \boldsymbol{x}) = \prod_{i=1}^n W(y_i \,|\, x_i)$, where $W(y \,|\, x)$ is the single-letter channel law.

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.

We focus on a complex-valued additive white Gaussian noise (AWGN) channel. The input alphabet $\mathcal{X}$ is a finite subset of the complex numbers and the channel output is given by

$$y_i = \sqrt{\mathrm{SNR}}\, x_i + z_i\,, \;\; i = 1, \dots, n \tag{1}$$

where $x_i$ are the symbols, $y_i$ the channel output values and SNR is the signal-to-noise ratio. The noise values $z_i$ are drawn from a circularly-symmetric complex-valued Gaussian random variable with zero mean and unit variance. Therefore, the symbol channel transition probability is given by

$$W(y \,|\, x) = \frac{1}{\pi} e^{-|y - \sqrt{\mathrm{SNR}}\, x|^2}. \tag{2}$$

The empirical average symbol energy of channel input sequence $\boldsymbol{x}$ is $\mathcal{E}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n |x_i|^2$. Let $Q$ be a distribution on the symbols in $\mathcal{X}$. We require that the constellation $\mathcal{X}$ is chosen such that the identities $\mathbf{E}[X] = 0$ and $\mathbf{E}[|X|^2] = 1$ hold, where the expectations are with respect to $Q$.

We use $\mathbf{P}[\mathcal{A}]$ to denote the probability of an event $\mathcal{A}$, and $\mathbf{E}[\cdot]$ is the expectation operator. We denote the Kullback-Leibler Divergence as $D(P\|Q)$, the set of all compositions on length-$n$ sequences drawn from $\mathcal{X}$ as $\mathcal{P}_n$, the set of all distributions on $\mathcal{X}$ as $\mathcal{P}$ and the image of set $\mathcal{A}$ under the function $f$ as $f[\mathcal{A}]$.

We denote the channel code by $\mathcal{C}_n$. The code consists of $M$ codewords $\boldsymbol{x}$, i.e. $\mathcal{C}_n = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\}$, and the corresponding code is said to be an $(n, M)$ block code of block length $n$. The encoder assigns to each message $\mathsf{m} \in \{1, \dots, M\}$ a codeword $\boldsymbol{x}^{(\mathsf{m})}$ from the codebook $\mathcal{C}_n$. We assume that the message $\mathsf{m}$ is drawn according to a uniform distribution. The rate of a code is defined as $R \triangleq \frac{\log_2 M}{n}$.

The decoder outputs an estimated message $\hat{\mathsf{m}}$ according to a maximum-metric rule

$$\hat{\mathsf{m}} = \arg\max_{i \in \mathcal{M}} q^n(\boldsymbol{x}^{(i)}, \boldsymbol{y}), \tag{3}$$

where $q^n$ denotes the metric that the decoder uses to estimate which message $\mathsf{m}$ has been sent. We focus on metrics $q^n$ that can be expressed in terms of the letter metric $q$ as in $q^n(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^n q(x_i, y_i)$. Further, we require $q$ to be positive.

An error occurs if the decoder's estimate differs from the sent message, i.e. $\hat{\mathsf{m}} \neq \mathsf{m}$. The error probability of a code $\mathcal{C}_n$ is $p_e(\mathcal{C}_n) = \frac{1}{M}\sum_{\mathsf{m} \in \mathcal{M}} \mathbf{P}[\hat{\mathsf{m}} \neq \mathsf{m} \,|\, \mathsf{m}]$, and we equivalently write $p_e(n, M) = p_e(\mathcal{C}_n)$. Finally, an error exponent $E(Q, R)$ is

said to be achievable if there exists a sequence of $(n, M)$-codes $\mathcal{C}_n$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log p_e(\mathcal{C}_n) \geq E(Q,R). \tag{4}$$

## III. Codeword Rescaling

### A. Rescaling setup

Let a rescaler $\eta$ be a block that performs the operation

$$\eta(\boldsymbol{x}) = \mathcal{E}(\boldsymbol{x})^{-\frac{1}{2}}\boldsymbol{x} \tag{5}$$

on a codeword $\boldsymbol{x}$; that is, it renormalizes the codeword $\boldsymbol{x}$ such that the empirical codeword energy is $\mathcal{E}\big(\eta(\boldsymbol{x})\big) = 1$.

We consider a coded modulation scheme that consists of an encoder and a rescaler $\eta$. The encoder maps message $\mathsf{m}$ into codeword $\boldsymbol{x}^{(\mathsf{m})} = \phi(\mathsf{m})$ using codebook $\mathcal{C}$, the rescaler outputs an energy-normalized version $\tilde{\boldsymbol{x}}^{(\mathsf{m})}$ of the codeword $\boldsymbol{x}^{(\mathsf{m})}$. The rescaling operation can be thought as part of the encoder's codebook, so that we use a code $\tilde{\mathcal{C}}$ with rescaled codewords $\tilde{\boldsymbol{x}}$ that consist of symbols $\tilde{x}$ from an expanded constellation $\tilde{\mathcal{X}}$.

The decoder outputs the estimate $\hat{\mathsf{m}}$ under the original codebook $\mathcal{C}$ by maximizing the metric $q^n$. With the choice of $q^n(\boldsymbol{x},\boldsymbol{y}) = W^n(\boldsymbol{y}\,|\,\boldsymbol{x})$, we have an instance of mismatched decoding, since the decoder does not account for the rescaling operation neither in the codebook $\mathcal{C}$ nor in the decoding metric. We consider a slightly more general choice given by $q^n(\boldsymbol{x},\boldsymbol{y}) \triangleq W^n(\boldsymbol{y}\,|\,\beta\boldsymbol{x})$, where $\beta$ may be optimized to minimize the error probability; however, it cannot depend on the codeword, and hence it cannot be used to undo the rescaling. Note that for a practical code, $\beta$ is fixed before deployment and such a decoding metric can be implemented without additional computational complexity.

We can build an equivalent model for the rescaling setup by removing the rescaling block from the transmitter and reinterpreting it as a channel property. With this model, the scaling function leads to a new channel law

$$\tilde{W}^n(\boldsymbol{y}\,|\,\boldsymbol{x}) = W^n(\boldsymbol{y}\,|\,\eta(\boldsymbol{x})). \tag{6}$$

Note that $\tilde{W}^n$ does not represent a memoryless channel.

### B. Scaling exponent

We study the i.i.d. random-coding error probability. We consider an ensemble of codebooks with block length $n$ and $M = 2^{nR}$ codewords. The ensemble consists of codebooks whose codewords $\boldsymbol{x}^{(i)}$, $i = 1,\dots,M$ are randomly generated. A codeword $\boldsymbol{x}^{(i)} = (x_1,\dots,x_n)$ at entry $i$ in the random codebook is generated by drawing its $n$ symbols according to the distribution $Q(x)$. We are interested in the achievable random-coding exponent $E_{\mathrm{r}}^{\mathrm{scl}}(Q,R)$ of the ensemble average of the error probability $\overline{p_e}(n,M) = \sum_{\mathcal{C}_n}\mathbf{P}[\mathcal{C}_n]p_e(\mathcal{C}_n)$.

*Theorem 1 (Scaling random-coding exponent):* The random-coding error probability in a rescaling setup satisfies

$$\lim_{n\to\infty} -\frac{1}{n}\log_2\overline{p_e}(n,2^{nR}) \geq E_{\mathrm{r}}^{\mathrm{scl},\beta}(Q,R), \tag{7}$$

where the scaling exponent is defined as

$$E_{\mathrm{r}}^{\mathrm{scl},\beta}(Q,R) \triangleq \sup_{\beta\geq 0}\min_{P\in\mathcal{P}}\sup_{\substack{\rho\in[0,1]\\ s\geq 0}}\big\{E_0^{\mathrm{scl},\beta}(Q,\rho,s,P) - \rho R\big\}, \tag{8}$$

and the corresponding $E_0^{\mathrm{scl},\beta}$ is defined as

$$E_0^{\mathrm{scl},\beta}(Q,\rho,s,P) \triangleq D(P\|Q) - \mathbf{E}\left[\log_2\mathbf{E}\left[\frac{\mathbf{E}[q(\bar{X},Y)^s\,|\,Y]^\rho}{q(X,Y)^{\rho s}}\,\middle|\,X\right]\right] \tag{9}$$

and the expectations are with respect to

$$(X,Y,\bar{X}) \sim P(x)W\big(y\,|\,\mathcal{E}(P)^{-\frac{1}{2}}x\big)Q(\bar{x}). \tag{10}$$

*Proof:* For the ensemble of codebooks with $M$ codewords of length $n$, chosen according to random-coding distribution $Q^n$, transmitted over a channel described by the arbitrary channel law $W^n$ and decoded according to the metric $q^n$, the average ensemble error probability is bounded by the random-coding union (RCU) bound [4], [5]

$$\mathrm{rcu}(n,M) = \mathbf{E}\left[\min\left\{1,(M-1)\,\mathbf{P}\left[\frac{q^n(\bar{\boldsymbol{X}},\boldsymbol{Y})}{q^n(\boldsymbol{X},\boldsymbol{Y})}\geq 1\,\middle|\,\boldsymbol{X},\boldsymbol{Y}\right]\right\}\right], \tag{11}$$

where $(\boldsymbol{X},\boldsymbol{Y},\bar{\boldsymbol{X}}) \sim Q^n(\boldsymbol{x})W^n(\boldsymbol{y}\,|\,\boldsymbol{x})Q^n(\bar{\boldsymbol{x}})$.

Weakening (11) by replacing the function $z\mapsto\min\{1,z\}$ with the function $z\mapsto z^{\rho(\boldsymbol{x})}$ and using Markov's inequality with parameter $s(\boldsymbol{x})$, leads to the definition of a parametrized upper bound on the the RCU bound

$$\mathrm{rcu}_{\rho,s}(n,M) \triangleq \mathbf{E}\left[\left(\frac{\mathbf{E}[q^n(\bar{\boldsymbol{X}},\boldsymbol{Y})^{s(\boldsymbol{X})}\,|\,\boldsymbol{X},\boldsymbol{Y}]}{q^n(\boldsymbol{X},\boldsymbol{Y})^{s(\boldsymbol{X})}(M-1)^{-1}}\right)^{\rho(\boldsymbol{X})}\right] \tag{12}$$

that holds for all pairs of functions $\big(\rho(\boldsymbol{x}),s(\boldsymbol{x})\big)$ such that $\rho[\mathcal{X}^n]\subseteq[0,1]$ and $s[\mathcal{X}^n]\subseteq[0,\infty)$.

We further weaken (12) by applying $M-1\leq 2^{nR}$, use the random-coding distribution $Q_{\mathrm{iid}}^n = \prod_{i=1}^n Q(x_i)$ and the equivalent scaling channel model (6), to obtain

$$\overline{p_e}(n,M) \leq \sum_{\boldsymbol{x}\in\mathcal{X}^n} 2^{n\rho(\boldsymbol{x})R}Q_{\mathrm{iid}}^n(\boldsymbol{x})f^n(\rho(\boldsymbol{x}),s(\boldsymbol{x}),\boldsymbol{x}), \tag{13}$$

where

$$f^n(\rho,s,\boldsymbol{x}) \triangleq \int_{\boldsymbol{y}} W\big(\boldsymbol{y}\,|\,\mathcal{E}(\boldsymbol{x})^{-\frac{1}{2}}\boldsymbol{x}\big)\left(\sum_{\bar{\boldsymbol{x}}\in\mathcal{X}^n}Q_{\mathrm{iid}}^n(\bar{\boldsymbol{x}})\frac{q^n(\bar{\boldsymbol{x}},\boldsymbol{y})^s}{q^n(\boldsymbol{x},\boldsymbol{y})^s}\right)^\rho \mathrm{d}\boldsymbol{y}. \tag{14}$$

We split the outer summation over the channel-input sequences in (13) into summations over sequences $\boldsymbol{x}$ of composition $P = \hat{P}_{\boldsymbol{x}}$ and obtain

$$\overline{p_e}(n,M) \leq \sum_{P\in\mathcal{P}_n} 2^{n\rho(P)R}\sum_{\boldsymbol{x}\in T(P)}Q_{\mathrm{iid}}^n(\boldsymbol{x})f^n(\rho(P),s(P),\boldsymbol{x}). \tag{15}$$

We also reduced the degrees of freedom for the parameters $\rho(\boldsymbol{x})$ and $s(\boldsymbol{x})$ such that they only depend on the composition $P = \hat{P}_{\boldsymbol{x}}$. This simplifies the analysis.

The codeword average symbol energy only depends on the codewords composition. Hence we write $\mathcal{E}(\boldsymbol{x}) = \mathcal{E}(\hat{P}_{\boldsymbol{x}})$ and use it in (14) to obtain $f^n(\rho, s, P, \boldsymbol{x})$. This, together with the product nature of decoding metric, channel law and $Q_{\text{iid}}^n$, allows us to factor $f^n$ as $f^n(\rho, s, P, \boldsymbol{x}) = \prod_{i=1}^n f(\rho, s, P, x_i)$, where

$$f(\rho, s, P, x) \triangleq \int_y W\big(y \,|\, \mathcal{E}(P)^{-\frac{1}{2}} x\big) \left( \sum_{\bar{x} \in \mathcal{X}} Q(\bar{x}) \frac{q(\bar{x}, y)^s}{q(x, y)^s} \right)^{\rho} \mathrm{d}y, \tag{16}$$

by invoking the distributive law. For the product in $f^n$, only the composition of $\boldsymbol{x}$ is relevant, that is, it can be expressed independent of $\boldsymbol{x}$ as $f^n(\rho, s, P) = \prod_{x \in \mathcal{X}} f(\rho, s, P, x)^{nP(x)}$. We use this form in (15) to obtain

$$\overline{p_e}(n, M) \leq \sum_{P \in \mathcal{P}_n} \sum_{\boldsymbol{x} \in T(P)} \frac{2^{n\rho(P)R} \prod_{x \in \mathcal{X}} f(\rho(P), s(P), P, x)^{nP(x)}}{2^{n\left(D(P\|Q) + H(P)\right)}}, \tag{17}$$

where we expressed the codeword probability in terms of its composition as in $Q_{\text{iid}}^n(\boldsymbol{x}) = 2^{-n\left(D(\hat{P}_{\boldsymbol{x}}\|Q) + H(\hat{P}_{\boldsymbol{x}})\right)}$ [6].

Since the summand is independent of the codeword $\boldsymbol{x}$ in (17), we can upper-bound the summation over the codewords by using the bound on the number of sequences in a composition class $|T(P)| \leq 2^{nH(P)}$ [6]. Doing some rearrangements and bringing all terms on a common exponent base, we obtain

$$\overline{p_e}(n, M) \leq \sum_{P \in \mathcal{P}_n} 2^{-n\xi(\rho(P), s(P), P, R)}, \tag{18}$$

where

$$\xi(\rho, s, P, R) \triangleq D(P\|Q) - \sum_{x \in \mathcal{X}} P(x) \log_2 f(\rho, s, P, x) - \rho R. \tag{19}$$

A simple upper bound on a sum is obtained by fixing its summands to the largest one, that is $\sum_{a \in \mathcal{A}} a \leq |\mathcal{A}| (\max_{a \in \mathcal{A}} a)$. We weaken (18) with this bound and the fact that the number of compositions is bounded by $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$ [6] and get

$$\overline{p_e}(n, M) \leq (n+1)^{|\mathcal{X}|}\, 2^{-n\left[\min_{P \in \mathcal{P}_n} \xi(\rho(P), s(P), P, R)\right]}. \tag{20}$$

The bound (20) is achievable for any pair of parameters $\big(\rho(P), s(P)\big)$, which we exploit by choosing them such that the bound gets as tight as possible. By definition of the $\min$ and $\sup$ operators, this is the case when we replace $\xi$ with

$$\xi^*(P, R) = \sup_{0 \leq \rho \leq 1, s \geq 0} \xi\big(\rho, s, P, R\big). \tag{21}$$

That is, place the supremum inside the minimum operator.

Finally, we observe the sub-exponential factor in (20) which suggests to transform the inequality as

$$-\frac{1}{n} \log_2 \overline{p_e}(n, 2^{nR}) \geq \min_{P \in \mathcal{P}_n} \xi^*(P, R) - |\mathcal{X}| \frac{\log_2(n+1)}{n} \tag{22}$$

and take the limit

$$\lim_{n \to \infty} -\frac{1}{n} \log_2 \overline{p_e}(n, 2^{nR}) \geq \min_{P \in \mathcal{P}} \xi^*(P, R). \tag{23}$$

■

### C. Swapped exponent

An inspection by example of the optimization parameters in (8) shows that $E_0^{\text{scl},\beta}$ is not convex in $P$ for fixed values of $\rho$, $s$ and $R$, which complicates its numerical computation. Further, note that it is crucial to find the true $P^*$ to guarantee the achievability of $E_{\text{r}}^{\text{scl},\beta}$, since $P^*$ is the result of a minimization over a variable that is not an arbitrary parameter. For these reasons, we introduce a lower bound on (8) that is computationally tractable.

*Theorem 2 (Swapped scaling random-coding exponent):* The swapped random-coding error exponent, defined as

$$E_{\text{r}}^{\text{swp},\beta}(Q, R) \triangleq \sup_{\beta \geq 0} \min_{\epsilon \in \mathcal{S}} \sup_{\substack{\rho \in [0,1] \\ s \geq 0}} \big\{ E_0^{\text{swp},\beta}(Q, \rho, s, \epsilon) - \rho R \big\} \tag{24}$$

with

$$E_0^{\text{swp},\beta}(Q, \rho, s, \epsilon) \triangleq \min_{P \in \mathcal{P}_\epsilon} \left\{ D(P\|Q) - \mathbf{E}\left[ \log_2 \mathbf{E}\left[ \frac{\mathbf{E}[q(\bar{X}, Y)^s \,|\, Y]^\rho}{q(X, Y)^{\rho s}} \,\middle|\, X \right] \right] \right\}, \tag{25}$$

is a lower bound on $E_{\text{r}}^{\text{scl},\beta}$, where $\epsilon$ is optimized over the interval $\mathcal{S} = [\min_{x \in \mathcal{X}} |x|^2, \max_{x \in \mathcal{X}} |x|^2]$, the set $\mathcal{P}_\epsilon$ is $\{P \in \mathcal{P} \mid \mathcal{E}(P) = \epsilon\}$ and the expectations are according to $(X, Y, \bar{X}) \sim P(x) W(y \,|\, \epsilon^{-\frac{1}{2}} x) Q(\bar{x})$.

*Proof:* First, let the energy level $\epsilon$ denote the expected symbol energy with respect to $P$ or equivalently $\epsilon = \mathcal{E}(P)$. We observe that the probability simplex $\mathcal{P}$ can be partitioned into disjoint subsets $\mathcal{P}_\epsilon$, where a subset consists of all distributions $P$ that obtain the energy level $\epsilon$, that is, we have $\mathcal{P} = \bigcup_{\epsilon \in \mathcal{S}} \mathcal{P}_\epsilon$ where $\mathcal{P}_\epsilon = \{P \in \mathcal{P} \mid \mathcal{E}(P) = \epsilon\}$ and $\mathcal{S} = [\min_{x \in \mathcal{X}} |x|^2, \max_{x \in \mathcal{X}} |x|^2]$. The subsets $\mathcal{P}_\epsilon$ are compact and convex since they are intersections of two hyperplanes and the closed positive orthant [7]. In correspondence with these observations, we rephrase the scaling exponent (8) in terms of energy levels and subsets as

$$E_{\text{r}}^{\text{scl},\beta}(Q, R) = \min_{\epsilon \in \mathcal{S}} \min_{P \in \mathcal{P}_\epsilon} \sup_{\substack{\rho \in [0,1] \\ s \geq 0}} \big\{ E_0^{\text{scl},\beta}(Q, \rho, s, P) - \rho R \big\}. \tag{26}$$

We swap the order of the inner two optimization operators and define $E_{\text{r}}^{\text{swp},\beta}$ and $E_0^{\text{swp},\beta}$ as in (24) and (25) respectively. Observing that the minimax inequality

$$\sup_x \min_y f(x, y) \leq \min_y \sup_x f(x, y) \tag{27}$$

holds we conclude that swapping the order of optimization results in a lower bound on (8). ■

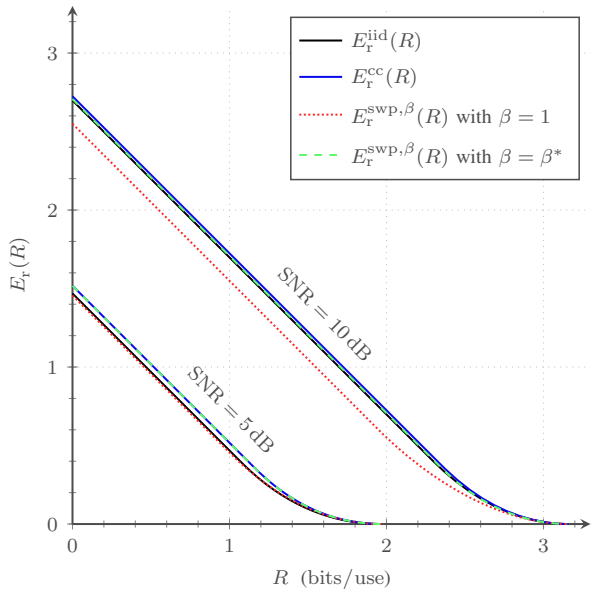



Fig. 1. Random-coding error exponents for 16QAM and fixed SNRs of 5 dB and 10 dB, respectively.

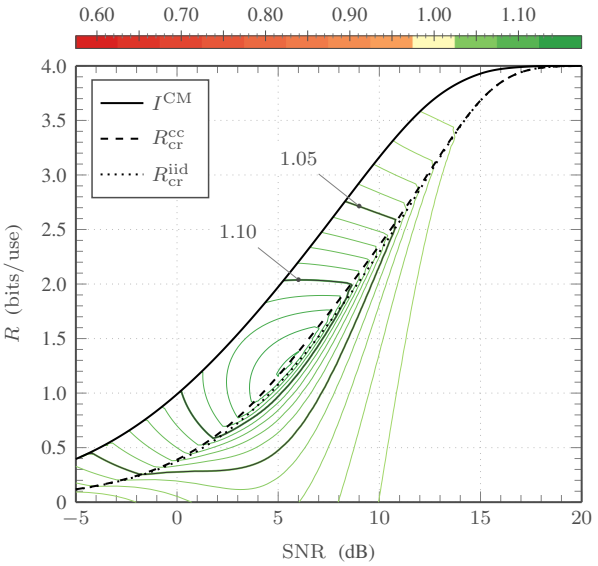



Fig. 2. Relative constant-composition exponent $\underline{E}_{\mathrm{r}}^{\mathrm{cc}}$ for 16QAM.

## IV. Numerical Results

We are interested in how the error exponent of channel-coding with rescaling performs with respect to setups with the iid and constant-composition exponents [1], [2].

For a discrete-input continuous-output memoryless channel with ML decoder and input distribution $Q$, the iid exponent is given by [1]

$$E_{\mathrm{r}}^{\mathrm{iid}}(Q,R) = \max_{\rho\in[0,1]} E_0^{\mathrm{iid}}(\rho) - \rho R, \tag{28}$$

where

$$E_0^{\mathrm{iid}}(Q,\rho) \triangleq -\log_2 \int_y \mathbf{E}\Big[W(y\,|\,X)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \mathrm{d}y \tag{29}$$

is the Gallager function [3]. The expectation is taken with respect to $Q$.

For a discrete-input continuous-output memoryless channel with ML decoder and input distribution $Q$, the constant-composition exponent is given by [4]

$$E_{\mathrm{r}}^{\mathrm{cc}}(Q,R) = \max_{\rho\in[0,1]} E_0^{\mathrm{cc}}(\rho) - \rho R, \tag{30}$$

where

$$E_0^{\mathrm{cc}}(Q,\rho) \triangleq \sup_{a(\cdot)} -\log_2 \int_y \mathbf{E}_Q\Big[\, W(y\,|\,X)^{\frac{1}{1+\rho}} e^{a(X)-\phi_a}\,\Big]^{1+\rho} \mathrm{d}y \tag{31}$$

and $\phi_a = \mathbf{E}_Q[\,a(X)\,]$, and the optimization in (31) is over all real-valued functions $a$. The expectations are taken with respect to $(X,Y,\bar{X}) \sim Q(x)W(y\,|\,x)Q(\bar{x})$.

Fig. 1 compares the error exponents (28), (30) and (24) at SNRs 5 dB and 10 dB. It suggests that $E_{\mathrm{r}}^{\mathrm{swp},\beta}(R)$ approaches the constant-composition exponent.

For a more detailed comparison, we introduce error exponent ratios with the iid exponent as the baseline. We call these the relative constant-composition exponent and the relative swapped scaling exponent, respectively given by

$$\underline{E}_{\mathrm{r}}^{\mathrm{cc}}(R) \triangleq \frac{E_{\mathrm{r}}^{\mathrm{cc}}(R)}{E_{\mathrm{r}}^{\mathrm{iid}}(R)} \quad \text{and} \quad \underline{E}_{\mathrm{r}}^{\mathrm{swp},\beta}(R) \triangleq \frac{E_{\mathrm{r}}^{\mathrm{swp},\beta}(R)}{E_{\mathrm{r}}^{\mathrm{iid}}(R)}. \tag{32}$$

Our figures show the relative random-coding exponents as a function of both SNR and the rate $R$ in a single contour plot for coded modulation with a 16QAM constellation and a uniform distribution $Q$. The figures also depict the mutual information (MI) as a solid line and the critical rate[1] $R_{\mathrm{cr}}^{\mathrm{iid}}$ of the iid exponent (resp. $R_{\mathrm{cr}}^{\mathrm{cc}}$ of the constant-composition exponent) as a dashed (resp. dotted) line.

*1) Constant-composition exponent:* The constant-composition exponent's gain with respect to the iid exponent provides a benchmark to assess the performance of the codeword rescaling. The relative exponent is shown in Fig. 2. The figure reveals that the constant-composition exponent exhibits the largest gains for low to moderate SNRs. At high SNRs above 12 dB, it is roughly equal to the iid exponent. A similar tendency is seen at very low SNRs. For our comparison, the most relevant part of the contour plot is located between the critical rates and the MI, since below the critical rate expurgated versions of the exponents lead to better bounds. The maximal gain with respect to the iid exponent is roughly 13 % and occurs around SNR $= 5.5\,$dB and $R =$1.25 bits per channel use.

*2) Swapped exponent with $\beta = 1$:* We discuss first the relative swapped exponent with $\beta = 1$, which corresponds to

[1]The critical rate is defined as the largest rate $R$ at which the optimal $\rho$ in the exponent optimization is one.

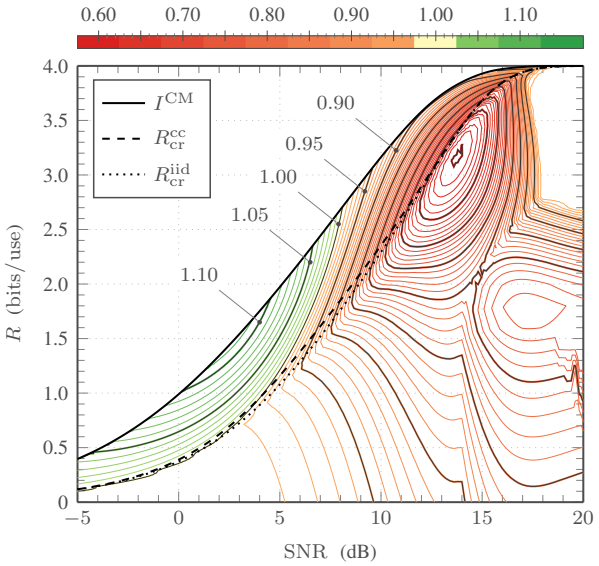



Fig. 3. Relative swapped exponent $\underline{E}_{\mathrm{r}}^{\mathrm{swp},\beta}$ with $\beta = 1$ for 16QAM.

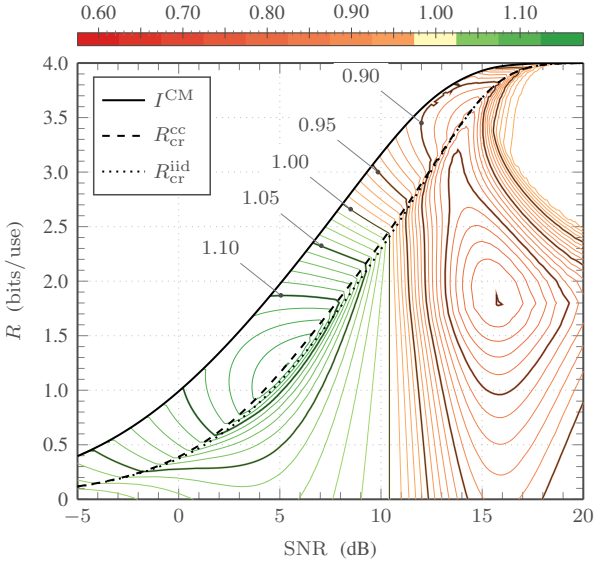



Fig. 4. Relative swapped exponent $\underline{E}_{\mathrm{r}}^{\mathrm{swp},\beta}$ for 16QAM.

the standard decoding rule of maximizing $W^n$. Fig. 3 shows this relative exponent. We observe two regimes in the relevant region between the critical rate and the capacity. Below about 7 dB, there is a gain with respect to the iid exponent, whereas above 7 dB the swapped exponent is worse.

In the low-SNR region and close to the capacity, the swapped exponent achieves similar gains as the constant-composition exponent, namely, about 10 % gain at 3 dB and 1.5 bits per channel use. However, it falls short as the rate approaches the critical rate. For example, there is no gain at 5.5 dB and 1.25 bits per channel use, the point where the constant-composition exponent achieves the highest gain. At high SNRs, especially around 14 dB, Fig. 3 unveils a large loss compared to the iid exponent. In this region, we observe not even at capacity a gain and the loss is up to 35 % close to the critical rate.

*3) Swapped exponent with optimal $\beta$:* Optimizing over $\beta$ to adapt the mismatched metric leads to respectable gains in some regions of the SNR-rate plane, as we can see in Fig. 4. At high SNRs, the swapped exponent with optimized $\beta$ exhibits a considerable improvement compared to the swapped exponent with $\beta = 1$, even though it is still below the iid exponent.

Most notably, for low SNRs, the swapped exponent with optimized $\beta$ achieves more than 90 % of the gain that the constant-composition exponent achieves with respect to the iid exponent. Especially in regions farther away from capacity, the fully optimized exponent shows a large improvement with respect to the one with $\beta = 1$.

## References

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

[2] A. G. D'yachkov, "Bounds on the average error probability for a code ensemble with fixed composition," *Prob. Inf. Transm.*, vol. 16, no. 4, pp. 3–8, 1980.

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

[4] J. Scarlett, A. Martinez and A. Guillén i Fàbregas, "Mismatched Decoding: Error Exponents, Second-Order Rates and Saddlepoint Approximations," *IEEE Trans. Inf. Theory,* vol. 60, no. 5, pp. 2647–2666, May 2014.

[5] 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.

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

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

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# Quantization and LLR Computation for Physical Layer Security

Oana Graur, Nazia Islam, Alexandra Filip, and Werner Henkel

Jacobs University Bremen
Electrical Engineering and Computer Science
Bremen, Germany
Emails: o.graur@jacobs-university.de, n.islam@jacobs-university.de, alexandra.filip@dlr.de, w.henkel@jacobs-university.de

***Abstract*—The problem of key reconciliation based on Low-Density Parity-Check (LDPC) codes and Slepian-Wolf coding for physical layer key generation is investigated. When using the channel-state-information (CSI) of a reciprocal wireless channel for key generation between two legitimate users, independent noise components, quantization, and synchronization errors at the end nodes give rise to key differences that need to be corrected by sending side information. We provide a comparison of three different quantization schemes in terms of key disagreement rate and output probability distributions and present the log-likelihood formulations required by a soft decision LDPC decoder to perform key reconciliation, for the investigated quantization methods.**

## I. Introduction

While computational security algorithms usually reside in upper protocol layers and rely on the assumption of limited processing capabilities of a potential eavesdropper, physical-layer key generation aims at providing secrecy in the more information-theoretic sense, as introduced by Shannon [1]. Thus, by sharing a previously known secret key, such as a one-time pad, two legitimate users, Alice and Bob, are able to exchange an encrypted message through an unsafe public channel, without leaking any information to a potential eavesdropper, Eve. As long as Alice and Bob share a secret common source of randomness from which they can generate a long uniformly distributed secret key, perfect secrecy is achieved, meaning that Eve has the same chances of guessing the original message with or without the ciphertext, or, in more theoretical terms, the eavesdropper's equivocation is equal to the entropy of the message. It soon became clear that such a common source of randomness could be provided by the fluctuating and reciprocal nature of the wireless medium and that the channel-state information (CSI) can be measured by both Alice and Bob and used to generate one-time pads, thus eliminating the problem of previous key distribution.

Since most wireless transmission standards, such as 802.11, Bluetooth, WiMAX, ZigBee, employ time division duplexing (TDD), probing in consecutive short time slots the forward and the reverse channel provides the possibility of obtaining nearly identical CSI on both sides. Such a method of key generation, solely based on the reciprocity property of wireless TDD systems, besides solving the problem of key distribution, comes with the significant benefit that it does not require the legitimate channel between Alice and Bob to have an SNR advantage over the eavesdropper's channels, such as [8], [9], nor does it assume Alice and Bob to have information about the channels to Eve. However, one important aspect that we address here is that due to independent noise on both ends, different transceiver circuitry, and quantization errors, key mismatches are very likely to occur, leading to the necessity of a key reconciliation scheme.

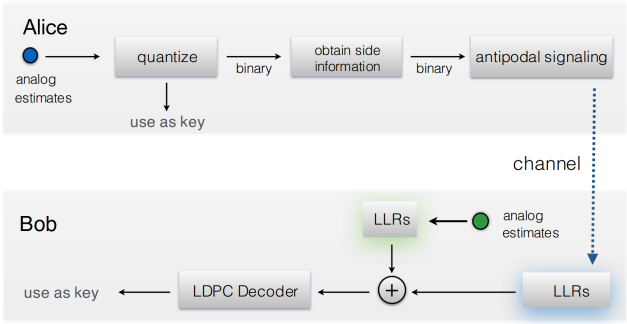



Fig. 1. System model - key reconciliation based on LDPC codes

The current paper is structured as follows: Section II offers an overview of the system model along with channel characterization aspects. In Section III, the impact of codebook sizes and possible quantization schemes is discussed in terms of key disagreement rates for the case when no key reconciliation takes place between the users. An exact formulation of the log-likelihood ratios as required, e.g., by LDPC decoding, is detailed in Section IV, for the quantization schemes analyzed in the previous section.

## II. System Description

A point from the channel distribution is measured by both Alice and Bob, disturbed by different noise components. The analog value obtained by Alice, is quantized and assumed to be correct. Employing Slepian-Wolf coding [13], Alice compresses her vector of quantized key symbols and sends Bob additional side information (parity or syndrome bits), as illustrated in Fig. 1. Bob obtains his own vector of channel estimates, along with the side information, possibly corrupted, sent by Alice, and proceeds to computing the log-likelihood

ratios required, e.g., for an LDPC decoder, in order to obtain the exact same key Alice generated after her quantization step. Although the reconciliation scheme presented in Fig. 1 shows an implementation with LDPC codes, the log-likelihood ratios presented in Section IV can be utilized by any soft decision non-binary decoder.

For key generation based on the channel-state information (CSI), a few conditions must be ensured. First, the channel has to be *reciprocal*, that is, if we denote by $h_B$ a forward channel sample from Alice to Bob, and by $h_A$ the corresponding reverse channel sample from Bob to Alice, then $h_A = h_B$. However, their estimates of the channel, $\hat{h}_A$ and $\hat{h}_B$, might differ due to independent noise. Reciprocity can be assumed in the case of TDD systems when the channel is quasi-static, i.e., the coherence time of the channel is larger than the measurement time. Thus, the vectors of estimates, $\hat{\mathbf{h}}_A$ and $\hat{\mathbf{h}}_B$ can be obtained during an initial measurement phase by sending pilot signals in consecutive TDD time slots.

A second assumption is that the channel follows a bivariate normal distribution. A Gaussian channel distribution, as shown in [2], [3], minimizes the number of vulnerable bits. If the wireless channel is static or a line-of-sight (LOS) channel, the randomness present is not sufficient such that the central limit theorem holds, leading to a normal distribution of the CSI, which would be ideal for key generation. However, since the wireless channel is also dependent on the radiation patterns of the antennas, random variations in the channel can be induced by using reconfigurable aperture antennas (RECAPs) and changing the capacitive loads at the reconfigurable elements. It has been shown that for a high number of RECAP antenna elements (e.g. 24), and a large number of states as discussed in [2], [7], the channel distribution is very close to a complex Gaussian, with the real and imaginary parts independent and identically distributed. Further details on the the validity of this assumption, RECAP configuration, as well as measurements description can be found in [4], [5]. For the rest of this paper, we will assume a circular symmetric Gaussian channel distribution with zero mean and variance $\sigma^2_{ch}$.

A third assumption refers to the antenna separation. Herein, we assume a sufficient separation between Eve and Alice and Bob such that the legitimate and the eavesdropper channels are not correlated. Recent studies [12] have shown that an antenna separation of half a wavelength is not sufficient, and that for the rate of the number of vulnerable key bits to the total number of key bits to go to zero, an eavesdropper separation of several wavelengths is necessary [15].

## III. Quantization

In order to study the effects of different quantization methods on the overall key disagreement rate, we first consider the case when no reconciliation is performed, and each legitimate user obtains a key by independently quantizing its own analog noisy CSI measurements. We will refer here to the symbol mismatch rate as the probability that given one sample measurement $\hat{h}_A$ at Alice and the corresponding measurement $\hat{h}_B$ at Bob, they are not quantized to the same region by both users.

Points from the channel distribution that are very close to quantization boundaries are very likely to result in a key mismatch. If a channel value $h$ falls within a certain region $\mathcal{R}_i$, but very close to a quantization boundary, its noisy measurements $\hat{h}_A$ and $\hat{h}_B$ might simply jump across the quantization threshold to a neighboring region. Both the quantization algorithm and the size of the codebook $N_q$, greatly impact the overall symbol agreement rate between Alice and Bob.

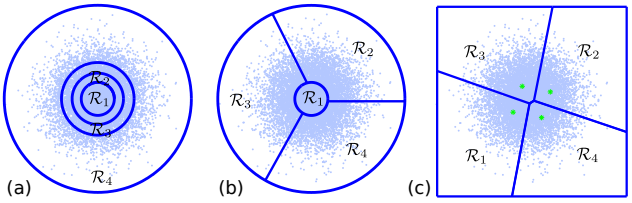



Fig. 2. Three quantization methods - (a) concentric quantization regions; (b) circles and slices; (c) Linde-Buzo-Gray algorithm, $N_q = 4$

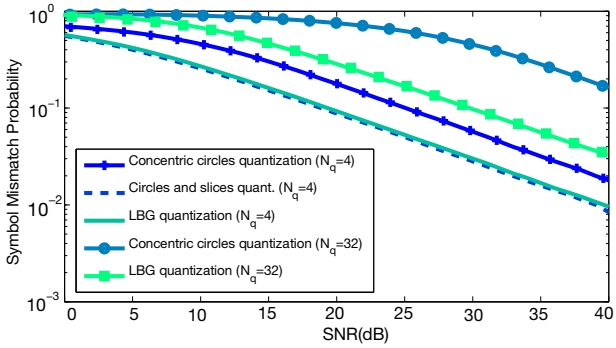



Fig. 3. Symbol mismatch probability between Alice and Bob for different quantization methods (no reconciliation)

We consider three vector quantizations schemes, the first based on concentric circles, as illustrated in Fig. 2-(a), the second based on concentric circles and slices, as shown in Fig. 2-(b), and the third based on the Linde-Buzo-Gray algorithm [5], [14], leading to the quantization areas shown in Fig. 2-(c), all for a codebook size of $N_q = 4$. The radii of the concentric circles in the first method are computed such that the number of measurement points end up uniformly distributed across all regions. The exact values are provided in Table I. While such a method leads to a simplification of the LLR formulation, as it will be explained in Section IV, when the codebook size is increased, a severe performance degradation in terms of symbol-error-ratio (SER), or symbol mismatch rate, is noticed. This is a direct consequence of the shape of the quantization regions. The narrower a region is, the more likely it is that noisy measurements will end up quantized to neighboring regions. Increasing the number of regions $N_q$ in the first quantization example will lead to even narrower regions, thus, such quantization method will be highly sensitive in terms of noise, leading to a high number of errors even at high signal-to-noise (SNR) ratios.

TABLE I
QUANTIZATION LIMITS; $N_q = 4$; $\sigma = \sqrt{\sigma_{ch}^2 + \sigma_A^2}$

| | METHOD (a) | | | METHOD (b) | | |
|---|---|---|---|---|---|---|
| Region | $r_{min_i}$ | $r_{max_i}$ | $r_{min_i}$ | $r_{max_i}$ | $\theta_{min_i}$ | $\theta_{max_i}$ |
| $\mathcal{R}_1$ | 0 | $0.758\ \sigma$ | 0 | $0.758\ \sigma$ | 0 | $2\pi$ |
| $\mathcal{R}_2$ | $0.758\ \sigma$ | $1.177\ \sigma$ | $0.758\ \sigma$ | $\infty$ | 0 | $\frac{2\pi}{3}$ |
| $\mathcal{R}_3$ | $1.177\ \sigma$ | $1.665\ \sigma$ | $0.758\ \sigma$ | $\infty$ | $\frac{2\pi}{3}$ | $\frac{4\pi}{3}$ |
| $\mathcal{R}_4$ | $1.665\ \sigma$ | $\infty$ | $0.758\ \sigma$ | $\infty$ | $\frac{4\pi}{3}$ | $2\pi$ |

The second quantization method introduces the so-called "slices" as a way to mitigate this effect and counteract the error performance degradation. This quantization method also leads to a uniform distribution of the key symbols, which is desirable for secrecy concerns, i.e., not to provide any redundancy to a potential eavesdropper. The boundaries for the four regions $\mathcal{R}_i$ illustrated in Fig. 2-(c) are also provided in Table I, in terms of radii $(r_{min_i}, r_{max_i})$ and angles $(\theta_{min_i}, \theta_{max_i})$, in polar coordinates. As seen in Fig. 3, this method shows a better performance in terms of symbol mismatch probability, as compared to the previous one, with an error reduction from $17\%$ to $9\%$ at an SNR of 20 dB, and from $1.78\%$ to $0.95\%$ at 40 dB, for the case of four quantization regions.

The third algorithm for channel quantization is the Linde-Buzo-Gray (LBG) vector quantization scheme, as described in [14]. The LBG algorithm is a sample version of the Lloyd-Max quantizer that does not require a closed form pdf of the channel distribution, but only the measurement samples. Given a length $M$ sequence of 2-dimensional channel samples and the desired number of code vectors $N_q$, the algorithm delivers the final codebook and the corresponding quantization region for each codeword vector.

The difference in mismatch rate between the first and third quantization method also becomes much more significant when increasing the size of the codebook vector. For an SNR of 20 dB and $N_q = 32$ quantization regions, we notice a probability of $75\%$ that Alice and Bob quantize to different regions when using first method (concentric circles), as compared to $24\%$ when using the LBG algorithm.

Once such partitioning boundaries have been determined, a Gray-like bit mapping can be assigned to the regions.

### A. Key Probability Distribution

The one-time pad perfect secrecy is achieved under two important assumptions, namely that the pad length is at least the size of the message to be encrypted, and that the key is selected at random with a uniform distribution. Thus, achieving a low SER is not sufficient, provided the uniform distribution requirement is not entirely satisfied. Since the key distribution is a parameter that is assumed to be known to the eavesdropper, a non-uniform distribution will result in some keys being more probable than others, facilitating potentially successful analytical attacks. When the output distribution provided by the quantizer is not uniform, the perfect secrecy condition requiring the eavesdroppers equivocation to be equal to the entropy of the message is not satisfied. While the first two quantization methods are constructed with a circularly symmetric zero-mean Gaussian input distribution in mind,

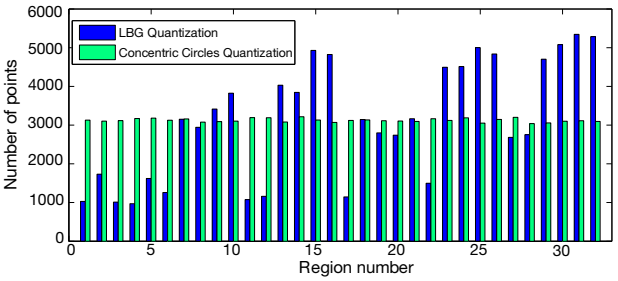



Fig. 4. Distribution of measurement points across $N_q = 32$ Voronoi regions resulting from concentric circles quantization and LBG quantization

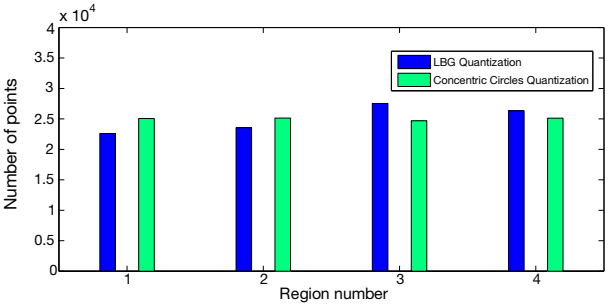



Fig. 5. Distribution of measurement points across $N_q = 4$ Voronoi regions resulting from concentric circles quantization and LBG quantization

with variance $\sigma$, and, by construction, deliver a uniform output distribution of the symbols, regardless of the codebook size[1], the LBG algorithm can be used for any arbitrary distribution. However, we show in Fig. 5 that the LBG quantization fails to deliver an exact uniform output distribution, even for small codebook lengths, i.e., $N_q = 4$. The non-uniformity of the distribution becomes even more pronounced for higher alphabets. Figure 4 shows the distribution of measurement points across 32 Voronoi regions delivered by the LBG algorithm. The non-uniformity can be expressed as a rate loss which we have taken into account in Fig. 3 by a corresponding right shift of the LBG curves. The rate loss corresponds to ideal source coding to make the distribution uniform. Overall, the best error performance among the three cases analyzed is provided by the second quantization method for the case of four regions, with only a slight advantage over the LBG algorithm.

## IV. KEY RECONCILIATION

For notation simplicity, we will denote an analog complex measurement estimate at Alice, $\hat{h}_A$, by $a = x_A + jy_A$, and a channel estimate at Bob, $\hat{h}_B$, by $b = x_B + jy_B$, where $(x, y)$ denote the real and imaginary parts, respectively. These values represent the AWGN-disturbed measurements of the ideal channel sample $c = x_{ch} + jy_{ch}$, at Alice and Bob, respectively. The input to the LDPC decoder consists of two sets of log-likelihood ratios (LLRs), one for the parity symbols received from Alice, and one for Bob's own estimates of the channel, assuming Alice's key bits as "correct" reference. In general,

[1]We only provide here a table of quantization boundaries for four quantization regions, although the limits for higher codebook sizes can be easily computed by integrating the complex Gaussian input distribution and imposing equal "volumes" in each region, i.e, uniform output distribution.

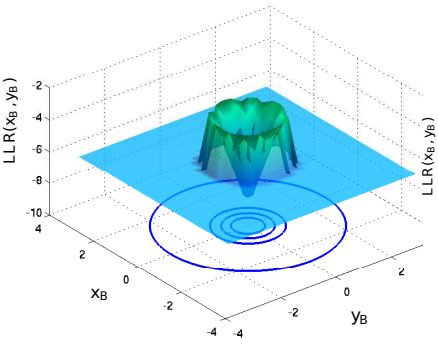



Fig. 6. Method (a): LLR(b)=ln $\frac{P(a\in\mathcal{R}_3|b)}{P(a\notin\mathcal{R}_3|b)}$

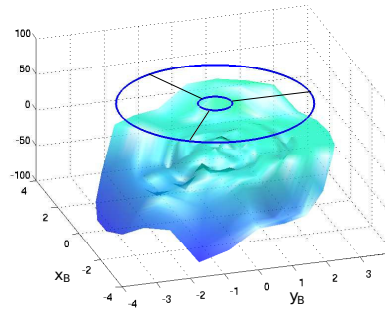



Fig. 7. Method (b): LLR(b)=ln $\frac{P(a\in\mathcal{R}_2|b)}{P(a\notin\mathcal{R}_2|b)}$

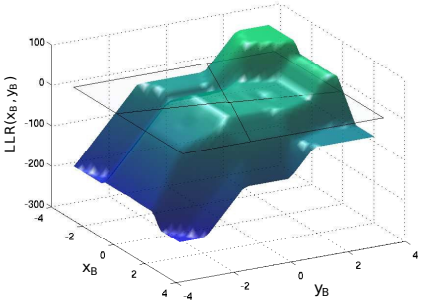



Fig. 8. LBG quant: LLR(b)=ln $\frac{P(a\in\mathcal{R}_3|b)}{P(a\notin\mathcal{R}_3|b)}$

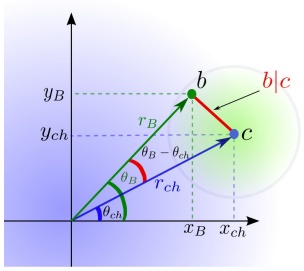



Fig. 9. Polar coordinates transformation; Given channel measurement $c$, $b$ represents Bob's noisy measurement of $c$. In polar coordinates $c$ is represented by $(r_{ch}, \theta_{ch})$.

the LLR for a measurement value at Bob can be computed according to (1), where $p(b|a \in \mathcal{R}_i)$ is the probability density function of Bob's measurement $b$ given that Alice quantized its corresponding value $a$ to region $\mathcal{R}_i$.

$$LLR(b) = \ln \frac{p(b|a \in \mathcal{R}_i)}{p(b|a \notin \mathcal{R}_i)} \tag{1}$$

Previous works, such as [13], consider a noiseless environment for the transmission of side information, or simply consider the same formulation for the information bits, as for the parity (syndrome) bits that are transmitted over the physical channel. This is, however, inaccurate, and leads to sub-optimum decoding.

While the LLR computation for the parity bits is trivial (2), since they might just experience a standard AWGN channel with variance $\sigma_B^2$, the calculation of the LLRs for Bob's information bits is much more problematic, due to the fact that Bob's decoding of the information bits is subject to what Alice quantized to.

$$LLR_{\,parity}(b) = \ln \left( \frac{e^{-\frac{(b-1)^2}{2\sigma_B^2}}}{e^{-\frac{(b+1)^2}{2\sigma_B^2}}} \right) = \frac{2b}{\sigma_B^2} \,. \tag{2}$$

The LLR formulation for the information bits at Bob has to account for the fact that Alice measures the channel with error determined by $\sigma_A^2$, and then quantizes. Nevertheless, Bob assumes that whatever Alice quantized to represents the correct key and it has to reconcile with her values. In [6], we have shown that the general formula for the LLR for the information symbols, assuming a uniform distribution of the quantized measurements across $N_q$ regions, to be as follows

$$LLR(b) = \ln \frac{(N_q - 1)P(a \in \mathcal{R}_i|b)}{P(a \notin \mathcal{R}_i|b)} \,, \tag{3}$$

where $P(a \in \mathcal{R}_i|b)$ represents the probability that Alice quantized its value $a$ to region $\mathcal{R}_i$, given current measurement value $b$ at Bob. In more intuitive terms, Eq. (3) is a measure of the log-likelihood that given a noisy value of the channel at Bob, Alice quantized her noisy counterpart to a certain region $\mathcal{R}_i$ and not the others. Now, by iterating through all the possible regions and computing LLRs for $b$, a vector of LLRs is produced for every variable node of the LDPC decoder that is associated with the channel measurements (information symbols). In [6] we provide a complete derivation of the exact LLR formulation when the channel distribution is a circularly symmetric Gaussian.

Equation (4) shows the LLR expression when the concentric circles quantization is used. For this specific quantization method, a transformation to polar coordinates, such as the one shown in Fig. 9, allows us to express parts of the LLR expression with modified Bessel functions of the first kind ($J_0$) as given by (4). Such a simplification leads to a significant reduction in the number of numerical integrations. However, this is only possible for the concentric circles quantization, which comes with the disadvantage of a worse SER performance than any of the other two methods. The LLR for the slices quantization is given in (5), while the one for arbitrary Voronoi regions is given by (6). The LLRs in (4) – (6) are functions of $b$ that can be viewed as resulting from the convolution of the noise and channel densities, and can be computed in advance by numerical integration, and stored, for a wide range of values of $b$ values and SNRs, in order to speed up the LDPC decoder.

## V. Numerical Results

In this section, we provide some numerical results for the intrinsic LLRs required by the LDPC decoder on Bob's side, necessary for key reconciliation, given the three different types of quantization discussed.

We show in Fig. 6 the numerical results obtained for the log-likelihood ratios for the first type of quantization, for an SNR=14 dB, where the SNR is defined as $\sigma_{ch}^2/\sigma_B^2$. Figure 6 shows for every possible value of $b$, in Cartesian coordinates, $(x_B, y_B)$, the probability that Alice quantized to region $\mathcal{R}_3$ and not to any other regions. As expected, for values of $b$ that would be also in region $\mathcal{R}_3$ and fall sufficiently far away

$$LLR_{\text{(a)}}(r_B) = \ln \frac{(N_q-1)\int_0^{\infty}\int_{r_{min_i}}^{r_{max_i}} r_A r_{ch} e^{-\frac{r_{ch}^2+r_A^2}{2\sigma_A^2}-\frac{r_{ch}^2+r_B^2}{2\sigma_B^2}-\frac{r_{ch}^2}{2\sigma_{ch}^2}} \cdot J_0(-\frac{r_{ch}r_A}{\sigma_A^2}) \cdot J_0(-\frac{r_{ch}r_B}{\sigma_B^2}) dr_A dr_{ch}}{\sum_{\mathcal{R}_k,k\neq i}\int_0^{\infty}\int_{r_{min_k}}^{r_{max_k}} r_A r_{ch} e^{-\frac{r_{ch}^2+r_A^2}{2\sigma_A^2}-\frac{r_{ch}^2+r_B^2}{2\sigma_B^2}-\frac{r_{ch}^2}{2\sigma_{ch}^2}} \cdot J_0(-\frac{r_{ch}r_A}{\sigma_A^2}) \cdot J_0(-\frac{r_{ch}r_B}{\sigma_B^2}) dr_A dr_{ch}} \tag{4}$$

$$LLR_{\text{(b)}}(r_B,\theta_B) = \ln \frac{(N_q-1)\int_0^{\infty}\int_0^{2\pi}\int_{r_{min_i}}^{r_{max_i}}\int_{\theta_{min_i}}^{\theta_{max_i}} r_A r_{ch} e^{-\frac{r_{ch}^2+r_A^2-2r_{ch}r_A\cos(\theta_{ch}-\theta_A)}{2\sigma_A^2}-\frac{r_{ch}^2+r_B^2-2r_{ch}r_B\cos(\theta_{ch}-\theta_B)}{2\sigma_B^2}-\frac{r_{ch}^2}{2\sigma_{ch}^2}} d\theta_A dr_A d\theta_{ch} dr_{ch}}{\sum_{\mathcal{R}_k,k\neq i}\int_0^{\infty}\int_0^{2\pi}\int_{r_{min_k}}^{r_{max_k}}\int_{\theta_{min_k}}^{\theta_{max_k}} r_A r_{ch} e^{-\frac{r_{ch}^2+r_A^2-2r_{ch}r_A\cos(\theta_{ch}-\theta_A)}{2\sigma_A^2}-\frac{r_{ch}^2+r_B^2-2r_{ch}r_B\cos(\theta_{ch}-\theta_B)}{2\sigma_B^2}-\frac{r_{ch}^2}{2\sigma_{ch}^2}} d\theta_A dr_A d\theta_{ch} dr_{ch}} \tag{5}$$

$$LLR_{\text{(c)}}(x_B,y_B) = \ln \frac{(N_q-1)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\mathcal{R}_{i_y}}\int_{\mathcal{R}_{i_x}} e^{-\frac{(x_A-x_{ch})^2+(y_A-y_{ch})^2}{2\sigma_A^2}-\frac{(x_{ch}-x_B)^2+(y_{ch}-y_B)^2}{2\sigma_B^2}-\frac{(x_{ch}^2+y_{ch}^2)}{2\sigma_{ch}^2}} dx_A\, dy_A\, dx_{ch}\, dy_{ch}}{\sum_{\mathcal{R}_k,k\neq i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\mathcal{R}_{k_y}}\int_{\mathcal{R}_{k_x}} e^{-\frac{(x_A-x_{ch})^2+(y_A-y_{ch})^2}{2\sigma_A^2}-\frac{(x_{ch}-x_B)^2+(y_{ch}-y_B)^2}{2\sigma_B^2}-\frac{(x_{ch}^2+y_{ch}^2)}{2\sigma_{ch}^2}} dx_A\, dy_A\, dx_{ch}\, dy_{ch}} \tag{6}$$

from any quantization boundaries, the LLR is maximum. As $b$ takes values closer to the quantization thresholds and into other regions, the LLR decreases to a minimum. This is the case for regions $\mathcal{R}_1, \mathcal{R}_2$, and $\mathcal{R}_4$. For the second quantization method (b), Fig. 7 shows the log-likelihood plot for one of the external slices. Figure 8 shows the log-likelihood for $\mathcal{R}_3$, for the case of the arbitrary Voronoi quantization regions, as provided by the LBG algorithm. The numbering of the regions is the one provided in Fig. 2. For our simulations, we used a discrete grid for $b$ with incremental values of $0.05$ between $[-3.5, 3.5]$ for both axes.

## VI. Conclusion

We have investigated the problem of physical-layer key generation and reconciliation with Slepian-Wolf coding and Low-Density Parity-Check (LDPC) codes in a wireless scenario when two users measure a reciprocal channel with independent noise on both sides. We offered an analysis on the effect of different quantization schemes on the overall error rate performance, or key disagreement rate, assuming imperfect channel measurements. Our results show that for higher codebook sizes, the Linde-Buzo-Gray quantizer does not output a uniform distribution of key symbols, which is of paramount importance for the secrecy aspect, and show a possible quantization scheme that guarantees a uniform output distribution and also provides a slightly lower key disagreement rate than the LBG quantizer. We have further shown the log-likelihood (LLR) formulation required by a soft-decision LDPC decoder for key reconciliation, for each of the quantization schemes analyzed and a circularly symmetric complex Gaussian channel distribution.

## Acknowledgment

This work is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG).

## References

[1] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, 28:656–715, 1949.

[2] R. Mehmood and J. Wallace, "Wireless Security Enhancement Using Reconfigurable Aperture Antennas," *European Conference on Antennas and Propagation (EuCAP'11)*, Rome, Italy, Apr. 12-16, 2011, pp. 1-5.

[3] R. Mehmood and J. Wallace, "MIMO Capacity Enhancement Using Parasitic Reconfigurable Aperture Antennas (RECAPs)," *IEEE Transactions on Antennas and Propagation*, vol. 60, pp. 665-673, Feb. 2012.

[4] J. Wallace, R. Kurma, "Automatic Secret Keys From Reciprocal MIMO Wireless Channels: Measurement and Analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5, pp. 381-392, Sept. 2010.

[5] A. Filip, R. Mehmood, J. Wallace, and W. Henkel, "Variable Guard Band Construction to Support Key Reconciliation," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May. 4-9, 2014.

[6] O. Graur, N. Islam, A. Filip, and W. Henkel, "Quantization Aspects in LDPC Key Reconciliation for Physical Layer Security" *10th International ITG Conference on Systems, Communications and Coding (SCC)*, Hamburg, Germany, February 2-5, 2015.

[7] A. Filip, R. Mehmood, J. Wallace, and W. Henkel, "Physical-Layer Key Generation Supported by RECAP Antenna Structures," proc. *9th International ITG Conference on Source and Channel Coding (SCC)*, Munich, Germany, 2013.

[8] M. Bloch, J. Barros, "Physical-Layer Security: From Information Theory to Security Engineering," *Cambridge University Press*, 2011.

[9] X. Zhou, L. Song, Y. Zhang, "Physical Layer Security in Wireless Communications," *CRC Press Inc.*, 2013.

[10] R. Wilson, D. Tse, R. .A. Scholz, "Channel Identification: Secret Sharing using Reciprocity in Ultrawideband Channels," *IEEE Transactions on Information Forensics and Security*, 2:364-375, Sept. 2007.

[11] C.Y. William Lee, "Mobile Communications Design Fundamentals," *John Wiley & Sons, Inc.,*, New York, NY, USA, 1992, pp. 227-240.

[12] A. J. Pierrot, R. A. Chou, M. R. Bloch, "Experimental Aspects of Secret Key Generation in Indoor Wireless Environments," in *2013 IEEE 14th Workshop on. Signal Processing Advances in Wireless Communications (SPAWC)*, 2013, pp. 669673.

[13] X. Sun, X. Wu, C. Zhao, M. Jiang, and W. Xu, "Slepian-Wolf Coding for Reconciliation of Physical Layer Secret Keys," proc. *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, 18-21 Apr. 2010.

[14] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design,"*IEEE Transactions on Communications*, 28:84-95, 1980.

[15] J.W. Wallace and R.K. Sharma, "Automatic Secret Keys from Reciprocal MIMO Wireless Channels: Measurement and Analysis," *IEEE Trans. Inf. Forensics and Security*, vol. 5, no. 3, pp. 381-392, Sep. 2010.

# An Importance Sampling Algorithm for the Ising Model with Strong Couplings

Mehdi Molkaraie
Dept. of Information and Communication Technologies
ETH Zurich
8092 Zürich, Switzerland
mehdi.molkaraie@alumni.ethz.ch

***Abstract*—We consider the problem of estimating the partition function of the two-dimensional ferromagnetic Ising model in an external magnetic field. The estimation is done via importance sampling in the dual of the Forney factor graph representing the model. Emphasis is on models at low temperature (corresponding to models with strong couplings) and on models with a mixture of strong and weak coupling parameters.**

## I. Introduction

The problem of estimating the partition function of the finite-size two-dimensional (2D) ferromagnetic Ising model in a consistent external field is considered. Applying factor graph duality to address the problem has been investigated in [1]–[4]. It was demonstrated in [1] that Monte Carlo methods based on the dual factor graph work very well for the Ising model at low temperature. In contrast, Monte Carlo methods in the primal/original graph suffer from critical slowing down and erratic convergence to estimate the partition function in the low-temperature regime [5]. Monte Carlo methods (based on uniform sampling and Gibbs sampling) in the dual factor graph were also proposed in [1] to estimate the partition function of the 2D Ising model without an external field.

In this paper, we continue this research to extend the results of [1], [2] to models with a mixture of strong and weak coupling parameters and in the presence of an external magnetic field. After defining an auxiliary probability mass function in the dual Forney factor graph of the model, we propose an importance sampling algorithm that can efficiently estimate the partition function. A similar importance sampling algorithm, designed specifically for models in a strong external field, was recently proposed in [2].

The paper is organized as follows. We review the Forney factor graph representation of the 2D Ising model in an external field in Section II. Section III discusses dual Forney factor graphs and the normal factor graph duality theorem. The importance sampling algorithm is described in Section IV. In Section V, we report numerical experiments.

## II. The Ising Model in an External Magnetic Field

Let $X_1, X_2, \ldots, X_N$ be a set of discrete binary random variables arranged on the sites of a 2D lattice. We suppose that interactions are restricted to adjacent (nearest-neighbor) variables (see Fig. 1). The real coupling parameter $J_{k,\ell}$ controls the strength of the interaction between adjacent variables $(X_k, X_\ell)$. The real parameter $H_m$ corresponds to the presence of an external field and controls the strength of the interaction between $X_m$ and the field. Each random variable takes on values in $\mathcal{X} = \{0, 1\}$. Let $x_i$ represent a possible realization of $X_i$, $\mathbf{x}$ stand for a configuration $(x_1, x_2, \ldots, x_N)$, and $\mathbf{X}$ stand for $(X_1, X_2, \ldots, X_N)$.

The energy of a configuration $\mathbf{x}$ is given by [6]

$$\mathcal{H}(\mathbf{x}) = -\sum_{(k,\,\ell)\,\in\,\mathcal{B}} J_{k,\ell} \cdot \big([x_k = x_\ell] - [x_k \neq x_\ell]\big) - \sum_{m=1}^{N} H_m \cdot \big([x_m = 1] - [x_m = 0]\big) \quad (1)$$

where $\mathcal{B}$ contains all the unordered pairs (bonds) $(k, \ell)$ with non-zero interactions, and $[\cdot]$ denotes the Iverson bracket [7], which evaluates to 1 if the condition in the bracket is satisfied and to 0 otherwise.

In this paper, the focus is on ferromagnetic Ising models characterized by $J_{k,\ell} > 0$ for each $(k, \ell) \in \mathcal{B}$. The external field is assumed to be consistent, i.e., it is either assigned to all positive or to all negative values.

The probability that the model is in configuration $\mathbf{x}$ is given by the Boltzmann distribution [6]

$$p_{\mathrm{B}}(\mathbf{x}) = \frac{e^{-\beta\mathcal{H}(\mathbf{x})}}{Z} \quad (2)$$

where the normalization constant $Z$ is the *partition function* $Z = \sum_{\mathbf{x}\in\mathcal{X}^N} e^{-\beta\mathcal{H}(\mathbf{x})}$ and $\beta$ is the inverse temperature. In the rest of this paper, we assume $\beta = 1$. With this assumption, large values of $J$ correspond to models at low temperature. Boundary conditions are assumed to be periodic.

For each adjacent pair $(x_k, x_\ell)$, let $\kappa : \mathcal{X}^2 \to \mathbb{R}_{>0}$

$$\kappa_{k,\ell}(x_k, x_\ell) = e^{J_{k,\ell}\cdot\left([x_k=x_\ell]-[x_k\neq x_\ell]\right)} \quad (3)$$

and for each $x_m$, let $\tau : \mathcal{X} \to \mathbb{R}_{>0}$

$$\tau_m(x_m) = e^{H_m\cdot\left([x_m=1]-[x_m=0]\right)} \quad (4)$$

We then define $f : \mathcal{X}^N \to \mathbb{R}_{>0}$ as

$$f(\mathbf{x}) \triangleq \prod_{(k,\,\ell)\,\in\,\mathcal{B}} \kappa_{k,\ell}(x_k, x_\ell) \prod_{m=1}^{N} \tau_m(x_m) \quad (5)$$

The corresponding Forney factor graph (normal graph) for the factorization in (5) is shown in Fig. 1, where the boxes labeled "=" are equality constraints [8], [9]. In Forney factor graphs variables are represented by edges.

From (5), $Z$ in (2) can also be expressed as

$$Z = \sum_{\mathbf{x} \in \mathcal{X}^N} f(\mathbf{x}) \tag{6}$$

At high temperature (i.e., for small $J$), the Boltzmann distribution (2) approaches the uniform distribution. In this case, Monte Carlo methods for estimating $Z$ usually perform well in the primal factor graph. Estimating $Z$ in the low-temperature regime is more challenging [5], [10], [11].

In this paper, we consider models at low temperature (i.e., with large $J$) and models with a mixture of strong and weak coupling parameters in an external magnetic field. To compute an estimate of $Z$ in this case, we propose an importance sampling algorithm in the dual of the Forney factor graph of the 2D Ising model.

## III. The Dual Forney Factor Graph

We can obtain the dual of the Forney factor graph in Fig. 1, by replacing each binary variable $x$ with its dual binary variable $\tilde{x}$, each factor $\kappa_{k,\ell}$ with its 2D Discrete Fourier transform (DFT), each factor $\tau_m$ with its one-dimensional (1D) DFT, and each equality constraint with an XOR factor, cf. [8], [12]–[14]. Fig. 2 shows the dual Forney factor graph of the 2D Ising model, where boxes containing "$+$" symbols represent XOR factors as

$$g(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k) = [\tilde{x}_1 \oplus \tilde{x}_2 \oplus \ldots \oplus \tilde{x}_k = 0] \tag{7}$$

the small boxes attached to each XOR factor are given by

$$\lambda_m(\tilde{x}_m) = \begin{cases} \cosh H_m, & \text{if } \tilde{x}_m = 0 \\ -\sinh H_m, & \text{if } \tilde{x}_m = 1 \end{cases} \tag{8}$$

and the unlabeled normal-size boxes attached to each equality constraint represent factors as

$$\gamma_k(\tilde{x}_k) = \begin{cases} 2\cosh J_k, & \text{if } \tilde{x}_k = 0 \\ 2\sinh J_k, & \text{if } \tilde{x}_k = 1 \end{cases} \tag{9}$$

Here, $J_k$ is the coupling parameter associated with each bond. See [1]–[3], for more details on constructing the dual Forney factor graph of the 2D Ising model.

In the dual domain, we denote the partition function by $Z_{\mathrm{d}}$. For the models that we study here, the normal factor graph duality theorem states that (see [13, Theorem 2])

$$Z_{\mathrm{d}} = |\mathcal{X}^N| Z \tag{10}$$

In order to design Monte Carlo methods in the dual Forney graph, we require factors (8) and (9) to be non-negative. In a 2D Ising model, $Z$ is invariant under the change of sign of the external field [6]. Therefore, without loss of generality, we will assume $H_m < 0$ for $1 \leq m \leq N$. Under the ferromagnetic assumption $J_{k,\ell} > 0$ for $(k, \ell) \in \mathcal{B}$. With these assumptions, (8) and (9) will be non-negative.

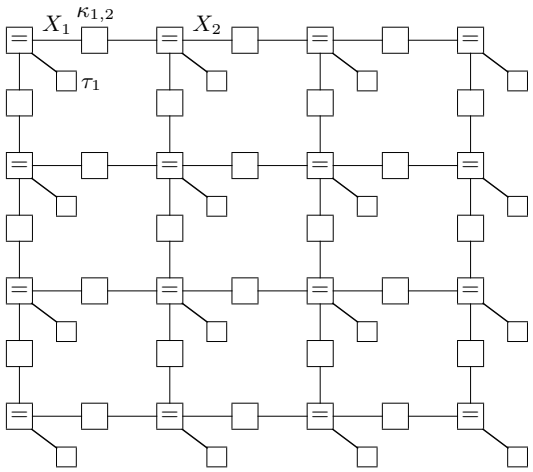



Fig. 1: Forney factor graph of the 2D Ising model in an external field, where unlabeled normal-size boxes represent (3), small boxes represent (4), and boxes containing " $=$ " symbols are equality constraints.

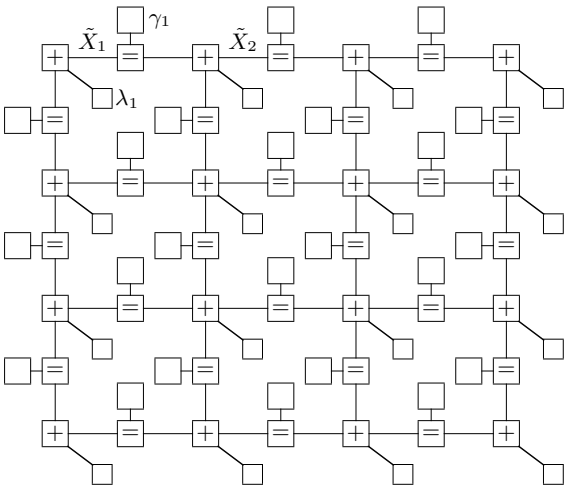



Fig. 2: The dual Forney factor graph of the 2D Ising model in an external field, where boxes containing " $+$ " symbols represent (7), small boxes represent (8), and unlabeled normal-size boxes represent (9).

## IV. The Importance Sampling Algorithm

The importance sampling algorithm is described on Fig. 2. We partition $\tilde{\mathbf{X}}$ into $\tilde{\mathbf{X}}_A$ and $\tilde{\mathbf{X}}_B$, with the condition that $\tilde{\mathbf{X}}_B$ is a linear combination (involving the XOR factors) of $\tilde{\mathbf{X}}_A$. In this set-up, a valid configuration in the dual factor graph can be created by assigning values to $\tilde{\mathbf{X}}_A$, followed by computing $\tilde{\mathbf{X}}_B$ as a linear combination of $\tilde{\mathbf{X}}_A$.

An example of such a partitioning is shown in Fig. 3, where $\tilde{\mathbf{X}}_A$ is the set of all the variables associated with the thick edges and $\tilde{\mathbf{X}}_B$ the set of all the variables associated with the remaining thin edges. Accordingly, let $\mathcal{B}_A \subset \mathcal{B}$ contain the indices of the bonds marked by thick edges and $\mathcal{B}_B = \mathcal{B} - \mathcal{B}_A$. For a valid configuration $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_A, \tilde{\mathbf{x}}_B)$, let

$\tilde{\mathbf{x}}_A = (\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$, where $\tilde{\mathbf{y}}$ contains all the thick edges attached to the small unlabeled boxes (involved in (8)) and $\tilde{\mathbf{z}}$ contains all the variables associated with the thick bonds (involved in (9)).

We prove that $\mathrm{w_H}(\tilde{\mathbf{y}})$, the Hamming weight of $\tilde{\mathbf{y}}$, is always even, where the Hamming weight of a vector is the number of non-zero components of that vector [15].

**Lemma 1.** If $\tilde{\mathbf{x}}$ is a valid configuration in the dual Forney factor graph, then $\mathrm{w_H}(\tilde{\mathbf{y}})$ is even.

*Proof.* We consider $c = \bigoplus_{t=1}^{N} \tilde{y}_t$ the component-wise XOR of $\tilde{\mathbf{y}}$. Each XOR factor imposes the constraint that all its incident variables sum to 0 in GF(2). Each $\tilde{y}_t$ in $c$ can thus be expanded as the XOR of the corresponding variables associated with the bonds, furthermore, the variables on the bonds each appear twice in this expansion. Hence $c = 0$, i.e., $\mathrm{w_H}(\tilde{\mathbf{y}})$ is even. ■

Lemma 1 implies that $Z_{\mathrm{d}}$, and thus $Z$ itself, are invariant under the change of sign of $H_m$. Indeed, regardless of the sign of $H_m$, i.e., assigned to all positive or to all negative values $\prod_{m=1}^{N} \lambda_m(\tilde{x}_m)$ takes the same positive value, cf. (8).

The importance sampling algorithm works as follows. To draw $\tilde{\mathbf{x}}^{(\ell)}$ at each iteration $\ell$, we first draw $\tilde{\mathbf{x}}_A^{(\ell)}$ according to a suitably defined auxiliary probability mass function on the bonds (see (13)). We then update $\tilde{\mathbf{x}}_B^{(\ell)}$ to create a valid configuration $\tilde{\mathbf{x}}^{(\ell)} = (\tilde{\mathbf{x}}_A^{(\ell)}, \tilde{\mathbf{x}}_B^{(\ell)})$. Updating $\tilde{\mathbf{x}}_B^{(\ell)}$ at each iteration is easy as $\tilde{\mathbf{x}}_B$ is a linear combination of $\tilde{\mathbf{x}}_A$.

Let us define

$$\Lambda(\tilde{\mathbf{x}}_B) \triangleq \prod_{k \in \mathcal{B}_B} \gamma_k(\tilde{x}_k) \tag{11}$$

$$\Psi(\tilde{\mathbf{x}}_A) \triangleq \prod_{k \in \mathcal{B}_A} \gamma_k(\tilde{x}_k) \prod_{m=1}^{N} \lambda_m(\tilde{x}_m) \tag{12}$$

$$q(\tilde{\mathbf{x}}_A) \triangleq \frac{\Psi(\tilde{\mathbf{x}}_A)}{Z_q}, \qquad \forall \tilde{\mathbf{x}}_A \in \mathcal{X}^{|\mathcal{B}_A|} \tag{13}$$

where $Z_q$ in (13) is available as

$$Z_q = \sum_{\tilde{\mathbf{x}}_A} \Psi(\tilde{\mathbf{x}}_A) = 2^{|\mathcal{B}_A|} \exp\Big( \sum_{k \in \mathcal{B}_A} J_k - \sum_{m=1}^{N} H_m \Big) \tag{14}$$

Here $|\mathcal{B}_A|$ is the cardinality of $\mathcal{B}_A$. Note that $H_m < 0$.

The product form of (12) suggests that to draw a sample $\tilde{\mathbf{x}}_A^{(\ell)} = (\tilde{\mathbf{y}}^{(\ell)}, \tilde{\mathbf{z}}^{(\ell)})$ according to $q(\tilde{\mathbf{x}}_A)$, two separate subroutines are required, one for the $\tilde{\mathbf{y}}^{(\ell)}$-part, and another for the $\tilde{\mathbf{z}}^{(\ell)}$-part. To draw the $\tilde{\mathbf{y}}^{(\ell)}$-part, we apply.

**repeat**
  *draw* $u_1^{(\ell)}, u_2^{(\ell)}, \ldots, u_N^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$
  **for** $m = 1$ **to** $N$
    **if** $u_m^{(\ell)} < \frac{1}{2}(1 + e^{2H_m})$
      $\tilde{y}_m^{(\ell)} = 0$
    **else**
      $\tilde{y}_m^{(\ell)} = 1$
    **end if**
  **end for**
**until** $\mathrm{w_H}(\tilde{\mathbf{y}}^{(\ell)})$ *is even*

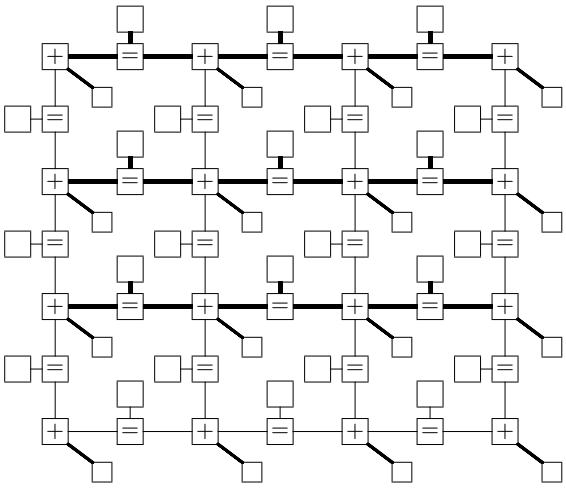


Fig. 3: A partitioning of variables in the dual Forney factor graph of the 2D Ising model. The thick edges represent $\tilde{\mathbf{X}}_A$ and the remaining thin edges represent $\tilde{\mathbf{X}}_B$.

The criteria to accept $\tilde{\mathbf{y}}^{(\ell)}$ is based on Lemma 1. The quantity $\frac{1}{2}(1 + e^{2H_m})$ is equal to $\lambda_m(0)/\big(\lambda_m(0) + \lambda_m(1)\big)$.

To draw the $\tilde{\mathbf{z}}^{(\ell)}$-part, the following subroutine is applied.

*draw* $u_1^{(\ell)}, u_2^{(\ell)}, \ldots, u_{|\mathcal{B}_A|}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$
**for** $k = 1$ **to** $|\mathcal{B}_A|$
  **if** $u_k^{(\ell)} < \frac{1}{2}(1 + e^{-2J_k})$
    $\tilde{z}_k^{(\ell)} = 0$
  **else**
    $\tilde{z}_k^{(\ell)} = 1$
  **end if**
**end for**

Here, $\frac{1}{2}(1 + e^{-2J_k})$ is equal to $\gamma_k(0)/\big(\gamma_k(0) + \gamma_k(1)\big)$. We can then create $\tilde{\mathbf{x}}_A^{(\ell)}$ as a concatenation of $\tilde{\mathbf{y}}^{(\ell)}$ and $\tilde{\mathbf{z}}^{(\ell)}$.

It is possible to compute the probability of rejection in the algorithm. E.g., if the model is in a constant external field $H$

$$\mathrm{P}\big(\mathrm{w_H}(\tilde{\mathbf{y}}) \text{ is odd}\big) = \sinh(N|H|)e^{-N|H|} \tag{15}$$

$$\leq 0.5 \tag{16}$$

The two previous subroutines will provide *i.i.d.* samples $\tilde{\mathbf{x}}_A^{(1)}, \tilde{\mathbf{x}}_A^{(2)}, \ldots, \tilde{\mathbf{x}}_A^{(\ell)}, \ldots$ according to (13). Updating $\tilde{\mathbf{x}}_B^{(\ell)}$ is easy after generating $\tilde{\mathbf{x}}_A^{(\ell)}$. The created samples are then used in the following importance sampling algorithm in order to estimate $Z_{\mathrm{d}}$.

**for** $\ell = 1$ **to** $L$
  *draw* $\mathbf{x}_A^{(\ell)}$ *according to* $q(\tilde{\mathbf{x}}_A)$
  *update* $\tilde{\mathbf{x}}_B^{(\ell)}$
**end for**
*compute*

$$\hat{Z}_{\mathrm{IS}} = \frac{Z_q}{L} \sum_{\ell=1}^{L} \Lambda(\tilde{\mathbf{x}}_B^{(\ell)}) \tag{17}$$

**Lemma 2.** $\hat{Z}_{\mathrm{IS}}$ is an unbiased estimator of $Z_{\mathrm{d}}$.

*Proof.*

$$\begin{aligned} \mathrm{E}_q[\,\hat{Z}_{\mathrm{IS}}\,] &= \mathrm{E}_q\Big[\,\frac{Z_q}{L}\sum_{\ell=1}^{L}\Lambda(\tilde{\mathbf{X}}_B^{(\ell)})\,\Big] \\ &= Z_q\cdot\mathrm{E}_q\,[\,\Lambda(\tilde{\mathbf{X}}_B)\,] \\ &= \sum_{\tilde{\mathbf{x}}_A}\Psi(\tilde{\mathbf{x}}_A)\cdot\Lambda(\tilde{\mathbf{x}}_B) \\ &= Z_{\mathrm{d}} \end{aligned}$$

■

The estimate of $Z_{\mathrm{d}}$ is then used to compute a Monte Carlo estimate of $Z$, as in (6), via the normal factor graph duality theorem (cf. Section III).

The accuracy of (17) depends on the fluctuations of $\Lambda(\tilde{\mathbf{x}}_B)$. If $\Lambda(\tilde{\mathbf{x}}_B)$ varies smoothly, $\hat{Z}_{\mathrm{IS}}$ will have a small variance. From (9) and (11), we expect to observe a small variance if $J_k$ is large for $k \in \mathcal{B}_B$ – as for large values of $J_k$, each factor (9) tends to a constant factor. For more details, see [4].

We emphasize that our choice of partitioning in Fig. 3 is not unique. Fig. 4 shows another example of a partitioning in the dual Forney factor graph whose corresponding partitioning in the primal factor graph is not cycle-free. A partitioning which gives rise to a slightly different importance sampling algorithm (with no rejections) is discussed in [4].

The proposed algorithm is applicable to the Ising model in the absence of an external field as well. Indeed, partitionings in Figs. 3 and 4 are valid even when the external field is not present. We will consider Ising models without an external field in our numerical experiments in Section V-A.

That being the case, to observe fast convergence in the dual domain, not all the coupling parameters need to be strong, but a restricted subset of them. The method of this paper can thus be regarded as supplementary to the ones presented in [1] and [2], where the focus is on models at low temperature (corresponding to models in which all the coupling parameters are strong) and on models in a strong external field.

## V. Numerical Experiments

We apply the importance sampling algorithm to estimate the log partition function per site, i.e., $\frac{1}{N}\ln Z$, of 2D Ising models. All simulation results show $\frac{1}{N}\ln Z$ vs. the number of samples for *one instance*[1] of the model with periodic boundaries.

We consider 2D ferromagnetic Ising models with spatially varying (edge-dependent) coupling parameters without an external field in Section V-A We will also compare the efficiency of the importance sampling algorithm with uniform sampling. Comparisons with Gibbs sampling and the Swendsen-Wang algorithm [16] are discussed in [4]. 2D ferromagnetic Ising models in an external field with spatially varying model parameters are considered in Section V-B.

[1]In statistical physics, estimating quantities for a fixed set of couplings (generated according to some distribution) is called the "quenched average".

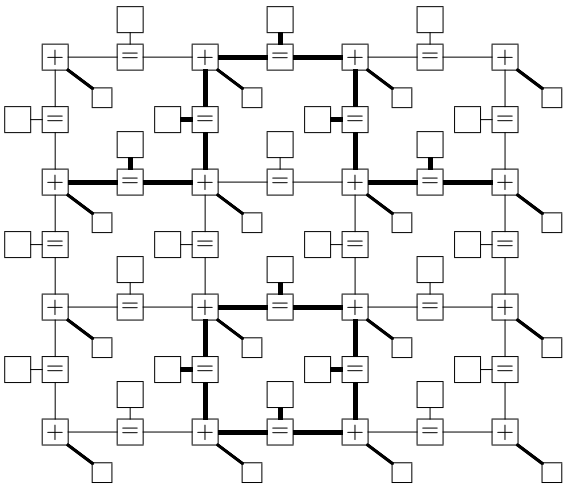


Fig. 4: Another example of a partitioning of variables in the dual Forney factor graph of the 2D Ising model.

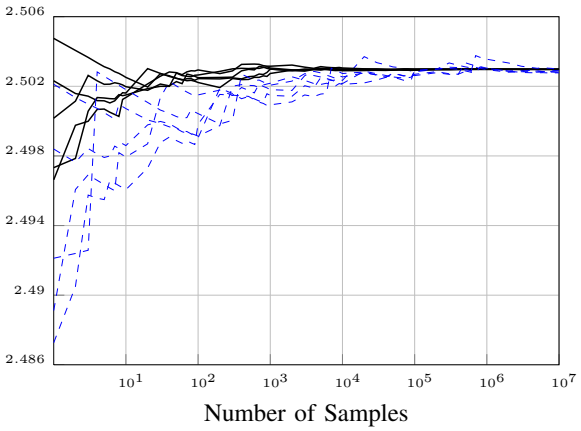



Fig. 5: Estimated log partition function per site vs. the number of samples for a $30\times 30$ Ising model, with $J_k \sim \mathcal{U}[1.0, 1.25]$ for $k \in \mathcal{B}_A$ and $J_k \sim \mathcal{U}[1.25, 1.5]$ for $k \in \mathcal{B}_B$. The plot shows five different sample paths obtained from importance sampling (solid black lines) and five different sample paths obtained from uniform sampling (dashed blue lines) on the dual factor graph.

### A. 2D Ising models without an external field

We consider a 2D Ising model of size $N = 30\times 30$ without an external magnetic field. For $k \in \mathcal{B}_A$, we set $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[1.0, 1.25]$ and for $k \in \mathcal{B}_B$, set $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[1.25, 1.5]$.

Fig. 5 shows simulation results obtained from importance sampling (solid lines) and from uniform sampling (dashed lines) in the dual Forney factor graph. From Fig. 5, the estimated log partition function per site is about 2.503.

We observe that importance sampling outperforms uniform sampling (with virtually the same amount of computation time); see also [2], [4].

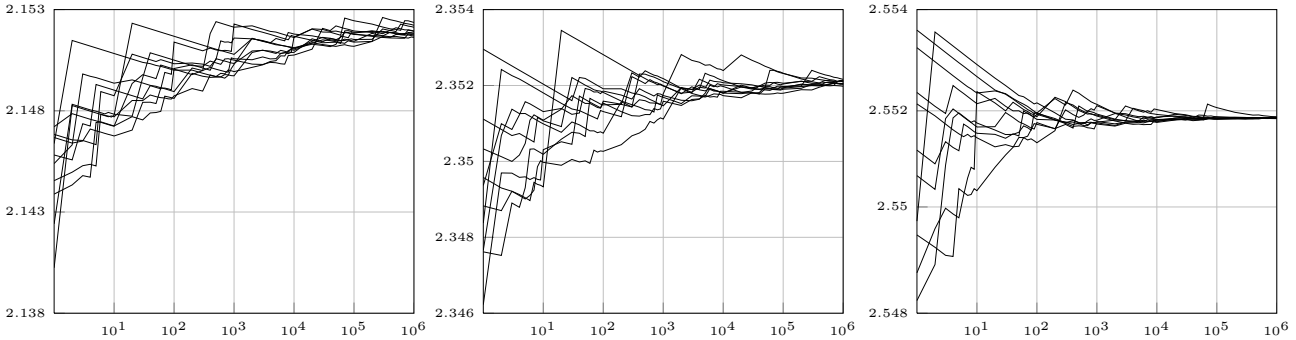



Fig. 6: Estimated log partition function per site vs. the number of samples for a $50 \times 50$ Ising model, with $J_k \sim \mathcal{U}[0.1, 1.0]$ for $k \in \mathcal{B}_A$ and $H_m \sim \mathcal{U}[-0.8, -0.2]$ for $1 \leq m \leq N$; for $k \in \mathcal{B}_B$ (left) $J_k \sim \mathcal{U}[1.0, 1.2]$, (middle) $J_k \sim \mathcal{U}[1.2, 1.4]$, and (right) $J_k \sim \mathcal{U}[1.4, 1.6]$. Each plot shows ten different sample paths obtained from importance sampling on the dual factor graph.

### *B. 2D Ising models in an external field*

We set $N = 50 \times 50$, $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0.1, 1.0]$ for $k \in \mathcal{B}_A$, and $H_m \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-0.8, -0.2]$ for $1 \leq m \leq N$ in all the experiments.

In the first experiment, $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[1.0, 1.2]$ for $k \in \mathcal{B}_B$. Simulation results obtained from importance sampling in the dual factor graph are shown in Fig. 6 (left). In the second experiment, $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[1.4, 1.5]$ for $k \in \mathcal{B}_B$. Fig. 6 (middle) shows simulation results. We set $J_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}[1.4, 1.6]$ for $k \in \mathcal{B}_B$ in the third experiment. Simulation results are shown in Fig. 6 (right), where the estimated $\frac{1}{N} \ln Z$ is about 2.5518. Notice that in Fig. 6 from left to right, the range of the $y$-axis is 0.015, 0.008, and 0.006, respectively.

In agreement with our analysis in Section IV, we observe that convergence improves as $J_k$ becomes larger for $k \in \mathcal{B}_B$.

## VI. Conclusion

An importance sampling algorithm was presented for estimating the partition function of the 2D ferromagnetic Ising model in a consistent external magnetic field. The algorithm is described in the dual Forney factor graph representing the model. After introducing a partitioning and an auxiliary importance sampling distribution, the method operates by first simulating a subset of the variables, followed by doing computations over the remaining ones. The algorithm can efficiently estimate the partition function when the model is at low temperature or when the model contains a mixture of strong and weak coupling parameters. The proposed algorithm is applicable to the 3D Ising model and the $q$-state Potts model in an external field as well. For duality results in the context of statistical physics, see, e.g., [17], [18], [19, Chapter 10].

## Acknowledgements

The author would like to thank Hans-Andrea Loeliger, David Forney, and Justin Dauwels for their helpful comments. The author would also like to thank Pascal Vontobel for proofreading an earlier version of this paper and for pointing out to him [18].

## References

[1] M. Molkaraie and H.-A. Loeliger, "Partition function of the Ising model via factor graph duality," *Proc. 2013 IEEE Int. Symp. on Inf. Theory,* Istanbul, Turkey, July 7–12, 2013, pp. 2304–2308.

[2] M. Molkaraie, "An importance sampling scheme for models in a strong external field," *Proc. 2015 IEEE Int. Symp. on Inf. Theory,* Hong Kong, June 14–19, 2015, pp. 1179–1183.

[3] A. Al-Bashabsheh and Y. Mao, "On stochastic estimation of the partition function," *Proc. 2014 IEEE Int. Symp. on Inf. Theory,* Honolulul, USA, June 29 – July 4, 2014, pp. 1504–1508.

[4] M. Molkaraie, "An importance sampling scheme on dual factor graphs. II. models with strong couplings," `arXiv:1404.5666`, 2014.

[5] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics.* Springer, 2010.

[6] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics.* Dover Publications, 2007.

[7] D. E. Knuth, "Two notes on notation," *Amer. Mathematical Monthly,* vol. 99, pp. 403–422, May 1992.

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

[9] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Proc. Mag.,* vol. 29, pp. 28–41, Jan. 2004.

[10] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods.* Methuen & Co., London, 1964.

[11] R. M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Techn. Report CRG-TR-93-1, Dept. Computer Science, Univ. of Toronto, Sept. 1993.

[12] G. D. Forney, Jr., "Codes on graphs: duality and MacWilliams identities," *IEEE Trans. Inf. Theory,* vol. 57, pp. 1382–1397, Feb. 2011.

[13] A. Al-Bashabsheh and Y. Mao, "Normal factor graphs and holographic transformations," *IEEE Trans. Inf. Theory,* vol. 57, pp. 752–763, Feb. 2011.

[14] G. D. Forney, Jr. and P. O. Vontobel, "Partition functions of normal factor graphs," *2011 Information Theory and Applications Workshop,* La Jolla, USA, Feb. 6–11, 2011.

[15] R. J. McEliece, *The Theory of Information and Coding: A Mathematical Framework for Communication.* Addison-Wesley, 1977.

[16] R. H. Swendsen and J. S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev.,* vol. 58, pp. 86–88, Jan. 1987.

[17] H. A. Kramers and G. H. Wannier, "Statistics of the two-dimensional ferromagnet. Part I," *Phys. Rev.,* vol. 60, pp. 252–262, Aug. 1941.

[18] R. Savit, "Duality in field theory and statistical systems," *Rev. of Modern Physics,* vol. 52, pp. 453–487, April 1980.

[19] H. Nishimoro and G. Oritz, *Elements of Phase Transition and Critical Phenomena.* Oxford University Press, 2011.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Capacity Scaling Bounds in Wideband Cellular Networks

Felipe Gómez-Cuba[†], *Member, IEEE*, Sundeep Rangan[‡], *Fellow, IEEE*,
Elza Erkip[‡], *Fellow, IEEE*, Francisco J. González-Castaño[†]
[†]AtlantTIC, University of Vigo, C.P. 36310 Vigo, España {fgomez,javier}@gti.uvigo.es
[‡]NYU Tandon School of Engineering, Brooklyn, NY 11201, USA {elza,srangan}@nyu.edu

***Abstract*—Future generations of cellular networks will have higher node density, larger number of antennas per node, and wider bandwidth. This paper develops capacity scaling laws for such networks by explicitly incorporating the impact of bandwidth on capacity. The main contribution is a capacity scaling upper bound for a cellular-type network architecture without base station cooperation. This upper bound is shown to be tight, as it provides the same scaling as an infrastructure multi-hop protocol introduced in previous work. The results show that large cellular networks transition from bandwidth-limited to power-limited capacities depending on the scaling of the bandwidth compared to the scaling of the number of nodes and that single hop protocols are suboptimal except when the bandwidth scaling is small.**

***Index Terms*—Wideband, capacity scaling, cellular network.**

## I. Introduction

Wireless standardization bodies have begun to consider deployment of millimeter wave (mmWave) technology with frequencies in the range $30 - 300$GHz, leading to large amounts of spectrum available for communication [1]. Other technologies that increase available system bandwidth, such as carrier aggregation across bands and opportunistic cognitive reuse of occupied bands, are also being considered for future generations of cellular networks.

The conventional analysis of a point-to-point wideband channel exhibits a transition from a high-SNR regime where rate grows with bandwidth to a low-SNR regime where rate is power-limited [2]. However, since practical networking protocols usually divide the available bandwidth among nodes in some form, the fact that system bandwidth is increasing does not necessarily imply that a network with large number of nodes will experience the effects described in the wideband point-to-point channel model.

While the exact capacity region of a large network is not known, capacity scaling laws provide a useful framework that characterizes the growth of the capacity region as the number of nodes increases [3]. In order to quantify the impact of increasing bandwidth on network capacity, considered in the light of other resources that also increase the available degrees of freedom, this paper provides a scaling law analysis of a cellular network operating in the wideband regime.

Work supported in part by COINS, FPU2012/01319, NSF Grants # 1302336 and 1547332, and NYU WIRELESS.

From an information theoretic perspective, mmWave, carrier aggregation, massive Multiple Input Multiple Output (MIMO) and dense cellular deployments are all, in essence, various ways to increase the fundamental degrees of freedom of the network: bandwidth, antennas and infrastructure density. To evaluate the potential value of each of these technologies, this paper derives an upper bound on the scaling of per node capacity of cellular networks under parametric scaling of these dimensions. Our analysis follows along the lines of the classic scaling laws results [3]–[9], but is applied to cellular networks rather than ad-hoc networks. Specifically, we consider a large cellular network with $n$ mobile nodes, with various scalings in $n$ of parameters such as the bandwidth, area, number of base stations (BSs) and number of BS antennas. In addition, we consider that BSs do not cooperate and nodes communicate with their closest BS, creating two traffic flows usually found in cellular networks, consisting of uplink and downlink transmissions. This produces different scaling properties than those in ad-hoc networks with infrastucture assistance [7]–[9].

Considering a deterministic channel model, our main result determines an upper bound to the throughput capacity scaling using cut-set arguments, by separating each single BS from the rest of the network. The cut of the network effectively forms a combined point-to-point MIMO system that displays a behavior resembling that of point-to-point channels. The upper bound meets a lower bound achieved by an infrastructure multi-hop (IMH) protocol whose throughput scaling was described for fading channels in [10], thereby establishing the capacity of cellular networks without BS cooperation, with two scaling regimes:

- In the *bandwidth limited* regime, bandwidth grows slower than the power that can be delivered towards the aggregate network, and throughput scales with the degrees of freedom of the network.
- In the *power limited* regime, bandwidth grows faster than the power transfer, and throughput scales at most with the power delivered to the nodes.

The rest of the paper is organized as follows. In Section II we provide a brief overview of the literature on capacity scaling laws analysis. In Section III we describe our models for a large cellular network with increasing number of users, and the channel between terminals. In Section IV we describe the

upper bound on the throughput capacity, establish the optimal throughput scaling and discuss different operating regimes. Finally, Section V concludes the paper.

## II. Overview of Capacity Scaling Laws

Capacity scaling law analysis started with the seminal work of Gupta and Kumar [3] where they studied the scaling of the rate $R(n)$ achieved by each node in an ad-hoc network with $n$ nodes distributed in a unit area (called *dense* network). Their protocols exhibit a scaling of $R(n) \propto \Theta(\frac{1}{\sqrt{n}})$[1]. This suggests that using the protocols in [3] it is not possible to increase the number of nodes in network arbitrarily without sacrificing rates. Similar results exist for *extended* network models [4], where $n$ users spread over an area $\Theta(n)$ with constant user density, instead of a fixed area with user density $\Theta(n)$.

Ozgur, Lévêque and Tse [4] introduced *hierarchical cooperation* (HC), and showed that HC achieves linear scaling $R(n) = \Theta(1)$ in dense networks and improves the Gupta-Kumar result to $\Theta(n^{2-\alpha/2})$ in extended networks with a low path-loss exponent $\alpha < 3$. Under HC, nodes cooperate to form virtual antenna arrays, and for a sufficiently high number of layers of cooperation, it would be possible to break the original limitation and grow the number of nodes arbitrarily without incurring any rate penalty. However, Franceschetti, Migliore and Minero [5] pointed out that the physical degrees of freedom of a signal within a bounded area are finite. It would be unrealistic to assume that as the number of nodes increases, all channel coefficients remain independent. Thus the result in [4] would be an artifact of optimistic independence assumptions. Considering these constraints, [5] obtained an ultimate limitation to scaling of $R(n)$ as $\Theta(\frac{\log(n)^2}{\sqrt{n}})$.

In [6], Ozgur, Johary, Tse and Lévêque proposed an argument to harmonize the results of [4] and [5]. Even though for very high $n$ channels do become dependent, this occurs at values of $n$ so high that there would exist first a transitory regime with high, but finite, values of $n$ where the linear scaling analysis holds. In [6] the same authors also introduced the concept of *operating regimes*, by allowing area to scale with an arbitrary exponent of $n$, $A \propto n^{\nu}$. A threshold on exponent $\nu$ separates two regimes: for small $\nu$ network capacity behaves similarly than the dense network model and for $\nu$ above the threshold network capacity behaves as in the extended network model.

Recently, the ability of HC to achieve linear scaling was put into question in [12]. It was found that under practical limitations, the optimal number of layers in a HC implementation would be small, contradicting the theoretical analysis where rate improves with the number of layers and for achieving the linear scaling a very large number of layers is necessary.

There have been other extensions of scaling law analysis in ad-hoc networks introducing cooperation, mobility, broadcast, infrastructure or large bandwidth. Readers are referred to [13] for a comprehensive review.

Most literature on scaling laws follows ad-hoc network models, which have different traffic demand than a cellular network. Even though [7]–[9] have modeled ad-hoc networks with infrastructure support, the use of infrastructure in these models is only as an intermediary to help the delivery of ad-hoc type communications. In these works, data always flows from one node to another, with destinations randomly picked across the network. More importantly, in such models it is always possible to fall back to pure ad-hoc protocols ignoring the presence of infrastructure when this is beneficial.

### A. Our Contributions

In this paper, rather than ad-hoc networks supported by infrastructure, we consider the traffic flows typical in cellular networks, where each node sustains uplink and downlink data flows with its closest BS. On the one hand, this reduces the typical distance between source-destination pairs; on the other hand, this model may cause rates of many users to concentrate at the same BS causing bottlenecks that cannot be avoided by dropping the infrastructure and falling back to pure ad-hoc protocols. Both our approach and that of [8] have in common the presence of infrastructure with arbitrary scaling density, but the difference in the traffic renders the ultimate scaling limitations very different.

The main innovation of our analysis is including the impact of very large bandwidths in capacity scaling. This provides a characterization of a bandwidth threshold beyond which a large network stops benefiting from the bandwidth increase and experiences power limitations, mirroring the well-known fact that point-to-point links become power-limited when bandwidth is large.

Most scaling analyses [3]–[9] considered a fixed bandwidth. However, a network with a fixed bandwidth would only exhibit low SNRs in long-distance links with a high path loss, scaling with the dimensions of the network and unrelated to bandwidth. This occurs because it is always possible to slice the constant bandwidth in small narrowband chunks as the number of nodes $n$ grows. In order to study large system bandwidth $W$, one could have $W \to \infty$ and *then* let $n$ grow, as in [14]. Nonetheless, this method forces the network to be always power-limited, rather than providing insights on the bandwidth scaling necessary to enter power-limitation, and its interplay with network architecture and rates. In our model, the goal is to investigate what happens between the two extremes; for this we take limits on $W$ and $n$ increasing to infinity at the same time, following an exponential relation:

$$\psi := \lim_{n, W \to \infty} \frac{\log W}{\log n}, \tag{1}$$

and the cases in the literature correspond to $\psi = 0$ and $\psi = \infty$.

## III. Network and Channel Models

### A. Network Model

We consider a sequence of cellular wireless networks indexed by $n$, where $n$ is the number of single-antenna nodes

[1] We use the standard $f(n) = O(g(n))$, $f(n) = \Omega(g(n))$ and $f(n) = \Theta(g(n))$ notations [11] to respectively represent that at sufficiently high $n$ function $f(n)$ becomes less than or equal than $g(n)$, greater than or equal to $g(n)$, and identical to $g(n)$ up to a constant factor.

uniformly distributed in an area $A$. The network is supported by $m$ BSs that do not cooperate, with $\ell$ antennas each, and communication takes place over an increasing bandwidth $W$.

Table I defines the scaling relation between $n$ and the different network parameters. Here $W_0$, $A_0$, $m_0$, $l_0$, $k_0$ are fixed constants. The exponents of the number of BSs and BS antennas are taken from [8]. The constraint $\beta+\gamma \leq 1$ ensures that the number of infrastructure antennas per node does not grow without bound. The scaling of the network area is as proposed by [6] to model a continuum of operating regimes between *dense* ($\nu = 0$) and *extended* ($\nu = 1$) networks. We introduce the bandwidth scaling exponent $\psi$ as shown in (1). Note that $\psi < 1$ represents that bandwidth per node decreases as the number of nodes increases, while $\psi > 1$ represents asymptotically infinite bandwidth per node.

We consider BSs that are placed at fixed distances of each other, dividing the network area in regular hexagonal cells around each BS with radius $r_{\text{cell}}$ with asymptotically $\frac{n}{m}$ nodes each. The downlink from the BS to the nodes and the uplink from the nodes to the BS operate independently in alternate time division duplex (TDD) frames. This imposes a $\frac{1}{2}$ penalty in rate but otherwise does not alter the scaling of capacity with $n$. Note that BSs cannot receive in the downlink phase or transmit in uplink, while nodes can do both.

Due to random node placement, the rate achievable by any individual user is a random variable which depends on the particular downlink or uplink protocol used. The following definitions are adapted from [3].

**Definition 1.** *A downlink (uplink) rate of $R_{\text{DL}}(n)$ ($R_{\text{UL}}(n)$) bits per second per node is feasible in a realization of the cellular network if there exists a protocol that achieves in all nodes.*

The feasible rate is evaluated on a realization of the random node locations, and feasibility of some rates will depend on distances in a specific layout. In the following definition we remove the randomness of Def. 1 by requiring a rate scaling at the frontier of feasible rates with probability one.

**Definition 2.** *The downlink (uplink) per node throughput capacity scaling $C_{\text{DL}}(n)$ ($C_{\text{UL}}(n)$) of a set of random cellular networks is of the order $\Theta(f(n))$ bits per second per node if there are constants $c_1 < c_2$ such that.*

$$\lim_{n\to\infty} P\left(R_{\text{DL}}(n) = c_1 f(n)\right) = 1 \tag{2}$$

$$\lim_{n\to\infty} P\left(R_{\text{DL}}(n) = c_2 f(n)\right) < 1 \tag{3}$$

Table I
SCALING EXPONENTS OF NETWORK PARAMETERS

| Exponent | Range | Parameter (vs. no. of nodes $n$) |
|---|---|---|
| $\psi$ | $[0,\infty)$ | Bandwidth $W = W_0 n^{\psi}$ |
| $\nu$ | $[0,1]$ | Area $A = A_0 n^{\nu}$ |
| $\beta$ | $[0,1]$ | No. of BSs $m = m_0 n^{\beta}$ |
| $\gamma$ | $[0,1-\beta]$ | No. of BS antennas $\ell = \ell_0 n^{\gamma}$ |

### B. Channel Model

Between a transmitter $t$ and receiver $r$, we consider a MIMO additive white Gaussian noise channel with deterministic full rank matrix $\mathbf{H}_{t,r} \in \mathbb{C}^{\ell_{\text{t}}\times\ell_{\text{r}}}$. Each entry of the channel matrix has unit gain and an arbitrary phase, $h_{t,r}^{(i,j)} = e^{2\pi j\theta_{i,j}}$, so that the channel squared norm satisfies $|\mathbf{H}|^2 = \ell_{\text{r}}\ell_{\text{t}}$. The distance between transmitter and receiver, $d_{t,r}$, defines the macroscopic pathloss gain $d_{t,r}^{-\frac{\alpha}{2}}$. Average transmission power constraints of nodes and BSs are $P$, and $P_{\text{BS}}$, respectively. Our results can be extended to random fading models with moderate effort. The signal at the receiver is given by

$$\mathbf{y}_r = d_{t,r}^{-\frac{\alpha}{2}}\mathbf{H}_{t,r}\mathbf{x}_t + \mathbf{z}_r \tag{4}$$

where $\mathbf{x}_t$ is the signal transmitted by $t$ with period $T_{\text{s}} = 1/W$, satisfying $\text{E}\left[|\mathbf{x}_t|^2\right] \leq \frac{P_t}{W}$. Here $P_t$ depends on the type of transmitter and the fraction of its power dedicated towards $r$. The thermal noise at the receiver is $\mathbf{z}_r \sim \mathcal{CN}(0, N_0\mathbf{I}_{\ell_{\text{r}}})$.

In practice it is expected that mmWave channels do not have a rich enough scattering to display full rank channel matrices in the *physical* arrays, but our results apply all the same to these non-full rank channels by appropriately projecting the dimensions of the antenna arrays to a smaller subspace and a small-dimensional full rank matrix that captures the equivalent *effective array dimensions*. Hereafter, we will use the term "number of antennas $\ell$" to refer to the *effective independent antenna array dimensions* and represent by $\ell_{\text{t}}$ and $\ell_{\text{r}}$ the effective number of transmit and receive antennas.

The upper bounds developed in this paper consider cuts separating one transmitter from the rest of the network, applying this channel model to one virtual transmitter and one virtual receiver containing all the antennas on each side of the cut. As we argue in Sec. IV, the upper bound is achievable in a scaling law sense and, even though we do not consider full BS cooperation, it represents the scaling with perfect interference suppression. The achievable schemes that illustrate this were presented originally for non-coherent fading channels in [10]; here they have been adapted to follow the channel model (4) with proper modifications to incorporate interference from other cells.

## IV. CHARACTERIZATION OF CAPACITY SCALING

In this section we first present an upper bound to the throughput capacity scaling of large cellular networks, and we next illustrate that an adaptation of the multi-hop protocol presented in [10] for fading channels to our channel model achieves this scaling.

**Theorem 1.** *The downlink throughput capacity scaling $C_{\text{DL}}(n)$ of random cellular networks is upper bounded by*

$$\Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(1-\nu)\frac{\alpha}{2}\right)}\right), \tag{5}$$

*and the uplink throughput capacity scaling $C_{\text{UL}}(n)$ by*

$$\Theta\left(n^{\min\left(\psi,(1-\nu)\frac{\alpha}{2}\right)}\right), \tag{6}$$

*both with probability 1 as $n \to \infty$.*

*Proof.* We introduce the detailed analysis for downlink. Uplink follows similarly.

Since our network model assumes no cooperation between BSs, we obtain an upper bound of the sum-rate of the users served by each BS by considering a cut separating that BS from the rest of the network. At the receiving side of the cut there is perfect cooperation among $n$ receiver nodes and another $m-1$ BS transmitters, resulting in perfectly-known interference that can be canceled.

This reduces the communication problem into a point-to-point MIMO channel consisting of a single transmitter-receiver pair with dimensions $\ell_{\mathrm{t}} = \ell$ $\ell_{\mathrm{r}} = n$. We represent the distance from each node $r$ to BS $t$ in a diagonal matrix $\mathbf{D}_t \triangleq \{D_t^{ii} = d_{t,i}^{-\frac{\alpha}{2}}\}$, and write the signals from all BSs to all nodes by extending (4) as

$$\mathbf{y} = \mathbf{D}_t\mathbf{H}_t\mathbf{x}_t + \underbrace{\sum_{t'\neq t}\mathbf{D}_{t'}\mathbf{H}_{t'}\mathbf{x}_{t'}}_{\text{known to all receivers}} + \mathbf{z} \tag{7}$$

where $\mathbf{H}_t$ refers to the small scale fading channel matrix between BS $t$ and all receivers. All nodes and the other BSs cooperate perfectly, canceling the second term and leaving a non-interfering point-to-point MIMO channel.

Hence, the DL sum rate on the cell of BS $t \in [1, m]$, denoted by $T_{\mathrm{DL}-t}(n)$ is bounded by

$$T_{\mathrm{DL}-t}(n) \leq \max_{\mathbf{Q}_{x_t}} W \log\det\left(\mathbf{I}_n + \frac{1}{WN_0}\mathbf{D}_t\mathbf{H}_t\mathbf{Q}_{x_t}\mathbf{H}_t^H\mathbf{D}_t^H\right) \tag{8}$$

where we represent by $\mathbf{Q}_{x_t}$ the normalized covariance matrix of the transmitted signal $\mathbf{x}_t$, with its power constraint expressed as $\mathrm{tr}\{\mathbf{Q}_{x_t}\} \leq P_{\mathrm{BS}}$.

By the assumption that channel matrices are full rank and $\ell \leq n$, following standard arguments in [2, (7.10)], it can be shown that the upper bound in (8) can be expressed as a power allocation over the eigenvalues $\lambda_i$ of the matrix $\mathbf{D}_t\mathbf{H}_t$

$$T_{\mathrm{DL}-t}(n) \leq \max_{P_i} W \sum_{i=1}^{\ell} \log\left(1 + \frac{P_i\lambda_i^2}{WN_0}\right) \tag{9}$$

We can use the trace of the matrix to upper bound each eigenvalue separately $\lambda_i^2 \leq \sum_{i=1}^{\ell}\lambda_i^2 = \mathrm{tr}\{\mathbf{D}_t\mathbf{H}_t\mathbf{H}_t^H\mathbf{D}_t^H\} \leq \ell\sum_{r=1}^{n} d_{t,r}^{-\alpha}$. Due to the fact that all terms in this upper bound are equal and the convexity of the logarithm, the power allocation $P_i^* = \frac{P_{\mathrm{BS}}}{\ell}$ maximizes this upper bound. Hence

$$T_{\mathrm{DL}-t}(n) \leq W\ell\log\left(1 + \frac{P_{\mathrm{BS}}}{WN_0}\sum_{r=1}^{n} d_{t,r}^{-\alpha}\right) \tag{10}$$

Notice that if $\lim_{n\to\infty}\frac{P_{\mathrm{BS}}\sum_{r=1}^{n} d_{t,r}^{-\alpha}}{WN_0} = \infty$, then the upper bound in (10) becomes degrees-of-freedom-limited. In this regime (10) scales as $\Theta(W\ell)$.

Conversely, only if all links produce a low power at the same time, satisfying $\lim_{n\to\infty}\frac{P_{\mathrm{BS}}\sum_{r=1}^{n} d_{t,r}^{-\alpha}}{WN_0} = 0$, the network is in the power-limited regime and (10) scales as $\Theta(\ell P_{\mathrm{BS}}\sum_{r=1}^{n} d_{t.r}^{-\alpha})$.

The sum $\sum_{r=1}^{n} d_{t,r}^{-\alpha}$ can be calculated using the exponential stripping method described in [15]. Consider a series of concentric rings centered at the BS $t$ with inner radius $r_i = n^{\frac{\nu}{2}}e^{\frac{-i}{2}}$ and outer radius $r_{i-1}$. Recall that the user density scales as $n^{1-\nu}$ and network area as $n^{\nu}$, thus the number of nodes contained in each disc is $S_i \leq ne^{1-i}$ with high probability. Using this, we can upper bound the sum over $n$ by summing over these discs and lower-bounding distance in each disc by the inner radius. Moreover, we have that an area of $n^{\nu-1}$ is the smallest that contains one node w.h.p. so the sum ends at $i \leq \lfloor\log n\rfloor + 1$, assigning distance $r^{\frac{\nu-1}{2}}$ to the last term instead.

$$\begin{aligned}\sum_{r=1}^{n} d_{t,r}^{-\alpha} &\leq \sum_{i=1}^{\lfloor\log n\rfloor+1} S_i r_i^{-\alpha} \\ &\leq \left[\sum_{i=1}^{\lfloor\log n\rfloor} ne^{1-i}n^{-\nu\frac{\alpha}{2}}e^{+i\frac{\alpha}{2}}\right] + en^{(1-\nu)\frac{\alpha}{2}} \\ &\leq n^{-\nu\frac{\alpha}{2}}\left[\log ne^{1+\frac{\alpha}{2}\log(n)}\right] + n^{(1-\nu)\frac{\alpha}{2}}e \\ &\leq (\log n + 1)n^{(1-\nu)\frac{\alpha}{2}}e\end{aligned} \tag{11}$$

where the third inequality is due to $e^{+i\frac{\alpha}{2}} \leq e^{\max(i)\frac{\alpha}{2}}$.

Examining Table I this leads to

$$T_{\mathrm{DL}-t}(n) = \begin{cases}\Theta\left(n^{\gamma+\psi}\right) & \psi \leq \frac{\alpha}{2}(1-\nu) \\ \Theta\left(n^{\gamma+\frac{\alpha}{2}(1-\nu)}\right) & \psi > \frac{\alpha}{2}(1-\nu)\end{cases} \tag{12}$$

Now, by symmetry, each of the $m$ single-BS cuts gives the same upper bound to the rate of the users served by that BS. By the requirement that feasible rate is guaranteed to all users, the throughput capacity of the network is upper bounded by

$$R_{\mathrm{DL}}(n) \leq \frac{m}{n}\min_t T_{\mathrm{DL}-t}(n) = \Theta(n^{\beta+\gamma-1+\min(\psi,(1-\nu)\frac{\alpha}{2})})$$

completing the proof of Theorem 1 for DL.

A similar set of arguments lead to the bound for the uplink. In this case we consider $n$ cuts, each separating one user node from the rest of the network. In this cut, all the BSs and the remaining $n-1$ nodes are receivers, and their mutual interference canceled. Due to the fact that the transmitting node has a single antenna (eigenvalue), the degrees of freedom are $\Theta(W)$; and the sum term over all receiving devices (equivalent of (11)) is

$$\sum_{r=1}^{n-1} d_{t,r}^{-\alpha} + \ell\sum_{r=1}^{m} d_{t,r}^{-\alpha} = \Theta(n^{(1-\nu)\frac{\alpha}{2}}).$$

Then the upper bound on uplink feasible rate becomes

$$R_{\mathrm{UL}}(n) \leq \min_t T_{\mathrm{UL}-t}(n) = \Theta\left(n^{\min(\psi,(1-\nu)\frac{\alpha}{2})}\right) \tag{13}$$

□

The next theorem shows that, after adapting the throughput scaling to the channel model in this paper, the Infrastructure Multi-Hop (IMH) protocol introduced in [10, Th. 2] can achieve the upper bound.

**Theorem 2.** *The IMH protocol achieves the upper bound in Theorem 1 and characterizes downlink throughput capacity as*

$$C_{\mathrm{DL}}(n) = \Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{14}$$

*and when $\beta+\gamma=1$ the uplink throughut capacity as*

$$C_{\mathrm{UL}}(n) = \Theta\left(n^{\min\left(\psi,(1-\nu)\frac{\alpha}{2}\right)}\right) \tag{15}$$

The downlink rate of Infrastructure Single-Hop (ISH) using direct transmissions [10, Th. 1] is only $\Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(\beta-\nu)\frac{\alpha}{2}\right)}\right)$ for the channel model used in this paper, and only achieves capacity if the single-receiver power at cell edge is high, when $\psi < (\beta-\nu)\frac{\alpha}{2}$, or in the particular case of maximum BS density $\beta = 1$. Both achievable schemes and the capacity scaling upper bound are compared in Fig. 1 for downlink transmission. A similar relation between the protocols and the bound exists for uplink.

Comparing the two cases of the capacity scaling on Theorem 2, we can distinguish two operating regimes.

*Bandwidth limited regime:* If $\psi < \frac{\alpha}{2}(1-\nu)$, the network capacity grows with $W$. It must be noted that this regime of the capacity scaling relies on the received power at all nodes in the network at once, and does not guarantee that any specific single node can receive a high-SNR in the absence of receiver cooperation. Indeed, as we argue above, the use of independent direct transmissions as in ISH protocol does not always achieve capacity scaling.

*Power limited Regime:* If $\psi > \frac{\alpha}{2}(1-\nu)$, the network capacity does not grow with $W$. Network capacity scaling is bounded by the power that can be transfered from one BS to all nodes in the network at once. The distances between BSs and the nodes are sufficiently far that the SNR in (10) goes to zero. In this regime no node can receive degrees-of-freedom limited rates even with cooperation.

Finally, note that these scaling laws make intuitive sense because, with probability 1 as $n \to \infty$, a disc with radius $\Theta(n^{\left(\frac{(\nu-1)}{2}\right)})$ around a BS contains one node, which combined with array gain $n^{\gamma}$ gives the best-case transfer of power between a single BS and the rest of the network. Also, the degrees of freedom of the cellular network cannot exceed $\Theta(Wm\ell)$.

## V. Conclusions

As cellular networks evolve, the node density, number of BS antennas and bandwidth increase. Wireless network capacity scales with these increasing degrees of freedom only if received power is not overspread. In this paper we have provided a characterization of cellular capacity scaling that exhibits a bandwidth-limited and a power-limited regime. Moreover, only in a fraction of the first regime capacity is achievable using independent non-cooperative direct transmission between the BS and each node in its cell, whereas for sufficiently large bandwidth cooperation or multi-hop is essential to achieve throughput capacity scaling. While traditional cellular networks typically operate in the bandwidth-limited regimes, future cellular networks with large bandwidth could experience power-limited scaling regimes, therefore necessitating multi-hop communications.

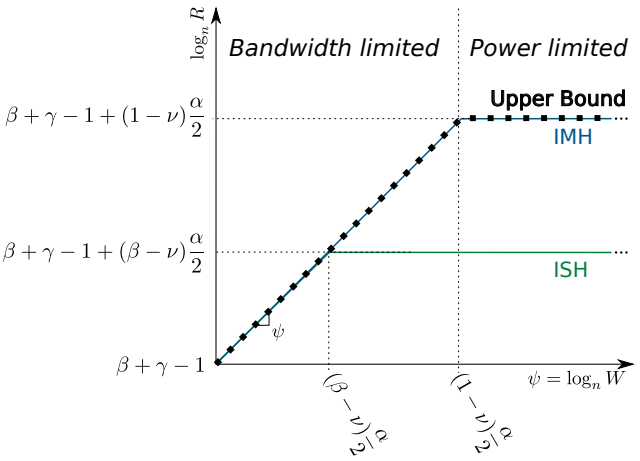



Figure 1. Exponents of downlnk rates vs bandwidth.

## References


[1] T. S. Rappaport, R. W. Heath Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Pearson Education, 2014.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge university press, 2005.

[3] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 11, p. 3117, 2003.

[4] A. Ozgur, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves linear capacity scaling in ad hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3549–3572, 2007.

[5] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3413–3424, aug 2009.

[6] A. Ozgur, R. Johari, D. N. C. Tse, and O. Lévêque, "Information-theoretic operating regimes of large wireless networks," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 427–437, jan 2009.

[7] U. C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in *International conference on Mobile computing and networking*, 2003.

[8] W.-y. Shin, S.-W. Jeon, N. Devroye, M. H. Vu, S.-y. Chung, Y. H. Lee, and V. Tarokh, "Improved capacity scaling in wireless networks with infrastructure," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5088–5102, aug 2011.

[9] C. Jeong and W.-Y. Shin, "Ad hoc networking with rate-limited infrastructure: Generalized capacity scaling," in *IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 61–65.

[10] F. Gómez-Cuba, S. Rangan, and E. Erkip, "Scaling laws for infrastructure single and multihop wireless networks in wideband regimes," in *IEEE International Symposium on Information Theory (ISIT)*, 2014.

[11] D. E. Knuth, "Big Omicron and big Omega and big Theta," *ACM SIGACT News*, vol. 8, no. 2, pp. 18–24, 1976.

[12] S.-N. Hong and G. Caire, "Demystifying the scaling laws of dense wireless networks: No linear scaling in practice," in *IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 71–75.

[13] N. Lu and X. S. Shen, "Scaling Laws for Throughput Capacity and Delay in Wireless Networks — A Survey," *IEEE Communications Surveys & Tutorials*, pp. 1–16, 2013.

[14] R. Negi and A. Rajeswaran, "Capacity of power constrained ad-hoc networks," in *IEEE INFOCOM*, 2004.

[15] M. Franceschetti, "A note on Lévêque and Telatar's upper bound on the capacity of wireless ad hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3207–3211, 2007.

International Zurich Seminar on Communications (IZS), March 2 – 4, 2016

# On the Secrecy Capacity of the Z-Interference Channel

Ronit Bustin
Dept. of Electrical Engineering
Tel Aviv University
Email: ronitbustin@post.tau.ac.il

Mojtaba Vaezi
Dept. of Electrical Engineering
Princeton University
Email: mvaezi@princeton.edu

Rafael F. Schaefer
Information Theory and Applications Group
Technische Universität Berlin
Email: rafael.schaefer@tu-berlin.de

H. Vincent Poor
Dept. of Electrical Engineering
Princeton University
Email: poor@princeton.edu

***Abstract*—The two-user Z-interference channel with an additional secrecy constraint is considered. The two transmitter-receiver pairs wish to reliably transmit their messages; however the transmission of the first pair both interferes with the transmission of the second pair and is also required to be completely secure from the second receiver. The focus here is on the capacity region of the above Z-interference channel in the Gaussian case under the standard power constraints. The maximum rates of the two users in this setting are described, and although the maximum rate of the transmission of the first pair has a single-letter expression, due to Wyner's secrecy capacity expression, its maximization is non-trivial. The significance of a stochastic encoder for the second transmitter, encoding a message which is not required to comply with any secrecy constraints, is noted. It is shown explicitly that constraining this encoder to be deterministic reduces the capacity region. Finally, a Sato-type outer bound on the capacity region is obtained under this additional deterministic encoder constraint.**

## I. Introduction

The interference channel is a central open problem in multi-user information theory. Understanding the effect of interference is critical to the understanding of the limitations of communication and essentially the interactions in any network. The basic aspects of interference appear already in the simplest setting - the two-user Z-interference channel. This channel comprises two independent inputs $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ and two outputs $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ (throughout the paper bold letters denote length $n$ random vectors) with a channel conditional distribution of the following form:

$$P_{\boldsymbol{Y}_1,\boldsymbol{Y}_2|\boldsymbol{X}_1,\boldsymbol{X}_2} = P_{\boldsymbol{Y}_1|\boldsymbol{X}_1} P_{\boldsymbol{Y}_2|\boldsymbol{X}_1,\boldsymbol{X}_2}. \tag{1}$$

The open question for this channel is: given two independent messages $W_1 \in [1, 2^{n\mathsf{R}_1}]$ and $W_2 \in [1, 2^{n\mathsf{R}_2}]$, what are the rates $\mathsf{R}_1$ and $\mathsf{R}_2$ that can be reliably transmitted through this channel? Due to the importance of this problem it had attracted considerable attention throughout the years. Many results have been obtained; however, in general, the problem is still open. We refer the reader to recent overviews of the problem given in the introductory sections of [1] and [2].

The difficulty of this problem lies in the fact that we do not have a good understanding of the effect of a transmission on other (unintended) users. Moreover, as can be seen in the additive Gaussian white noise (AWGN) setting there is a rivalry between the two users [3], meaning that when one transmits at its maximum rate it also causes the maximum disturbance on the other user (a phenomenon known as the "worst additive noise" result [4]).

In this work we place an additional requirement on the interfering signal. Beyond its reliable decoding at its receiver we also require complete secrecy of this message at the interfered-with receiver. This additional requirement is relevant to many practical settings, in which our transmission can both be received by other, unintended receivers, but still we would like it to remain secure. As will be shown here this additional requirement provides interesting observations and many open questions.

We begin by formally stating the complete secrecy requirement:

$$\lim_{n\to\infty} \frac{1}{n} I\left(W_1; \boldsymbol{Y}_2\right) \to 0 \tag{2}$$

which assures complete secrecy of the interfering message at $\boldsymbol{Y}_2$. Note that due to this complete secrecy constraint we may consider only "weak interference". Figure 1 depicts the model.

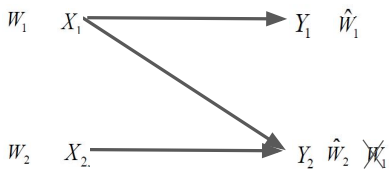



Fig. 1. The Z-Interference channel with a complete secrecy constraint for the interfering message at the interfered-with reciever.

This initial work focuses on the AWGN setting; however, as will be evident, many of the claims can be extended to discrete memoryless channels. Thus, we consider the following model for a single use of the channel:

$$\begin{aligned} Y_1 &= \sqrt{\mathsf{snr}_1} X_1 + N_1 \\ Y_2 &= \sqrt{\mathsf{snr}_2} X_2 + \sqrt{a\mathsf{snr}_1} X_1 + N_2 \end{aligned} \tag{3}$$

where $N_1$ and $N_2$ are standard additive Gaussian noise terms which can be assumed to be independent of each other and

The work of R. Bustin was supported in part by the women postdoctoral scholarship of Israel's Council for Higher Education (VATAT) 2014-2015, in part by the the U. S. Army Research Office under MURI Grant W911NF-11-1-0036, and in part by the U. S. National Science Foundation under Grants CMMI-1435778 and ECCS-1343210.

from channel use to channel use. They are also independent of the transmissions of the two users $X_1$ and $X_2$ which are independent of each other (no cooperation between the transmitters). We assume $a \in [0,1)$ which is the "weak interference" regime. There is also an average power constraint of 1 on both channel inputs.

In this work we distinguish between two scenarios. First, we do not limit either encoder, and allow stochastic encoders at both transmitters. Under this assumption we first examine the bounding box of the capacity region, that is, the maximum possible rates either user can obtain. This proves to be an interesting problem for which we can show, using the methods in [3] and [5] that the rate of the interfering message can increase beyond that obtained by simply having the interfered-with transmitter transmit Gaussian noise. Second, we consider a very probable case in which the interfered-with transmitter which has no secrecy requirements of its own is limited to a deterministic encoder. We observe that this limitation reduces the capacity region and provide a Sato-type [6] outer bound on its capacity region.

## II. Stochastic Encoders - Capacity Region Bounding Box

The first question that comes to mind when considering the above problem concerns the bounding box of its capacity region. That is, we wish to know what is the maximum rate for either $W_1$ or $W_2$ regardless of the rate of the other user. For $W_2$ if we take $\mathsf{R}_1 = 0$ we comply with the secrecy constraint in the trivial sense (no information is transmitted), and in addition there is no interference; thus $\mathsf{R}_2 = \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_2)$ can be achieved, which is, of course, the maximum possible rate. On the other hand, the maximum value of $\mathsf{R}_1$ is an open problem. Examining (3) we can see that $\boldsymbol{Y}_2$ is a *degraded* version of $\boldsymbol{Y}_1$ (since $a \in [0,1)$), and thus we have that $\boldsymbol{X}_1 - \boldsymbol{Y}_1 - \boldsymbol{Y}_2$ is a Markov chain regardless of the distribution of $\boldsymbol{X}_2$. Using the wiretap result for a *degraded* channel [7] we have a single-letter expression

$$\mathsf{R}_{1,\max} = \max_{P_{X_1}P_{X_2}} \left\{ I\left(X_1;Y_1\right) - I\left(X_1;Y_2\right) \right\} \tag{4}$$

where the maximization is over both distributions, as $Y_2$ depends on $X_2$ as well.

Note that if $P_{X_1}$ is a Gaussian distribution then the optimal choice for $P_{X_2}$ is also Gaussian due to the "worst additive noise" lemma [4]. However, there is a dependence between the two distributions and they must be optimized jointly. In order to break this dependence one can invoke the entropy power inequality (EPI), which provides an upper bound that is attained with equality if and only if both $X_1$ and $X_2$ are Gaussian. However, attempting to optimize this upper bound results in a Gaussian distribution for $X_1$ but not for $X_2$. This observation leads us to the next result following the approach of Abbe and Zheng [5], a proof of which is given in the appendix.

**Theorem 1.** *For any* $\mathsf{snr}_1 > 0, \mathsf{snr}_2 > 0$ *and any* $a \in [0,1)$, $\mathsf{R}_{1,\max}$ *is obtained by non-Gaussian distributions in (4), meaning*

$$\mathsf{R}_{1,\max} > \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_1) - \frac{1}{2}\mathsf{log}\left(1+\frac{a\mathsf{snr}_1}{1+\mathsf{snr}_2}\right). \tag{5}$$

## III. Deterministic Interfered-With Encoder

In this section we restrict the encoder of the interfered-with transmitter to the class of deterministic encoders. Note that this transmitter has no secrecy constraints on its message, thus making this a very reasonable assumption. The advantage of this restriction is in allowing us to follow the approach of Sato [6] and Costa [8] and provide a good outer bound on the capacity of this channel. However, as we will show, this restriction, although reasonable from a practical viewpoint, reduces the capacity region of this channel.

We begin this section with the following result that extends the result of Costa [8] to our setting:

**Lemma 1.** *The Gaussian Z-Interference channel with secrecy constraint and a deterministic encoder for the message* $W_2$, *meaning* $H(\boldsymbol{X}_2|W_2) = 0$, *is equivalent, in the sense that they have the same capacity region, to the* degraded *Gaussian interference channel:*

$$\begin{aligned} \boldsymbol{Y}_1' &= \sqrt{\mathsf{snr}_1}\boldsymbol{X}_1 + \sqrt{\frac{\mathsf{snr}_2}{a}}\boldsymbol{X}_2 + \boldsymbol{N}_1 \\ \boldsymbol{Y}_2' &= \sqrt{\mathsf{snr}_1}\boldsymbol{X}_1 + \sqrt{\frac{\mathsf{snr}_2}{a}}\boldsymbol{X}_2 + \boldsymbol{N}_1 + \boldsymbol{N}_2' \end{aligned} \tag{6}$$

*where* $\boldsymbol{N}_1$ *is as defined above, standard additive Gaussian noise, whereas* $\boldsymbol{N}_2'$ *is additive Gaussian noise of variance* $\frac{1-a}{a}$.

*Proof:* Following the proof of Costa [8] we refer to [8, Figure 6]. Note that the equivalence between [8, Figure 6-(a)] and [8, Figure 6-(c)] holds for the same reasons as in [8]. The equivalence to [8, Figure 6-(d)], the *degraded* Gaussian interference channel requires more delicacy. Note that as claimed in [8, Appendix A] the capacity region of [8, Figure 6-(a)] contains the capacity region of [8, Figure 6-(d)], with the additional secrecy constraints. This is due to the fact that $\boldsymbol{Y}_1$ is a better version than the equivalent output in the [8, Figure 6-(d)]. The reverse claim follows if we assume that $H(\boldsymbol{X}_2|W_2) = 0$ by following the proof in [8, Appendix A]. ∎

The above equivalence is limited to the case of deterministic encoders for the interfered-with transmitter. This transmitter is not the one transmitting a message that is required to be completely secure. Nonetheless, we will now show that this restriction limits the capacity region and is thus a sub-region of the capacity region of the original problem.

**Theorem 2.** *By restricting the encoder of the interfered-with user to a deterministic encoder we strictly reduce the capacity region.*

*Proof:* In order to show the above we show that the capacity region given the restriction to a deterministic encoder for the interfered-with transmitter does not contain a point

that can be achieved without this restriction. The point that we consider is the one obtained by $\mathsf{R}_2 = 0$, where $\boldsymbol{X}_2$ is simply Gaussian noise (random noise created ad-hoc in the transmitter, which is an "empty" stochastic encoder). Given that this is our choice of $\boldsymbol{X}_2$ it is evident that

$$\mathsf{R}_1 = \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_1) - \frac{1}{2}\mathsf{log}\left(1+\frac{a\mathsf{snr}_1}{1+\mathsf{snr}_2}\right) \tag{7}$$

is achievable using a standard optimal code sequence for the Gaussian wiretap channel.

We now need to show that this is not attainable when we limit the encoder of the interfered-with transmission to a deterministic encoder. This observation also provides us with the exact bounding box of the capacity region of this limited case. Note that by limiting to a deterministic encoder and due to the fact that we have reliable communication of $W_2$ to $\boldsymbol{Y}_2$, we have the following equality in the limit, due to the requirement of complete secrecy:

$$\begin{aligned}\mathsf{R}_{1,\max} &= \lim_{n\to\infty}\left[I\left(\boldsymbol{X}_1;\boldsymbol{Y}_1\right) - I\left(\boldsymbol{X}_1;\boldsymbol{Y}_2\right)\right] \\ &= \lim_{n\to\infty}\left[I\left(\boldsymbol{X}_1;\boldsymbol{Y}_1\right) - I\left(\boldsymbol{X}_1;\boldsymbol{Y}_2|\boldsymbol{X}_2\right)\right. \\ &\quad \left. + I\left(\boldsymbol{X}_2;\boldsymbol{Y}_2|\boldsymbol{X}_1\right) - I\left(\boldsymbol{X}_2;\boldsymbol{Y}_2\right)\right] \\ &= \lim_{n\to\infty}\left[I\left(\boldsymbol{X}_1;\boldsymbol{Y}_1\right) - I\left(\boldsymbol{X}_1;\boldsymbol{Y}_2|\boldsymbol{X}_2\right)\right] \\ &= \lim_{n\to\infty}\left[I\left(\boldsymbol{X}_1;\sqrt{\mathsf{snr}_1}\boldsymbol{X}_1+\boldsymbol{N}_1\right)\right. \\ &\quad \left. - I\left(\boldsymbol{X}_1;\sqrt{a\mathsf{snr}_1}\boldsymbol{X}_1+\boldsymbol{N}_2\right)\right]\end{aligned} \tag{8}$$

where the second equality is due to reliable communication of $W_2$ and the deterministic encoder restriction $W_2 \to \boldsymbol{X}_2$. The last transition is due to the independence of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (no cooperation). Maximizing the above expression over $P_{\boldsymbol{X}_1}$ (it no longer depends on $P_{\boldsymbol{X}_2}$) gives us the maximum rate

$$\mathsf{R}_{1,\max} = \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_1) - \frac{1}{2}\mathsf{log}(1+a\mathsf{snr}_1) \tag{9}$$

which is strictly less than (7). Thus, for this setting we know the bounding box is defined by the above and $\mathsf{R}_2 = \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_2)$. Moreover, according to the results in [2] $\mathsf{R}_2 = \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_2)$ is obtained when $\mathsf{R}_1 = 0$, since reliable decoding of $\boldsymbol{X}_1$ is required for the maximum $\mathsf{R}_2$, and thus complete secrecy cannot be attained. In other words, the pair $(0, \frac{1}{2}\mathsf{log}(1+\mathsf{snr}_2))$ is a corner point of the capacity region in this setting. This concludes the proof. ∎

Given the above equivalence we have a *degraded* Gaussian interference channel. For this channel we wish to provide a Sato-type outer bound, meaning we wish to follow the approach in [6]. The approach in [6] was to observe that the *degraded* Gaussian interference channel is upper bounded by the Gaussian broadcast channel (BC) capacity since in the Gaussian BC the two transmitters cooperate and only have a general power constraint (may split the power between themselves as they wish). Thus, the specific power split of the *degraded* Gaussian interference channel is a special case of the Gaussian BC.

Our approach is to follow the same logic and employ the known results of the Gaussian BC with confidential messages (BCC), for which we have the capacity region [9]. This leads to the following result:

**Theorem 3.** *The capacity region of the Gaussian Z-interference channel with a secrecy constraint on the interfering message and a deterministic encoder at the interfered-with transmitter is contained in the following region:*

$$\begin{aligned}(\mathsf{R}_1,\mathsf{R}_2) = \Bigg(&\frac{1}{2}\mathsf{log}\left(\frac{1+\beta(\mathsf{snr}_1+\mathsf{snr}_2/a)}{1+\beta a(\mathsf{snr}_1+\mathsf{snr}_2/a)}\right), \\ &\frac{1}{2}\mathsf{log}\left(\frac{1+a(\mathsf{snr}_1+\mathsf{snr}_2/a)}{1+\beta a(\mathsf{snr}_1+\mathsf{snr}_2/a)}\right)\Bigg) \\ \mathsf{R}_1 &\le \frac{1}{2}\mathsf{log}\left(1+\mathsf{snr}_1\right) - \frac{1}{2}\mathsf{log}(1+a\mathsf{snr}_1) \\ \mathsf{R}_2 &\le \frac{1}{2}\mathsf{log}\left(1+\mathsf{snr}_2\right)\end{aligned} \tag{10}$$

*for some* $\beta \in [0,1]$.

As discussed in the proof of Theorem 2, the point $(0, \frac{1}{2}\mathsf{log}\left(1+\mathsf{snr}_2\right))$ is a corner point of the capacity region. However this is not a point on the above outer bound. Assume that

$$\mathsf{R}_2 = \frac{1}{2}\mathsf{log}\left(1+\mathsf{snr}_2\right) \tag{11}$$

then

$$\beta = \frac{a\mathsf{snr}_1}{(a\mathsf{snr}_1+\mathsf{snr}_2)(1+\mathsf{snr}_2)}. \tag{12}$$

Substituting the above in the bound on $\mathsf{R}_1$ we obtain

$$\mathsf{R}_1 = \frac{1}{2}\mathsf{log}\left(1+\frac{\mathsf{snr}_1}{1+\mathsf{snr}_2}\right) - \frac{1}{2}\mathsf{log}\left(1+\frac{a\mathsf{snr}_1}{1+\mathsf{snr}_2}\right). \tag{13}$$

The other corner point of the above outer bound is when

$$\mathsf{R}_1 = \frac{1}{2}\mathsf{log}\left(1+\mathsf{snr}_1\right) - \frac{1}{2}\mathsf{log}(1+a\mathsf{snr}_1) \tag{14}$$

for which

$$\beta(a\mathsf{snr}_1+\mathsf{snr}_2) = a\mathsf{snr}_1. \tag{15}$$

Substituting the above in the bound on $\mathsf{R}_2$ we obtain

$$\mathsf{R}_2 = \frac{1}{2}\mathsf{log}\left(1+\frac{\mathsf{snr}_2}{1+a\mathsf{snr}_1}\right). \tag{16}$$

This, of course, is an attainable point by using for $W_1$ a Gaussian wiretap code sequence, and for $W_2$ a Gaussian point-to-point code sequence. At $\boldsymbol{Y}_2$ we first consider $\boldsymbol{X}_1$ as additive Gaussian noise and decode $\boldsymbol{X}_2$ ($W_2$). After removing it, we still have for $W_1$ a Gaussian wiretap code sequence designed for complete secrecy at $\boldsymbol{Y}_2$.

In order to get a better feeling for the above outer bound, we compare it with three possible inner bounds. The first most basic inner bound is obtained by time-sharing between the two schemes that attain the corner points of the capacity region mentioned above. The second bound is the time/frequency division multiplexing (TDM/FDM) bound given in Lemma 2 and the third bound, given in Lemma 3, improves on the

TDM/FDM bound by allowing the interfered-with transmitter to transmit over both subbands.

**Lemma 2.** *The set of non-negative rate pairs* $(\mathsf{R}_1, \mathsf{R}_2)$ *satisfying*

$$\mathsf{R}_1 \le \frac{\lambda}{2}\mathsf{log}\Big(1+\frac{\mathsf{snr}_1}{\lambda}\Big) - \frac{\lambda}{2}\mathsf{log}\Big(1+\frac{a\mathsf{snr}_1}{\lambda}\Big), \quad (17)$$

$$\mathsf{R}_2 \le \frac{\bar{\lambda}}{2}\mathsf{log}\Big(1+\frac{\mathsf{snr}_2}{\bar{\lambda}}\Big), \quad (18)$$

*in which* $0 \le \lambda \le 1$ *and* $\bar{\lambda} = 1-\lambda$, *is achievable for the Gaussian Z-Interference channel with secrecy constraint and a deterministic encoder.*

*Proof:* This region is the TDM/FDM region. To achieve this region we divide the available time/frequency into two orthogonal parts, respectively proportional to $\lambda$ and $\bar{\lambda}$. Then, let user 1 be the only active user for $\lambda$ fraction of time/frequency. As a result, we will have a degraded Gaussian wiretap channel and the achievable rate of secure communication is given by (17). Note that the average SNR of transmitter 1 in the $\lambda$-subband is equal to $\frac{snr_1}{\lambda}$. Similarly, let user 2 transmit only for $\bar{\lambda}$ fraction of time/frequency. Then, (18) gives the achievable rate. ∎

We can improve the TDM/FDM inner bound of Lemma 2 by allowing the interfered-with transmitter to split its power over both subbands.

**Lemma 3.** *The set of non-negative rate pairs* $(\mathsf{R}_1, \mathsf{R}_2)$ *satisfying*

$$\mathsf{R}_1 \le \frac{\lambda}{2}\mathsf{log}\Big(1+\frac{\mathsf{snr}_1}{\lambda}\Big) - \frac{\lambda}{2}\mathsf{log}\Big(1+\frac{a\mathsf{snr}_1}{\lambda}\Big), \quad (19)$$

$$\mathsf{R}_2 \le \frac{\lambda}{2}\mathsf{log}\Big(1+\frac{\mathsf{snr}_{21}}{1+a\frac{\mathsf{snr}_1}{\lambda}}\Big) + \frac{\bar{\lambda}}{2}\mathsf{log}\Big(1+\mathsf{snr}_{22}\Big), \quad (20)$$

*in which* $0 \le \lambda \le 1$, $\bar{\lambda} = 1-\lambda$, *and* $\lambda\mathsf{snr}_{21} + \bar{\lambda}\mathsf{snr}_{22} = \mathsf{snr}_2$ *is achievable for the Gaussian Z-Interference channel with secrecy constraint and a deterministic encoder.*

*Proof:* Similar to the TDM/FDM inner bound, we divide the available time/frequency into two orthogonal parts proportional to $\lambda$ and $\bar{\lambda}$. The main difference here is to split $\mathsf{snr}_2$ into $\mathsf{snr}_{21}$ and $\mathsf{snr}_{22}$ such that $\lambda\mathsf{snr}_{21} + \bar{\lambda}\mathsf{snr}_{22} = \mathsf{snr}_2$ and let user 2 consume them in the $\lambda$ and $\bar{\lambda}$ fraction of time/frequency, respectively. However, user 1 transmits only in the $\lambda$-subband. Therefore, in the $\lambda$-subband both users are active. Clearly, (19) is still achievable for user 1 since receiver 1 is free of interference. The achievable rate of user 2 has two terms, each corresponding to one of the subbands. In the $\lambda$-subband receiver 2 treats interference as noise to achieve $\mathsf{R}_{21} = \frac{1}{2}\mathsf{log}\big(1+\frac{\mathsf{snr}_{21}}{1+a\frac{\mathsf{snr}_1}{\lambda}}\big)$. In the $\bar{\lambda}$-subband user 2 is the only active user and thus the interference-free rate $\mathsf{R}_{22} = \frac{1}{2}\mathsf{log}\big(1+\mathsf{snr}_{22}\big)$ is achievable. Therefore, $\mathsf{R}_2 = \lambda\mathsf{R}_{21} + \bar{\lambda}\mathsf{R}_{22}$ is obtained for user 2 in (20). ∎

Figure 2 depicts both the Sato-type outer bound (dashed) and the three possible inner-bounds.

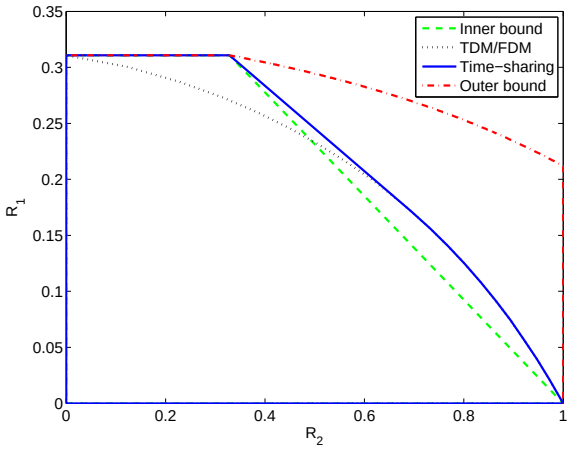



Fig. 2. The Sato-type outer bound (dash-dot line), the basic time-sharing inner bound (dashed), the TDM/FDM inner bound (dotted) and the improved TDM/FDM inner bound (solid).

## APPENDIX

*Sketch Proof of Theorem 1:* We follow the approach proposed by Abbe and Zheng in [5] which examines the optimality of the Gaussian inputs by analysis of the information theoretic equation in the vicinity of the Gaussian input distributions using permutations depicted by Hermite polynomials.

The single-letter expression which we are considering is the following:

$$I(X_1;Y_1) - I(X_1;Y_2) = I(X_1;\sqrt{\mathsf{snr}_1}X_1 + N_1) - I(X_1;\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2).$$

Similar to [5] we denote the above function using

$$\begin{aligned} &S_{a,\mathsf{snr}_1,\mathsf{snr}_2,p}(X_1,X_2) \quad (21)\\ &= h(\sqrt{\mathsf{snr}_1}X_1 + N_1) - h(N_1)\\ &\quad - h(\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2) + h(\sqrt{\mathsf{snr}_2}X_2 + N_2), \end{aligned}$$

where $p$ is defined below. Now the idea is to examine the following function:

$$F_k(a,\mathsf{snr}_1,\mathsf{snr}_2,p) = \lim_{\delta\to 0}\lim_{\epsilon\to 0}\frac{2}{\epsilon^2}\left[S_{a,\mathsf{snr}_1,\mathsf{snr}_2,p}(X_1,X_2) - S_{a,\mathsf{snr}_1,\mathsf{snr}_2,p}(X_1^G,X_2^G)\right] \quad (22)$$

where $X_1^G \sim g_p$, $X_2^G \sim g_p$, $X_1 \sim g_p(1+\epsilon\widetilde{H}_k)$ and $X_2 \sim g_p(1-\epsilon\widetilde{H}_k)$ with $\widetilde{H}_k$ defined in [5, Lemma 2] as a function of $H_k^{[p]}$ (and the $\delta H_{4k}^{[p]}$ correction term), the normalized Hermite polynomials for the Gaussian distribution having variance $p$, $g_p$. Recall from [5] that $\{H_k^{[p]}\}_{k\ge 0}$ is an orthonormal basis for $L_2(g_p;\Re)$. Moreover, $H_1^{[p]}$ and $H_2^{[p]}$ perturb a Gaussian distribution into another Gaussian distribution with a different first and second moments, respectively. For $k \ge 3$ the permutations move away from the Gaussian distribution. To

simplify notation we denote the Hermite polynomials as $H_k$ whenever the variance $p$ is clear from context.

Given the above we wish to analyze the behavior of the function $F_k(a, \mathsf{snr}_1, \mathsf{snr}_2, p)$ for $k \geq 3$. We have three differential entropies to consider. We begin with $h(\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2)$. As shown in [5, Equation (20)] the density of $\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2$ is given by

$$g_{a\mathsf{snr}_1 p}(1 + \epsilon\left[H_k + \delta H_{4k}\right]) \star g_{\mathsf{snr}_2 p}(1 - \epsilon\left[H_k - \delta H_{4k}\right]) \star g_1 \tag{23}$$

where $\star$ denotes the convolution operator. From [5, Theorem 1] the above is equal to

$$\begin{aligned} & g_{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\left(1 + \epsilon\left\{\left[\left(\frac{a\mathsf{snr}_1 p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^{\frac{k}{2}} H_k + \right.\right.\right. \\ & \left.\delta\left(\frac{a\mathsf{snr}_1 p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^{2k} H_{4k}\right] \\ & - \left[\left(\frac{\mathsf{snr}_2 p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^{\frac{k}{2}} H_k \right. \\ & \left.\left.\left. -\delta\left(\frac{\mathsf{snr}_2 p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^{2k} H_{4k}\right] - \epsilon L\right\}\right) \end{aligned} \tag{24}$$

where

$$L = \frac{g_{a\mathsf{snr}_1 p}\left[H_k + \delta H_{4k}\right] \star g_{\mathsf{snr}_2 p}\left[H_k - \delta H_{4k}\right] \star g_1}{g_{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}}. \tag{25}$$

Using [5, Lemma 3] one can show that $L$ is a linear combination of several Hermite polynomials $H_\ell$ of power $a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1$ with $\ell \geq 2k$. Thus, the density of $\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2$ can be written as a Gaussian $g_{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}$ perturbed by the direction $H_k$ on the order of $\epsilon$ and several $H_\ell$'s with $\ell \geq 2k$ on the order of $\epsilon^2$. Using [5, Lemma 2] and denoting $Y_2 = \sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2$ and $Y_2^G = \sqrt{a\mathsf{snr}_1}X_1^G + \sqrt{\mathsf{snr}_2}X_2^G + N_2$, we have

$$\begin{aligned} & h(\sqrt{a\mathsf{snr}_1}X_1 + \sqrt{\mathsf{snr}_2}X_2 + N_2) = \\ & h(\sqrt{a\mathsf{snr}_1}X_1^G + \sqrt{\mathsf{snr}_2}X_2^G + N_2) - D(Y_2||Y_2^G) \end{aligned} \tag{26}$$

and using [5, Lemma 1] we have

$$\begin{aligned} D(Y_2||Y_2^G) &= \frac{\epsilon^2}{2}o(\delta) \\ &+ \frac{\epsilon^2}{2}\left((a\mathsf{snr}_1)^{\frac{k}{2}} - (\mathsf{snr}_2)^{\frac{k}{2}}\right)^2\left(\frac{p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^k \end{aligned} \tag{27}$$

Following similar steps we have that $h(\sqrt{\mathsf{snr}_1}X_1 + N_1)$ is

$$h(\sqrt{\mathsf{snr}_1}X_1^G + N_1) - \frac{\epsilon^2}{2}\left(\frac{\mathsf{snr}_1 p}{\mathsf{snr}_1 p + 1}\right)^k + \frac{\epsilon^2}{2}o(\delta) \tag{28}$$

and $h(\sqrt{\mathsf{snr}_2}X_2 + N_2)$ is

$$h(\sqrt{\mathsf{snr}_2}X_2^G + N_2) - \frac{\epsilon^2}{2}\left(\frac{\mathsf{snr}_2 p}{\mathsf{snr}_2 p + 1}\right)^k + \frac{\epsilon^2}{2}o(\delta). \tag{29}$$

Putting everything together we have that

$$S_{a,p}(X_1, X_2) - S_{a,p}(X_1^G, X_2^G) \tag{30}$$

$$\begin{aligned} &= \frac{\epsilon^2}{2}\left[\left(\frac{\mathsf{snr}_1 p}{\mathsf{snr}_1 p + 1}\right)^k + \left(\frac{\mathsf{snr}_2 p}{\mathsf{snr}_2 p + 1}\right)^k \right. \\ &\left. - \left((a\mathsf{snr}_1)^{\frac{k}{2}} - (\mathsf{snr}_2)^{\frac{k}{2}}\right)^2\left(\frac{p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^k\right] + \frac{\epsilon^2}{2}o(\delta). \end{aligned}$$

As noted above, the condition for a non-Gaussian distribution to improve on the Gaussian one is that there exists some $k \geq 3$ for which

$$\begin{aligned} & \left(\frac{\mathsf{snr}_1 p}{\mathsf{snr}_1 p + 1}\right)^k + \left(\frac{\mathsf{snr}_2 p}{\mathsf{snr}_2 p + 1}\right)^k \\ & - \left((a\mathsf{snr}_1)^{\frac{k}{2}} - (\mathsf{snr}_2)^{\frac{k}{2}}\right)^2\left(\frac{p}{a\mathsf{snr}_1 p + \mathsf{snr}_2 p + 1}\right)^k > 0. \end{aligned}$$

Moreover, since we have $\mathsf{snr}_1$ and $\mathsf{snr}_2$ we can take $p = 1$ in the above. Examining this expression (with $p = 1$) we observe that we can lower bound it with

$$\left(\frac{\mathsf{snr}_1}{\mathsf{snr}_1 + 1}\right)^k + \left(\frac{\mathsf{snr}_2}{\mathsf{snr}_2 + 1}\right)^k - \frac{a^k\mathsf{snr}_1^k + \mathsf{snr}_2^k}{(a\mathsf{snr}_1 + \mathsf{snr}_2 + 1)^k}.$$

Noticing that

$$\left(\frac{\mathsf{snr}_2}{\mathsf{snr}_2 + 1}\right)^k - \frac{\mathsf{snr}_2^k}{(a\mathsf{snr}_1 + \mathsf{snr}_2 + 1)^k} \geq 0 \tag{31}$$

and for any $a \in [0, 1]$ also

$$\left(\frac{\mathsf{snr}_1}{\mathsf{snr}_1 + 1}\right)^k - \frac{a^k\mathsf{snr}_1^k}{(a\mathsf{snr}_1 + \mathsf{snr}_2 + 1)^k} \geq 0 \tag{32}$$

as it is a monotonically decreasing function in $a \in [0, 1]$ and for $a = 1$ it is non-negative. Thus, we can conclude that for any set of parameters $\mathsf{snr}_1 > 0, \mathsf{snr}_2 > 0$ and $a \in [0, 1)$ a non-Gaussian distribution would outperform the Gaussian one in the maximization problem given in (4). This concludes the proof. ■

## REFERENCES


[1] I. Sason, "On the corner points of the capacity region of a two-user Gaussian interference channel," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 3682–3697, July 2015.

[2] R. Bustin, H. V. Poor, and S. Shamai (Shitz), "The effect of maximal rate codes on the interfering message rate," submitted to the *IEEE Transactions on Information Theory, April 2015*, 2015, arXiv:1404.6690.

[3] E. A. Abbe and L. Zheng, "A coordinate system for Gaussian networks," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 721–733, February 2012.

[4] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 3072–3081, November 2001.

[5] E. A. Abbe and L. Zheng, "Coding along Hermite polynomials for interference channels," in *Proc. IEEE Information Theory Workshop, (ITW 2009)*, pp. 584–588, Taormina, Sicillia, Italy, 11-16 October 2009.

[6] H. Sato, "On degraded Gaussian two-user channels," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 637–640, September 1977.

[7] A. D. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, October 1975.

[8] M. H. M. Costa, "On the Gaussian interference channel," *IEEE Transactions on Information Theory*, vol. 31, no. 5, pp. 607–615, September 1985.

[9] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Secure communication over fading channels," *IEEE Transactions on Information Theory, Special Issue on Information Theoretic Security*, vol. 54, no. 6, pp. 2470–2492, June 2008.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Two Applications of the Gaussian Poincaré Inequality in the Shannon Theory

Silas L. Fong
Department of ECE,
National University of Singapore,
Email: silas_fong@nus.edu.sg

Vincent Y. F. Tan
Department of ECE/Mathematics,
National University of Singapore
Email: vtan@nus.edu.sg

***Abstract*—We employ the Gaussian Poincaré inequality for two tasks in the Shannon theory. First, we show that the Gaussian broadcast channel admits a strong converse. Second, we demonstrate that the empirical output distribution of a delay-limited code for the AWGN channel with quasi-static fading and with non-vanishing probability of error converges to the maximum mutual information output distribution (in the normalized relative entropy sense).**

## I. Introduction

The Poincaré inequality for Gaussian measures [1] is one of the most prominent results in the theory of concentration of measure. Roughly speaking, it states that if $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function and $\phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ is the standard Gaussian density, then the variance of $f$ can be bounded in terms of the expectation of the squared derivative of $f$, i.e.,

$$\mathsf{Var}_\phi[f] \leq \mathbb{E}_\phi[\|\nabla f\|^2]. \tag{1}$$

In the present work, we employ a modification of the Gaussian Poincaré inequality for two tasks in Shannon theory. These are described briefly in the following sections.

## II. Gaussian Broadcast Channels

The Gaussian broadcast channel [2, Ch. 5] is a basic model for the downlink of a communication system. Two messages $W_1 \in [2^{nR_1}]$ and $W_2 \in [2^{nR_2}]$ are to be encoded into a codeword $X^n = f^{(n)}(W_1, W_2)$. This codeword is power constrained, i.e., $\|X^n\|_2^2 \leq nP$. It is transmitted through two AWGN channels with variances $\sigma_1^2$ and $\sigma_2^2$ respectively, i.e.,

$$Y_1^n = X^n + Z_1^n, \quad \text{and} \quad Y_2^n = X^n + Z_2^n. \tag{2}$$

Decoder $j$, which observes $Y_j^n$, is required to estimate message $W_j$ where $j = 1, 2$. The average probability of error is defined to be $\Pr((\hat{W}_1, \hat{W}_2) \neq (W_1, W_2))$ where $\hat{W}_j$ is decoder $j$'s estimate of $W_j$. The capacity region $\mathcal{C}_{\text{BC}}$ is well known and is given by

$$\mathcal{C}_{\text{BC}} = \bigcup_{\alpha\in[0,1]} \left\{ (R_1, R_2) \in \mathbb{R}_+^2 \,\middle|\, \begin{matrix} R_1 \leq \mathrm{C}\left(\frac{\alpha P}{\sigma_1^2}\right) \\ R_2 \leq \mathrm{C}\left(\frac{(1-\alpha)P}{\alpha P+\sigma_2^2}\right) \end{matrix} \right\}, \tag{3}$$

where $\mathrm{C}(x) := \frac{1}{2}\log(1 + x)$. This region is achieved using superposition coding [3]. Recall that the capacity region is the set of all rate pairs for which the error probability vanishes.

The central question of our investigation in [4] is whether the region in (3) is enlarged if we relax the condition that the error probability vanishes. We allow the error probability to be upper bounded by a non-vanishing constant $\varepsilon \in (0, 1)$. We show that the $\varepsilon$-capacity region is precisely the region in (3). The main technicality in the proof involves bounding a certain variance of the log-likelihood of the messages using (1).

## III. Good Delay-Limited Codes

In [5], we used (1) to investigate quasi-static fading channels [6, Sec. 5.4.1] where the fading coefficient $H$ is random but remains constant during the course of transmission. We are interested in the so-called *delay-limited capacity* [7], which is the maximum achievable rate under the assumption that the maximal error probability over all non-zero fading coefficients vanishes as the blocklength grows.

We adopt a long-term power constraint [8] and the max-over-messages error criterion for delay-limited decoding. It is known (e.g., [7, Sec. III-B]) that the delay-limited capacity is $\mathrm{C}(P_{\text{DL}})$ where $P_{\text{DL}} := \frac{P}{\mathbb{E}[1/H]}$. We show in [5] that for any sequence of codes that is capacity-achieving and whose error probability is upper bounded by some $\varepsilon \in [0, 1)$ is such that sequence of induced output distributions $\{p_{Y^n}\}_{n=1}^\infty$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} D(p_{Y^n} \| p_{Y^*}^n) = 0 \tag{4}$$

where $p_{Y^*}(y) = \mathcal{N}(y; 0, 1 + P_{\text{DL}})$.

## References

[1] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités*, volume 33, pages 120–216, 1999.
[2] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, Cambridge, U.K., 2012.
[3] T. Cover. Broadcast channels. *IEEE Trans. on Inform. Th.*, 18(1):2–14, 1972.
[4] S. L. Fong and V. Y. F. Tan. A proof of the strong converse theorem for Gaussian broadcast channels via the Gaussian Poincaré inequality. `arXiv:1509.01380 [cs.IT]`, Sep 2015.
[5] S. L. Fong and V. Y. F. Tan. Empirical output distribution of good delay-limited codes for quasi-static fading channels. `arXiv:1510.08544 [cs.IT]`, Oct 2015.
[6] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, Cambridge, U.K., 2005.
[7] S. V. Hanly and D. N. C. Tse. Multiaccess fading channels – Part II: Delay-limited capacities. *IEEE Trans. on Inform. Th.*, 44(7):2816–2831, 1998.
[8] G. Caire, G. Taricco, and E. Biglieri. Optimum power control over fading channels. *IEEE Trans. on Inform. Th.*, 45(5):1468–1489, 1999.

# Cooperation Strategies for the Broadcast and Multiple Access Channels

Yossef Steinberg
Technion - Israel Inst. of Technology
Haifa, Israel 3200003
Email: ysteinbe@ee.technion.ac.il

***Abstract*—It has long been observed that cooperation between users in a communication network can improve its performance and simplify the coding schemes. In this talk I will describe recent results on cooperation for the broadcast and multiple access channels, and discuss few of the difficulties that arise when the number of users is large. New results and insights will be presented for specific scenarios.**

## I. Introduction

Due to their potential advantage, coding schemes that employ cooperation between users have attracted considerable attention in recent years. Among the most studied models are the multiple access channel (MAC) with cooperating encoders, and the broadcast channel (BC) with cooperating decoders. For the MAC there are two possible forms of cooperation - conference links and cribbing. In [9] Willems introduced and studied a two users MAC model where the encoders can communicate via conference links of limited capacity before transmission begins, and derived its capacity region. In a subsequent work [10], Willems and Van Der Meulen introduced the MAC with cribbing encoders, and derived its capacity region for all possible cribbing scenarios. In [1], Dabora and Servetto introduced the two users degraded BC with conferencing decoders, and derived its capacity region. The model of Dabora and Servetto can be viewed as a special case of the relay-broadcast channel (RBC) of Liang and Veeravalli [6], where the link from the relay to the destination is replaced by a noiseless bit pipe. The model of Dabora and Servetto was extended in [2], [3] to state-dependent channels. A new coding scheme was introduced by Dikstein *et. al.*, where binning replaces the block-Markov approach that was used by Dabora and Servetto in [1].

The coding schemes developed in [1], [2], [3], [6], [9] and [10], rely on the cooperation links, and cannot be used in their absence. In many modern ad-hoc communication systems, the cooperation resources are not allocated a priori, and their availability depends on many factors of which the system designer does not have control - e.g., weather conditions, presence of users that serve as relays, etc. A typical situation is that a user who wishes to transmit messages to a remote destination, is aware of the possibility that intermediate nodes in the network will act as relays, but their help is not guaranteed. Moreover, in most cases the transmitting user cannot be informed whether the cooperation resources are indeed available during transmission. It is therefore desired to devise coding schemes that take advantage of the cooperation resources when they are available, but can operate also when they are absent, although possibly at reduced rates. This set of problems can be viewed as a channel coding analog of some well known source coding problems, like multiple description [11], [12] and rate distortion when side information may be absent [4].

## II. Recent results

In [7] new models for the broadcast and multiple access channels with uncertain cooperation were presented. Specifically, for the broadcast channel, the model of [1] was extended to the case where the cooperation link is unstable and may be absent. The capacity region for this case was characterised. A model of a MAC with a cribbing link that may be absent was presented, and an achievable region derived based on a combination of super position coding and the block-Markov construction of [10]. Recently, an outer bound for the MAC model of [7] was derived, and the capacity region fully characterised for special cases of interest [5]. In [8], the BC with degraded message sets and conference link was presented, and its capacity region fully characterised. In this talk I will describe recent results on the MAC and BC with unstable cooperation, and discuss a few of the difficulties that arise when extending the basic cooperation models.

## Acknowledgement

This research was supported by the ISAREL SCIENCE FOUNDATION (grant No. 684/11).

## References

[1] R. Dabora and S. Servetto, "Broadcast channels with cooperating decoders," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5438-5454, December 2006.
[2] L. Dikstein, H. Permuter, and Y. Steinberg, "The state-dependent broadcast channel with cooperation," in *Proc. of the 51st Annual Allerton Conference on Communication, Control, and Computing*, Allerton House, Monticello, Illinois, October 2–4, 2013.
[3] L. Dikstein, H. Permuter, and Y. Steinberg, "On state-dependent degraded broadcast channels with cooperation," *IEEE Trans. Inf. Theory*, accepted.
[4] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 6, pp. 727-734, Nov. 1985.
[5] W. Huleihel and Y. Steinberg, "Channels with cooperation links that may be absent," in preparation.
[6] Y. Liang and V. V. Veeravalli, "Cooperative relay broadcast channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 900-1028, March 2007.
[7] Y. Steinberg, "Channels with cooperation links that may be absent," in *Proc. IEEE Int. Symp. Information Theory*, Honolulu, HI, June 29-July 4, 2014.
[8] Y. Steinberg, "Instances of the relay-broadcast channel and cooperation strategies," in *Proc. IEEE Int. Symp. Information Theory*, Hong Kong, China, June 14-19, 2015.
[9] France M. J. Willems, "The discrete memoryless multiple access channel with partially cooperating encoders," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 441-445, May 1983.
[10] F.M.J. Willems and E. C. Van Der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 313-327, May 1985.
[11] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 518-521, Sept. 1980.
[12] J. K. Wolf, A. D. Wyner, J. Ziv, "Source coding for multiple descriptions," *Bell Syst. Tech. J.*, vol. 59, no. 8, pp. 1417-1436, October 1980.

# Mismatched Decoding: DMC and General Channels

Anelia Somekh-Baruch
Bar-Ilan University, Faculty of Engineering
Ramat-Gan 52900, Israel
Email: somekha@biu.ac.il

***Abstract*—The setup of mismatched decoding is considered. By analyzing multi-letter expressions and bounds on the mismatch capacity of a general channel, several results pertaining to the mismatched discrete memoryless channel $W$ with an additive metric $q$ are deduced: it is shown that Csiszár and Narayan's "product-space" improvement of the random coding lower bound on the mismatched capacity, $C_q^{(\infty)}(W)$, is equal to the mismatched threshold capacity with a constant threshold level. It is also proved that $C_q^{(\infty)}(W)$ is the highest rate achievable when the average probability of error converges to zero at a certain specified rate, which is $o(1/n)$ in the case of $q$ which is a bounded rational metric. Finally a lower bound on the average probability of error at rates above the erasures-only capacity of the DMC is derived.**

## I. Introduction

In [1], the mismatch capacity of the DMC with decoding metric $q$, denoted $C_q(W)$, is considered. It is shown that the lower bound derived previously in [2] and [3] is not tight in general. This is established by proving that the random coding bound for the product channel $W^K$, denoted $C_q^{(k)}(W)$, referred to as the "product-space" lower bound, may result in strictly higher achievable rates. The rate $C_q^{(k)}(W)$ is given by $C_q^{(k)}(W) = \max_{P_{X^k}} \min_{P_{\tilde{Y}^k|X^k}} \frac{1}{k} I(X^k;\tilde{Y}^k)$ , where the minimization is over $P_{\tilde{Y}^k} = P_{Y^k}, \mathbb{E}(q(X^k,\tilde{Y}^k)) \geq \mathbb{E}(q(X^k,Y^k))$ and $(X^k,Y^k) \sim P_{X^k} \times W^k$. Consequently, $C_q^{(\infty)}(W) = \limsup_{k\to\infty} C_q^{(k)}(W)$ is an achievable rate as well. In the special case of erasures-only capacity, $C_q^{(\infty)}(W)$ is shown to be a tight bound.

## II. Multiletter Expressions for the Mismatch Capacity and Consequent Results for the DMC

Consider a DMC with a finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, which is governed by the conditional p.m.f. $W$. A rate-$R$ block-code of length $n$ consists of $2^{nR}$ $n$-vectors $x^n(m)$, $m = 1, 2, ..., 2^{nR}$, which correspond to $2^{nR}$ equiprobable messages. An additive mismatched decoder for the channel is defined by function $q_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^n q(x_i, y_i)$ where $q$ is a mapping, referred to as metric, from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. The decoder declares that message $i$ was transmitted iff $q_n(x^n(i), y^n) > q_n(x^n(j), y^n), \forall j \neq i$, and if no such $i$ exists, an error is declared. The mismatch capacity of the DMC is defined as the supremum of achievable rates using the mismatched decoder.

In [4], general multi-letter expressions and bounds for the mismatch capacity of general channel with a general metric were derived. In this work we describe briefly results that were deduced for the mismatched DMC, whose derivations rely on these multi-letter expressions.

The first result refers to a threshold decoder. A $(q_n, \tau_n)$-threshold decoder decides that $i$ is the transmitted message iff $q_n(x^n(i), y^n) \geq \tau_n$ and $q_n(x^n(j), y^n) < \tau_n$, $\forall j \neq i$.

In [5] we prove that for a bounded metric $q$, $C_q^{(\infty)}(W)$ is equal to the constant-threshold capacity, that is, the supremum of achievable rates with a threshold decoder of a constant value $\tau_n = \tau$. An implication of this result is that Csiszár and Narayan's conjecture [1] that $C_q(W) = C_q^{(\infty)}(W)$ is equivalent to the statement that $C_q(W)$ is equal to the constant threshold capacity. In [4, Theorem 6] a multi-letter expression for the constant threshold capacity was derived for a general channel and a general metric sequence. Specifying this expression for the DMC case, we obtain an alternative expression for $C_q^{(\infty)}(W)$ in [5]. In [6] we prove that for a bounded metric $q$, every code-sequence of rate $R$, whose average probability of error with a decoding metric $q$ employed on the output of a DMC vanishes faster than $\eta_n$, where $\eta_n = \min_{x^n,\tilde{x}^n,y^n:\, q_n(\tilde{x}^n,y^n)\neq q_n(x^n,y^n)} |q_n(\tilde{x}^n, y^n) - q_n(x^n, y^n)|$, must satisfy $R \leq C_q^{(\infty)}(W)$. In particular, for rational metrics, it identifies $C_q^{(\infty)}(W)$ as the highest rate achievable with average probability of error which is $o(1/n)$. Since at rates below $C_q^{(\infty)}(W)$ exponential decay of the average probability of error is feasible [1], one can deduce that for a bounded rational metric $q$, $C_q(W) = C_q^{(\infty)}(W)$ iff for all $R < C_q(W)$ the error exponent is positive.

Finally it is shown in [7] that the erasures only capacity of the DMC, $C_{q_{eo}}(W)$, satisfies $\mathcal{E}_{eo}(R,\Theta) \geq 1 - \frac{C_{q_{eo}}(W)}{R}$, where $\mathcal{E}_{eo}(R)$ is the lowest achievable average probability of error at rate $R$ in the erasures-only setup.

## References

[1] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 35–43, Jan. 1995.
[2] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.
[3] J. Hui, "Fundamental issues of multiple accessing," *PhD dissertation, MIT*, 1983.
[4] A. Somekh-Baruch, "A general formula for the mismatch capacity," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4554–4568, Sept. 2015.
[5] ——, "Properties of the Csiszár-narayan product-space achievable rate to the mismatch capacity," Sept. 2015, submitted to the IEEE Trans. Inf. Theory.
[6] ——, "Multi-letter converse bounds for the mismatched discrete memoryless channel with an additive metric," in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 531–535.
[7] ——, "On mismatched list decoding," in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 526–530.

# Classical and Classical-Quantum Sphere Packing Bounds: Rényi vs Kullback and Leibler

Marco Dalai
Department of Information Engineering
University of Brescia
Via Branze 38, 25123, Brescia, Italy
Email: marco.dalai@unibs.it

1 ***Abstract*—We review the use of binary hypothesis testing for the derivation of the sphere packing bound in channel coding, pointing out a key difference between the classical and the classical-quantum setting. In the first case, two ways of using the binary hypothesis testing are known, which lead to the same bound written in different analytical expressions. The first method (historically) compares output distributions induced by the codewords with an auxiliary fixed distribution, and naturally leads to an expression using the Rényi divergence. The second method compares the given channel with an auxiliary one and leads to an expression using the Kullback-Leibler divergence. In the classical-quantum case, due to a fundamental difference in the quantum binary hypothesis testing, these two approaches lead to two different bounds, the first being the "right" one. We discuss the details of this phenomenon, which suggests the question of whether auxiliary channels are used in the optimal way in second approach.**

## I. Classical Hypothesis Testing

We start by recalling that in classical binary hypothesis testing between two distributions $P_0$ and $P_1$ on some set $\mathcal{V}$, based on $n$ independent extractions, the trade-off of the best possible exponents of error probabilities of the first and second kind can be expressed parametrically, for $0 < s < 1$, as

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|0} = -\mu(s) + s\mu'(s) + o(1)$$
$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|1} = -\mu(s) - (1-s)\mu'(s) + o(1)\,,$$

where

$$\mu(s) = \log \sum_{v\in\mathcal{V}} P_0(v)^{1-s} P_1(v)^s$$

is a scaled version of the Rényi divergence usually defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\sum_{v\in V} P^\alpha(v) Q^{1-\alpha}(v)\,,$$

so that $\mu(s) = -sD_{1-s}(P_0\|P_1)$. An explicit computation - or just a different way of deriving the bound - shows that an equivalent expression is

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|0} = D(P_s\|P_0) + o(1)$$
$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|1} = D(P_s\|P_1) + o(1)\,,$$

where $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence

$$D(P,Q) = \sum_v P(v)\log\frac{P(v)}{Q(v)}$$

and $P_s$ is the tilted mixture

$$P_s(v) = \frac{P_0(v)^{1-s}P_1(v)^s}{\sum_{v'} P_0(v)^{1-s}P_1(v)^s}\,.$$

This second representation gives an intuitive interpretation of the bound. Roughly speaking, the probability of error for the optimal test is essentially due to those sequences in $\mathcal{V}^n$ with empirical distribution close to $P_s$, whose total probabilities under $P_0$ and $P_1$ vanish with exponents given by $D(P_s\|P_0)$ and $D(P_s\|P_1)$ respectively. One can notice that the problem of determining the trade-off of the error exponents in the test between $P_0$ and $P_1$ is essentially reduced to the problem of testing $P_s$ against $P_i$, $i = 0, 1$, in the Stein regime where $\mathsf{P}_{\mathrm{e}|s}$ is bounded away from one.

## II. Classical Sphere-Packing

Given a discrete memoryless channel $W : \mathcal{X} \to \mathcal{Y}$ with capacity $C$, the sphere packing bound gives an exponential lower bound on the probability of error of codes at rate $R < C$ in the form

$$\mathsf{P}_{\mathrm{e}} \geq e^{-n(E_{\mathrm{sp}}(R)+o(n))}\,,$$

where $R$ is the coding rate, $n$ the block length and $E_{\mathrm{sp}}(R)$ is the so called sphere packing exponent. Two proofs are known for the classical version of the bound, which naturally lead to two equivalent yet different analytical expressions for the function $E_{\mathrm{sp}}(R)$. A preliminary technical feature common to both procedures is that they both focus on some constant-composition sub-code which has virtually the same rate as the original code, but where all codewords have the same empirical composition $P$. In both cases, then, the key ingredient is binary hypothesis testing (BHT).

### A. The MIT proof

The first proof (see [1], [2]) is based on a binary hypothesis test between the output distributions $\boldsymbol{W}_{\boldsymbol{x}_m}$ induced by the codewords $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ and an auxiliary output product distribution $\boldsymbol{Q} = Q^{\otimes n}$ on $\mathcal{Y}^n$. Let $\boldsymbol{\mathcal{Y}}_m \subseteq \mathcal{Y}^n$ be the decision region for message $m$. Since $\boldsymbol{Q}$ is a distribution, for at least one $m$ we have

$$\boldsymbol{Q}(\boldsymbol{\mathcal{Y}}_m) \leq 1/M \tag{1}$$
$$= e^{-nR}. \tag{2}$$

Considering a binary hypothesis test between $\boldsymbol{W}_{\boldsymbol{x}_m}$ and $\boldsymbol{Q}$, with $\boldsymbol{\mathcal{Y}}_m$ as decision region for $\boldsymbol{W}_{\boldsymbol{x}_m}$, equation (1) gives an exponential upper bound on the probability of error under hypothesis $\boldsymbol{Q}$ which implies a lower bound on the probability of error under hypothesis $\boldsymbol{W}_{\boldsymbol{x}_m}$, which is $\boldsymbol{W}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m})$, the probability of error for message $m$. Here the BHT is considered in the regime where both probabilities decrease exponentially. The standard procedure uses the first form of the bound mentioned in the previous section based on the Rényi divergence. The bound can be extended to the case of testing products of non identical distributions; for the pair of distribution $\boldsymbol{W}_{\boldsymbol{x}_m} = W_{x_{m,1}} \otimes \cdots W_{x_{m,n}}$ and $\boldsymbol{Q} = Q \otimes \cdots \otimes Q$ it gives the performance of an optimal test in the form

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\boldsymbol{W}_{\boldsymbol{x}_m}} = -\mu(s) + s\mu'(s) + o(1) \tag{3}$$

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\boldsymbol{Q}} = -\mu(s) - (1-s)\mu'(s) + o(1) \tag{4}$$

where

$$\mu(s) = \sum_x P(x) \left[\log \sum_{y\in\mathcal{Y}} W_x(y)^{1-s} Q(y)^s\right].$$

At this point the arguments in [1] and [2] diverge a bit; while the former is not rigorous, it has the advantage of giving the tight bound for the arbitrary codeword composition $P$. The latter is instead rigorous but only gives the tight bound for the optimal composition $P$. In [3] we proposed a variation which we believe to be rigorous and at the same time gives the tight bound for an arbitrary composition $P$. The need for this variation will be clear in the discussion of classical-quantum channels in the next section.

For the test based on the decoding region $\boldsymbol{\mathcal{Y}_m}$, the left hand side of (4) is lower bounded by $R$ due to (1). So, if we choose $s$ and $Q$ in such a way that the right hand side of (4) is roughly $R-\epsilon$, then $-(1/n)\log \mathsf{P}_{\mathrm{e}|\boldsymbol{W}_{\boldsymbol{x}_m}}$ must be smaller than the right hand side of (3) computed for those same $s$ and $Q$ (for otherwise the decision region $\boldsymbol{\mathcal{Y}_m}$ would give a test strictly better than the optimal one).

This is obtained by choosing $Q$, as a function of $s$, as the minimizer of $-\mu(s)$ and then selecting $s$ which makes the right hand side of (4) equal to $R-\epsilon$ (whenever possible). Extracting $\mu'(s)$ from (4) in terms of $\mu(s)$ and $R$ and using it in (3), the probability of error for message $m$ is bounded in terms of $R$. After some tedious technicalities, cf. [3, Appendix A], we get

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\boldsymbol{W}_{\boldsymbol{x}_m}} \le \sup_{0<s<1}\left[E_0(s,P) - \frac{s}{1-s}(R-\epsilon)\right] + o(1)\,, \tag{5}$$

where

$$E_0(s,P) = \min_Q \left[\frac{1}{s-1}\sum_x P(x)\log\sum_y W_x(y)^{1-s}Q(y)^s\right] \tag{6}$$

$$= \min_Q \left[\frac{s}{1-s}\sum_x P(x) D_{1-s}(W_x\|Q)\right] \tag{7}$$

$$= \frac{s}{1-s} I_{1-s}(P,W)\,, \tag{8}$$

the minimum being over distributions $Q$ and $I_\alpha(P,W)$ being the $\alpha$-mutual information as defined by Csiszár [4]. We thus find the bound, valid for codes with constant composition $P$

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e,max}} \le \sup_{0<s<1}\frac{s}{1-s}\left[I_{1-s}(P,W) - R + \epsilon\right] + o(1)\,.$$

It is worth pointing out that the chosen $Q$, which achieves the minimum in the definition of $E_0(s,P)$, satisfies the set of constraints (cf [1, eqs. (9.23), (9.24), (9.50)], [5, Cor. 3])

$$Q(y) = \sum_x P(x) V_x(y) \tag{9}$$

if we define $V_x(y)$ as

$$V_x(y) = \frac{W_x^{1-s}(y)Q^s(y)}{\sum_{y'} W_x^{1-s}(y')Q^s(y')}\,. \tag{10}$$

So, the chosen $Q$ is such that its tilted mixtures with the distributions $W_x$ induce $Q$ itself on the output set $\mathcal{Y}$. Using the second representation of the error exponents in binary hypothesis testing mentioned in Section I (extended for independent extractions from non-identical distributions), we observe thus that the chosen $Q$ induces the construction of an auxiliary channel $V$ such that $I(P,V) = \sum_x P(x) D(V_x\|Q) = R-\epsilon$. The second proof of the sphere packing bound, which is summarized in the next section, takes this line of reasoning as starting point.

### *B. Haroutunian's proof*

In the second proof (see [6], [7]) one considers the performance of the given coding scheme for channel $W$ when used for an auxiliary channel $V$ such that $I(P,V) < R$. Due to the strong converse for channel coding, when used with channel $V$ the coding scheme will incur an error probability $1-o(1)$, which means that for at least one codeword $m$ we must have $\boldsymbol{V}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m}) = 1-o(1)$. Applying the data processing inequality for the Kullback-Leibler divergence one thus finds that

$$\boldsymbol{V}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m})\log\frac{\boldsymbol{V}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m})}{\boldsymbol{W}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m})} + \boldsymbol{V}_{\boldsymbol{x}_m}(\boldsymbol{\mathcal{Y}}_m)\log\frac{\boldsymbol{V}_{\boldsymbol{x}_m}(\boldsymbol{\mathcal{Y}}_m)}{\boldsymbol{W}_{\boldsymbol{x}_m}(\boldsymbol{\mathcal{Y}}_m)} \le nD(V\|W|P)\,,$$

from which

$$\log \boldsymbol{W}_{\boldsymbol{x}_m}(\overline{\boldsymbol{\mathcal{Y}}_m}) \ge -\frac{nD(V\|W|P)+1}{1+o(1)}\,.$$

So, the error exponent for channel $W$ is bounded as

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\boldsymbol{W}_{\boldsymbol{x}_m}} \le \min_{V:I(P,V)\le R} D(V\|W|P)(1+o(1)).$$

Note that, thanks to the use of the strong converse, the data processing inequality is enough to get the desired result, but any converse for $V$ would work if followed by the more powerful Stein lemma.

The bound derived is precisely the same as in the previous section, and for the optimal choice of the channel $V$, if we

define the output distribution $Q = PV$ as in (9), then (10) is satisfied for some $s$. So, we notice that the two proofs really rely on a comparison between the original channel and equivalent auxiliary channels/distributions. In the first procedure we start with an auxiliary distribution $Q$, but we find that the optimal choice of $Q$ is such that the tilted mixtures with the $W_x$ distributions are the $V_x$ which give $PV = Q$. In the second procedure we start with the auxiliary channel $V$ but we find that the optimal $V$ induces an output distribution $Q$ whose tilted mixtures with the $W_x$ are the $V_x$ themselves. It is worth noticing that in this second procedure we use a converse for channel $V$; hidden in this step we are using the output distribution $Q$ induced by $V$.

These observations point out that while the MIT proof follows the first formulation of the binary hypothesis testing bound in terms of Rényi divergences, Haroutunian's proof exploits the second formulation based on Kullback-Leiblrer divergences, but the compared quantities are equivalent. There seems to be no reason to prefer the first procedure given the simplicity of the second one.

## III. Quantum Hypothesis Testing

In a binary hypothesis testing between two density operators $\sigma_0$ and $\sigma_1$, based on $n$ independent extractions (but with global measurement), the error exponents of the first and second kind can be expressed parametrically as (see [8])

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\sigma_0} = -\mu(s) + s\mu'(s) + o(1) \tag{11}$$

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\sigma_1} = -\mu(s) - (1-s)\mu'(s) + o(1)\,, \tag{12}$$

where, in complete analogy with the classical case,

$$\mu(s) = \log \operatorname{Tr} \sigma_0^{1-s}\sigma_1^s\,.$$

Upon differentiation, one finds for example for (11)

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\sigma_0} = -\log \operatorname{Tr}(\sigma_0^{1-s}\sigma_1^s) + \operatorname{Tr}\left[\frac{\sigma_0^{1-s}\sigma_1^s}{\operatorname{Tr}\sigma_0^{1-s}\sigma_1^s}\left(\log\sigma_1^s - \log\sigma_0^s\right)\right] + o(1)\,.$$

When $\sigma_0$ and $\sigma_1$ commute, that is, in the classical case, we can define the density operator

$$\sigma_s = \frac{\sigma_0^{1-s}\sigma_1^s}{\operatorname{Tr}\sigma_0^{1-s}\sigma_1^s}$$

and use the property $\log\sigma_1^s - \log\sigma_0^s = \log\sigma_0^{1-s}\sigma_1^s - \log\sigma_0$ to obtain

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\sigma_0} = \operatorname{Tr}\sigma_s(\log\sigma_s - \log\sigma_0) + o(1) = D(\sigma_s\|\sigma_0) + o(1)\,.$$

In a similar way we find

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\sigma_1} = D(\sigma_s||\sigma_1) + o(1)\,.$$

This is indeed the second form of the bound as mentioned already in Section I. However, if $\sigma_0$ and $\sigma_1$ do not commute, the above simplification is not possible. Hence, the two error exponents cannot be expressed in terms of the Kullback-Leibler divergence. So, unlike in the classical binary hypothesis testing, the problem of determining the trade-off of the error exponents in the test between $\sigma_0$ and $\sigma_1$ cannot be reduced to the problem of testing some $\sigma_s$ against $\sigma_i$, $i = 0, 1$, in the Stein regime.

To verify that this is really a property of the quantum binary hypothesis testing and not an artificial effect of the used procedure, it is useful to consider the case of pure states, that is when operators $\sigma_0$ and $\sigma_1$ have rank 1, say $\sigma_0 = |\psi_0\rangle\langle\psi_0|$ and $\sigma_1 = |\psi_1\rangle\langle\psi_1|$, with non-orthogonal $\psi_0$ and $\psi_1$. In this case, $\sigma_0^{1-s} = \sigma_0$ and $\sigma_1^s = \sigma_1$, so that one simply has

$$\mu(s) = \log \operatorname{Tr}\sigma_0\sigma_1 = \log|\langle\psi_0|\psi_1\rangle|^2$$

and consequently the two error exponents both equal[1] $-\log|\langle\psi_0|\psi_1\rangle|^2$. These quantity cannot be expressed as $D(\sigma_s\|\sigma_i)$, $i = 0, 1$, for any $\sigma_s$ because

$$D(\rho\|\sigma_i) = \begin{cases} 0 & \rho = \sigma_i \\ +\infty & \rho \neq \sigma_i \end{cases}, i = 0, 1\,,$$

since $\sigma_0$ and $\sigma_1$ are pure.

## IV. Classical-Quantum Sphere-Packing

The different behavior of binary hypothesis testing in the quantum case with respect to the classical has a direct impact on the sphere packing bound for classical-quantum channels. Both the MIT and Haroutunian's approaches can be extended to this setting, but the resulting bounds are different. In particular, since the binary hypothesis testing is correctly handled with the Rényi divergence formulation, the MIT form of the bound extends to what one expects as the right generalization (in particular, it matches known achievability bounds for pure-state channels), while Haroutunian's form extends to a weaker bound. It was already observed in [9] that the latter gives a trivial bound for all pure state channels, which is a direct consequence of what already shown for the simple binary hypothesis testing in the previous section.

It is useful to investigate this weakness at a deeper level in order to see clearly where the problem really is. Let $W_x$, $x \in \mathcal{X}$, be now density operators, let $\boldsymbol{W}_{\boldsymbol{x}} = W_{x_1} \otimes \cdots \otimes W_{x_n}$ be the state associated to a sequence $\boldsymbol{x}$ and thus $\boldsymbol{W}_{\boldsymbol{x}_1}, \ldots, \boldsymbol{W}_{\boldsymbol{x}_M}$ the states associated to the $M$ messages, where $M = e^{nR}$. Let $\{\boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \ldots, \boldsymbol{\Pi}_M\}$ be the POVM used at the receiver for channel $W$, which means that the probability of decoding $m'$ when $m$ is sent is $\operatorname{Tr}\boldsymbol{\Pi}_{m'}\boldsymbol{W}_{\boldsymbol{x}_m}$. Consider then an auxiliary classical-quantum channel with states $V_x$ and with capacity $C < R$. The strong converse still holds for channel $V$ which implies that for any decoding rule, for at least one message the probability of error is $1 - o(1)$. In particular for the given POVM, for at least one $m$,

$$\operatorname{Tr}(I - \boldsymbol{\Pi}_m)\boldsymbol{V}_{\boldsymbol{x}_m} = 1 - o(1)\,.$$

[1] More precisely, a correct formulation is that at least one of the two error exponents is not larger than $-\log|\langle\psi_0|\psi_1\rangle|^2$.

Using again a data processing inequality for the quantum Kullback-Leibler divergence one then finds as in the classical case that

$$\log \mathrm{Tr}(I-\mathbf{\Pi}_m)\boldsymbol{W}_{\boldsymbol{x}_m} \geq -\frac{nD(V\|W|P)+1}{1+o(1)},$$

and thus

$$-\frac{1}{n}\log \mathsf{P}_{\mathrm{e}|\boldsymbol{W}_{\boldsymbol{x}_m}} \leq \min_{V:I(P,V)\leq R} D(V\|W|P)(1+o(1)).$$

The problem is that if $W$ is a pure state channel, at rates $R < C$ any auxiliary channel $V \neq W$ gives $D(V\|W|P) = \infty$, so that the bound is trivial for all pure state channels. It is important to observe that this is not due to a weakness in the used data processing inequality. In a binary hypothesis test between the the pure state $\boldsymbol{W}_{\boldsymbol{x}_m}$ and a state $\boldsymbol{V}_{\boldsymbol{x}_m}$ built from a different channel $V$, one can notice that the POVM $\{A, I-A\}$ with $A = \boldsymbol{W}_{\boldsymbol{x}_m}$ satisfies

$$\mathrm{Tr}(I-A)\boldsymbol{V}_{\boldsymbol{x}_m} = 1-o(1), \quad \mathrm{Tr}(I-A)\boldsymbol{W}_{\boldsymbol{x}_m} = 0.$$

So, it is really impossible to deduce a positive lower bound for $\mathrm{Tr}(I-\mathbf{\Pi}_m)\boldsymbol{W}_{\boldsymbol{x}_m}$ using only the fact that $\mathrm{Tr}(I-\mathbf{\Pi}_m)\boldsymbol{V}_{\boldsymbol{x}_m} = 1-o(1)$.

It is also worth checking what happens with the MIT procedure. All the steps can be extended to the classical-quantum case (see [3] for details) leading to a bound which has the same form as (5) where $E_0(s,P)$ is defined in analogy with (6) as

$$\begin{aligned} E_0(s,P) &= \min_Q \left[\frac{1}{s-1}\sum_x P(x)\log \mathrm{Tr}\, W_x^{1-s}Q^s\right] \\ &= \min_Q \left[\frac{s}{1-s}\sum_x P(x) D_{1-s}(W_x\|Q)\right], \end{aligned}$$

the minimum being over all density operators $Q$, and $D_{1-s}(\cdot\|\cdot)$ being the quantum Rényi divergence. However, as far as we know, there is no analog of equations (9) and (10), and the optimizing $Q$ does not induce an auxiliary $V$ such that $I(P,V) = R-\epsilon$.

## V. Auxiliary Channels and Strong Converses

We have presented the two main approaches to sphere packing as different procedures which are equivalent in the classical case but not in the classical-quantum case. However, it is actually possible to consider the two approaches as particular instances of one general approach where the channel $W$ is compared to an auxiliary channel $V$, since the auxiliary distribution/state $Q$ can be considered as a channel with constant $V_x = Q$. This principle is very well described in [10], where it is shown that essentially all known converse bounds in classical channel coding can be cast in this framework.

According to this interpretation, the starting point in Haroutunian's proof is general enough to include the MIT approach as a special case. So, the weakness of the method in the classical-quantum case must be hidden in one of the intermediate steps. It is not difficult to notice that the key point is how the strong converse is used in Haroutunian's proof. The general auxiliary channel $V$ is only assumed to have capacity $C < R$, and the strong converse for $V$ which is used is of the simple form $\mathsf{P}_{\mathrm{e}} = 1-o(1)$, which is good enough in the classical case. In the MIT proof, instead, the auxiliary channel is such that $C = 0$, so that the strong converse takes another simple form, $\mathsf{P}_{\mathrm{e}} \geq 1-e^{-nR}$. The critical point is that in the classical-quantum setting a converse of the form $\mathsf{P}_{\mathrm{e}} = 1-o(1)$ for $V$ does not lead to a lower bound on $\mathsf{P}_{\mathrm{e}}$ for $W$ in general. What is needed is a sufficiently fast exponential convergence to 1 of $\mathsf{P}_{\mathrm{e}}$ for channel $V$, which essentially suggests that $V$ should be chosen with capacity not too close to $R$ and that the exact strong converse exponent for $V$ should be used.

The natural question to ask at this point is what the optimal[2] auxiliary channel is when the exact exponent of the strong converse is used. At high rates the question is not really meaningful for all those cases where the known versions of the sphere packing bound coincide with achievability results, that is, for classical channels and for pure state channels. However in the remaining cases, that is, in the low rate region for the mentioned channels or in the whole range of rates $0 < R < C$ for general non commuting mixed-state channels, the question is legitimate. In the classical case, since the choice of an (optimal) auxiliary channel with $C = 0$ or $C = R^-$ leads to the same result, one might expect that any other intermediate choice would give the same result. However, to the best of our knowledge this has never been clarified in the literature.

For classical-quantum channels, the question is perhaps not trivial; it is worth pointing out that even the exact strong converse exponent has been determined only very recently [11]. What is very interesting is that while in the classical case the strong converse exponent for $R > C$ is expressed in terms of Rényi divergence [12], [13], similarly as error exponents for $R < C$, for classical-quantum channels the strong converse exponents are expressed in terms of the so called "sandwiched" Rényi divergence defined by

$$\tilde{D}_\alpha(\rho,\sigma) = \frac{1}{\alpha-1}\log \mathrm{Tr}\left(\sigma^{\frac{1-\alpha}{2\alpha}}\rho\sigma^{\frac{1-\alpha}{2\alpha}}\right)^\alpha.$$

The problem to consider would thus be more or less as follows. Consider an auxiliary channel $V$ with capacity $C < R$ and evaluate its strong converse exponent in terms of sandwiched Rényi divergences. Fix this exponent as the probability of error under hypothesis $\boldsymbol{W}_{\boldsymbol{x}_m}$ in a test between $\boldsymbol{W}_{\boldsymbol{x}_m}$ and $\boldsymbol{V}_{\boldsymbol{x}_m}$, where $\mathbf{\Pi}_m$ is the operator in favor of $\boldsymbol{W}_{\boldsymbol{x}_m}$ and $I-\mathbf{\Pi}_m$ is the one in favor of $\boldsymbol{V}_{\boldsymbol{x}_m}$. Then deduce a lower bound for the probability of error under hypothesis $\boldsymbol{W}_{\boldsymbol{x}_m}$ using the standard binary hypothesis testing bound in terms of Rényi divergences. It is not entirely clear, to this author, that the optimal auxiliary channel should necessarily always be one such that $C = 0$ as used up to now. Since for non commuting mixed-state channels the current known form of sphere packing bound is not yet matched by any achievability result, one cannot exclude that it is not the tightest possible form.

[2]Here we mean optimal memoryless channel for bounding the error exponent in the asymptotic regime.

## REFERENCES


[1] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*. Wiley, New York, 1961.

[2] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding in Discrete Memoryless Channels. I," *Information and Control*, vol. 10, pp. 65–103, 1967.

[3] M. Dalai and A. Winter, "Constant Composition in the Sphere Packing Bound for Classical-Quantum Channels," *arXiv:1509.00715 [cs.IT]*, 2014.

[4] I. Csiszár, "Generalized Cutoff Rates and Rényi's Information Measures," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 26 –34, Jan. 1995.

[5] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 405–417, 1974.

[6] E. A. Haroutunian, "Estimates of the error exponents for the semi-continuous memoryless channel," *(in Russian) Probl. Peredachi Inform.*, vol. 4, no. 4, pp. 37–48, 1968.

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

[8] K. Audenaert, M. Nussbaum, A. Szkoła, and F. Verstraete, "Asymptotic error rates in quantum hypothesis testing," *Communications in Mathematical Physics*, vol. 279, pp. 251–283, 2008, 10.1007/s00220-008-0417-5. [Online]. Available: http://dx.doi.org/10.1007/s00220-008-0417-5

[9] A. Winter, "Coding Theroems of Quantum Information Theory," *Ph.D. dissertation, Uni Bielefeld, arXiv:quant-ph/9907077*.

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

[11] M. Mosonyi and T. Ogawa, "Strong converse exponent for classical-quantum channel coding," *arXiv:1409.3562v5 [quant-ph]*, 2014.

[12] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 357–359, 1973.

[13] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, Sept 2010, pp. 1327–1333.

# Converses from non-signalling codes and their relationship to converses from hypothesis testing

William Matthews
Department of Applied Mathematics and Theoretical Physics, University of Cambridge,
Wilberforce Road, Cambridge, CB3 0WA.
Email: will@northala.net

***Abstract*—Finite blocklength converses for classical and quantum channel coding can be obtained by relaxing the optimization over independent encoding and decoding procedures to procedures which are merely "non-signalling". This approach, inspired by quantum information theory, results in converses which are closely related to the hypothesis testing-based converse of Polyanskiy-Poor-Verdú. Indeed, in the classical case they are equivalent. I will give an overview of the non-signalling codes method and describe its relationship to the hypothesis testing approach.**

## I. Linear transformations of conditional distributions

Consider the following situation. Given a symbol $M$, Alice applies some, possibly randomised, process to produce symbols $X$ and $F$. $F$ is sent to Bob over a noiseless channel. Alice uses $X$ as the input to some discrete channel, from which Bob receives output $Y$. Bob applies some process to $F$ and $Y$ to obtain a symbol $W$. We assume that $M-(F,X)-(F,Y)-W$ is a Markov chain. Letting $N(y|x) = P_{Y|X}(y|x)$, $E(f,x|m) = P_{FX|M}(f,x|m)$, and $D(w|f,y) = P_{W|FY}(w|f,y)$, we have

$$P_{WYXM}(w,y,x,m) = Z(x,w|m,y)N(y|x)P_M(m) \quad (1)$$

where

$$Z(x,w|m,y) := \sum_f D(w|f,y)E(f,x|m). \quad (2)$$

The conditional distribution (2) is what one would have if $N(y|x) = Q_Y(y)$ for some distribution $Q$. It is *non-signalling from Bob to Alice*, which means that

$$\forall x,m,y,y' : \sum_w Z(x,w|m,y) = \sum_w Z(x,w|m,y') =: Z_{X|M}(x|m), \quad (3)$$

in particular

$$\forall x,m,y : \sum_w Z(x,w|m,y) = \sum_f E(f,x|m). \quad (4)$$

Conversely, any bipartite conditional distribution which is non-signalling from Bob to Alice has a (non-unique) decomposition of the form (2) (see [9]). Operationally, this means that it can be implemented by local operations and one-way communication from Alice to Bob. Note that $P_{WYXM}$ depends on $E$ and $D$ only through the distribution $Z$.

The distribution of $W$ given $M$ in the present scenario is

$$P_{W|M}(w|m) = \sum_{x,y} Z(x,w|m,y)N(y|x). \quad (5)$$

Clearly, any linear transformation which takes conditional distributions for $Y$ given $X$ to conditional distributions for $W$ given $M$ can be written in the form (5) if we allow $Z(x,w|m,y)$ to be arbitrary numbers. In fact, the map will have the property that it transforms every conditional distribution to a conditional distribution if and only if $Z$ is a conditional distribution which is non-signalling from Bob to Alice (see [9]).

Naturally, we can write

$$P_{WYXM}(w,y,x,m) = \hat{Z}(m,w|x,y)P_{XY}(x,y) \quad (6)$$

where, for $x$ such that $P_X(x) > 0$ we define

$$\hat{Z}(m,w|x,y) := P_{MW|XY}(m,w|x,y) \quad (7)$$

$$= \frac{P_{WYXM}(w,y,x,m)}{P_{Y|X}(y|x)P_X(x)} \quad (8)$$

$$= Z(x,w|m,y)\frac{P_M(m)}{P_X(x)}. \quad (9)$$

The final equality follows from (1). Note that $P_X(x) = \sum_m Z_{X|M}(x|m)P_M(m)$, so $\hat{Z}$ depends only on $Z$ and $P_M$ (*not* on $N_{Y|X}$). For $x$ such that $P_X(x) = 0$ we let

$$\hat{Z}(m,w|x,y) := P_M(m)P_W(w). \quad (10)$$

It follows that

$$P_{MW}(m,w) = \sum_{m,w} \hat{Z}(m,w|x,y) \sum_x N_{Y|X}(y|x)P_X(x). \quad (11)$$

We will make use of this expression in the next section, and give a quantum generalisation of it in Section V. Note that $\hat{Z}(m,w|x,y)$ is non-signalling from Bob to Alice.

## II. Non-signalling codes

We can regard channel coding as a special case of the scenario described in the previous section. Let $M$ and $W$ take values in the same set of size $k$. We can interpret $M$ as the message and $W$ as the estimate of that message made by the decoder. Let $M$ be uniformly distributed. The average probability of error is $\Pr(M \neq W)$. With the arbitrary noiseless communication from Alice to Bob allowed in the

previous section one can obviously find zero-error codes of arbitrary size for any channel $N_{Y|X}$. A conventional code corresponds to the situation where

$$Z(x,w|m,y) = E(x|m)D(w|y). \tag{12}$$

For these types of code, $Z$ is non-signalling not only from Bob to Alice but also from Alice to Bob, that is

$$\forall w,y,m,m' : \sum_x Z(x,w|m,y) = \sum_x Z(x,w|m',y) =: Z_{W|Y}(w|y). \tag{13}$$

We call any code with this property a *non-signalling code* [7]. The condition (13) implies that $\hat{Z}$ satisfies

$$\sum_x \hat{Z}(m,w|x,y)P_X(x) = P_M(m)Z_{W|Y}(w|y) \tag{14}$$

and, if $\hat{Z}$ satisfies this condition then the corresponding $Z$ is non-signalling from Alice to Bob. The success probability of the code for channel $N_{Y|X}$ is

$$\Pr(M=W) = \sum_{m,x,y} Z(x,m|m,y)N_{Y|X}(y|x)P_M(m) \tag{15}$$

$$= \sum_{m,x,y} \hat{Z}(m,m|x,y)N_{Y|X}(y|x)P_X(x). \tag{16}$$

**Remark 1.** *Fixing $N_{Y|X}$, the success probability (15) is a linear functional of $Z$ and, since the constraints which make $Z$ non-signalling are linear, maximising the success probability over all non-signalling codes is a linear program. Using symmetry, this can be simplified to one whose size is independent of $k$ [9].*

If we use a non-signalling code and take a channel $R_{Y|X}$ where $Y$ and $X$ are independent, i.e. $R_{Y|X}(y|x) = Q_Y(y)$ then, using (14), the distribution of $(M,W)$ is

$$Q_{MW}(m,w) = \sum_{x,y} \hat{Z}(m,w|x,y)P_X(x)Q_Y(y) \tag{17}$$

$$= P_M(m)\sum_y Z_{W|Y}(w|y)Q_Y(y), \tag{18}$$

that is $W$ and $M$ are independent. In this situation, for any choice of $Q_Y$, $\Pr(M=W) = 1/k$, that is

$$\forall Q_Y : \sum_{m,x,y} \hat{Z}(m,m|x,y)P_X(x)Q_Y(y) = 1/k. \tag{19}$$

## III. Hypothesis testing converse

Consider the following hypothesis testing problem. The null hypothesis is that $X$ and $Y$ are distributed according to $P_{XY}$. The alternative hypothesis is a composite hypothesis, which states that $X$ and $Y$ are distributed according to $P_XQ_Y$ for some arbitrary $Q_Y$. A hypothesis test is specified by

$$T[x,y] := \Pr(\text{Accept null}|X=x,Y=y). \tag{20}$$

The minimum type-II error which can be attained by a test with type-I error no more than $\epsilon$ is

$$\beta^*_\epsilon(P_{XY}) := \min_T \max_{Q_Y} \sum_{xy} T[x,y]P_X(x)Q_Y(y) \tag{21}$$

$$\text{subject to} \tag{22}$$

$$\sum_{yx} T[x,y]P_{XY}(x,y) \geq 1-\epsilon. \tag{23}$$

Let us define for distributions $p$ and $q$,

$$\beta_\epsilon(p\|q) := \min_T \left\{ \sum_z T[z]q(z) : \sum_z T[z]p(z) \geq 1-\epsilon \right\}.$$

The set of distributions for $Y$ and the set of tests are both compact, convex sets and the objective function on the RHS of (21) is a bilinear function of the distribution and test. Therefore, by von Neumann's minimax theorem

$$\beta^*_\epsilon(P_{XY}) = \max_{Q_Y} \beta_\epsilon(P_{XY}\|P_XQ_Y). \tag{24}$$

**Proposition 2.** *There is a non-signalling code of size $k$, input distribution $P_X$, and error probability $\epsilon$ for channel $N_{Y|X}$ if and only if there is a test $T$ with*

$$\sum_{xy} T[x,y]N_{Y|X}(y|x)P_X(x) = 1-\epsilon, \text{ and} \tag{25}$$

$$\forall Q_Y : \sum_{xy} T[x,y]Q_Y(y)P_X(x) = 1/k. \tag{26}$$

*Proof.* Suppose that we have a non-signalling code of size $k$ which attains error probability $\epsilon$ for channel $N_{Y|X}$. The distribution of $X$ is fixed by $Z$ and the fact that $M$ is uniformly distributed. For the direct part, let $Z$ be the bipartite conditional distribution for a non-signalling code satisfying the stated properties. If we let

$$T[x,y] = \sum_{m=1}^k \hat{Z}(m,m|x,y), \tag{27}$$

then using (16) we obtain (25) and, using (19) in addition, we obtain (26).

For the converse, let $T$ be a test satisfying (25) and (26), and let

$$\hat{Z}(m,w|x,y) = \frac{1}{k}\delta_{mw}T[x,y] + \frac{1}{k(k-1)}(1-\delta_{mw})(1-T[x,y]). \tag{28}$$

This clearly satisfies (3). Using (26) we have

$$\sum_x \hat{Z}(m,w|x,y)P_X(x) = \frac{1}{k}\delta_{mw}\sum_x T[x,y]P_X(x) \tag{29}$$

$$+\frac{1}{k(k-1)}(1-\sum_x T[x,y]P_X(x))(1-\delta_{mw}) = 1/k^2 \tag{30}$$

so $\hat{Z}$ also satisfies (14). It follows that $Z$ satisfies (3) and (13), so it is a non-signalling code. Furthermore, by (25),

$$\Pr(M=W) = \sum_{m,x,y} \hat{Z}(m,m|x,y)N_{Y|X}(y|x)P_X(x) = 1-\epsilon. \tag{31}$$

□

A constraint on tests of the form (26) is a rather unusual in the context of hypothesis testing. In [10], tests with this property (or more generally, property (35)) are called "$P_X$-balanced"), and as noted there, we may relax this condition without changing the minimax type-II error probability: Suppose we have a test $T'$ which satisfies

$$\sum_{xy} T'[x,y]N_{Y|X}(y|x)P_X(x) \geq 1-\epsilon, \text{ and} \tag{32}$$

$$\forall Q_Y : \sum_{xy} T'[x,y]Q_Y(y)P_X(x) \leq \beta. \tag{33}$$

The later condition is equivalent to

$$\forall y : \sum_{x} T'[x,y]P_X(x) =: c_y \leq \beta. \tag{34}$$

If we let

$$T[x,y] = (1-\lambda_y)T'[x,y] + \lambda_y,$$

where $\lambda_y = \frac{\beta - c_y}{1-c_y}$, then

$$\forall y : \sum_{x} T[x,y]P_X(x) = \beta, \tag{35}$$

and since $T'[x,y] \leq T[x,y] \leq 1$ for all $x, y$

$$\sum_{xy} T[x,y]N_{Y|X}(y|x)P_X(x) \geq 1-\epsilon. \tag{36}$$

It follows that there is a non-signalling code of size $k$ and input distribution $P_X$ with error probability $\epsilon$ for $N_{Y|X}$ if and only if

$$1/k \geq \beta_\epsilon^*(N_{Y|X}(y|x)P_X(x)). \tag{37}$$

**Theorem 3.** *There is a non-signalling code of size $k$ and input distribution $P_X$ and error probability $\epsilon$ for $N_{Y|X}$ if and only if*

$$k \leq \min_{Q_Y} \beta_\epsilon(N_{Y|X}P_X \| Q_Y P_X)^{-1}. \tag{38}$$

*There is a non-signalling code of size $k$ and error probability $\epsilon$ for $N_{Y|X}$ if and only if*

$$k \leq \max_{P_Y} \min_{Q_Y} \beta_\epsilon(N_{Y|X}P_X \| Q_Y P_X)^{-1}. \tag{39}$$

As an upper-bound this is exactly the "minimax" converse given (for conventional codes) in [6] and further studied in [10].

## IV. A little background

For any two systems $\mathsf{Q}$ and $\tilde{\mathsf{Q}}$ of equal dimension $d$ we define $|\Phi^+\rangle_{\tilde{\mathsf{Q}}\mathsf{Q}} := \sum_{0\leq j<d} |j\rangle_{\tilde{\mathsf{Q}}} \otimes |j\rangle_{\mathsf{Q}}$ and $\Phi^+_{\tilde{\mathsf{Q}}\mathsf{Q}} = |\Phi^+\rangle\langle\Phi^+|_{\tilde{\mathsf{Q}}\mathsf{Q}}$. The vector $|\Phi^+\rangle_{\tilde{\mathsf{Q}}\mathsf{Q}}$ has the property that for any operator $L_{\tilde{\mathsf{Q}}}$

$$L_{\tilde{\mathsf{Q}}}|\Phi^+\rangle_{\tilde{\mathsf{Q}}\mathsf{Q}} = L_{\mathsf{Q}}^{\mathrm{T}}|\Phi^+\rangle_{\tilde{\mathsf{Q}}\mathsf{Q}}. \tag{40}$$

where $L_{\mathsf{Q}}^{\mathrm{T}}$ is the transpose of $L_{\mathsf{Q}} := \mathbf{id}^{\mathsf{Q}\leftarrow\tilde{\mathsf{Q}}} L_{\tilde{\mathsf{Q}}}$ in the computational basis ($\mathbf{id}^{\mathsf{Q}\leftarrow\tilde{\mathsf{Q}}}$ is the linear map which takes the computational basis for operators on $\tilde{\mathsf{Q}}$ to that for $\mathsf{Q}$, i.e. $\mathbf{id}^{\mathsf{Q}\leftarrow\tilde{\mathsf{Q}}} : |i\rangle\langle j|_{\tilde{\mathsf{Q}}} \mapsto |i\rangle\langle j|_{\mathsf{Q}}$). This fact is sometimes referred to as the 'transpose trick'. We also note that $\mathrm{Tr}_{\tilde{\mathsf{Q}}}\Phi^+_{\tilde{\mathsf{Q}}\mathsf{Q}} = \mathbb{1}_{\mathsf{Q}}$ and $\mathrm{Tr}_{\mathsf{Q}}\Phi^+_{\tilde{\mathsf{Q}}\mathsf{Q}} = \mathbb{1}_{\tilde{\mathsf{Q}}}$. From this property it follows that, for any density operator $\rho_{\mathsf{A}}$, $\rho_{\mathsf{A}}^{1/2}\Phi_{\mathsf{A}\tilde{\mathsf{A}}}\rho_{\mathsf{A}}^{1/2}$ is a purification of $\rho_{\mathsf{A}}$. Let $\mathcal{H}_{\mathsf{A}}$ and $\mathcal{H}_{\mathsf{B}}$ be Hilbert spaces of finite dimension. Any linear map $\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}}$ from operators on $\mathcal{H}_{\mathsf{A}}$ to operators on $\mathcal{H}_{\mathsf{B}}$, has an *operator representation* $\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}}\Phi^+_{\tilde{\mathsf{A}}\mathsf{A}}$. We note that

$$\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}} : \kappa_{\mathsf{A}} \mapsto \mathrm{Tr}_{\tilde{\mathsf{A}}}\kappa_{\tilde{\mathsf{A}}}^{\mathrm{T}}\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}}\Phi^+_{\tilde{\mathsf{A}}\mathsf{A}}. \tag{41}$$

(This correspondence between linear maps between operators and operators is known as the 'Choi-Jamiołkowski isomorphism'.) Complete positivity of a map corresponds to its operator representation being positive semidefinite. $\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}}$ is trace preserving if and only if $\mathrm{Tr}_{\mathsf{B}}\mathcal{L}^{\mathsf{B}\leftarrow\mathsf{A}}\Phi^+_{\tilde{\mathsf{A}}\mathsf{A}} = \mathbb{1}_{\tilde{\mathsf{A}}}$. A quantum operation from system $\mathsf{A}$ to system $\mathsf{B}$ is a linear map from $\mathcal{H}_{\mathsf{A}}$ to $\mathcal{H}_{\mathsf{B}}$ which is completely positive and trace-preserving.

Given any density operator $\rho_{\mathsf{AB}}$ we can write

$$\rho_{\mathsf{AB}} = \mathcal{W}^{\mathsf{B}\leftarrow\tilde{\mathsf{A}}}\rho_{\mathsf{A}\tilde{\mathsf{A}}} \tag{42}$$

where $\rho_{\mathsf{A}\tilde{\mathsf{A}}} = \rho_{\mathsf{A}}^{1/2}\Phi^+_{\mathsf{A}\tilde{\mathsf{A}}}\rho_{\mathsf{A}}^{1/2}$ and $\mathcal{W}^{\mathsf{B}\leftarrow\tilde{\mathsf{A}}}$ is an operation which we may specify explicitly in terms of its operator representation: Let $\rho_{\mathsf{A}}^{-1/2}$ denote the generalised inverse of $\rho_{\mathsf{A}}^{1/2}$, which is the unique operator such that $\rho_{\mathsf{A}}^{1/2}\rho_{\mathsf{A}}^{-1/2}$ and $\rho_{\mathsf{A}}^{1/2}\rho_{\mathsf{A}}^{-1/2}$ are equal to the orthogonal projection operator, $\rho_{\mathsf{A}}^0$, onto the support of $\rho_{\mathsf{A}}$. Then, for any state $\tau_{\mathsf{B}}$, the operation

$$\mathcal{W}^{\mathsf{B}\leftarrow\tilde{\mathsf{A}}}\Phi_{\mathsf{A}\tilde{\mathsf{A}}} = \rho_{\mathsf{A}}^{-1/2}\rho_{\mathsf{AB}}\rho_{\mathsf{A}}^{-1/2} + (\mathbb{1} - \rho_{\mathsf{A}}^0) \otimes \tau_{\mathsf{B}} \tag{43}$$

satisfies equation (42).

## V. Linear transformations of quantum operations

We will now develop the quantum generalisation of the classical results given earlier, starting with Section I.

Alice has some system $\mathsf{M}$ to which she applies an operation $\mathcal{E}^{\mathsf{FX}\leftarrow\mathsf{M}}$. System $\mathsf{F}$ is transferred noiselessly to Bob, while an operation $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ is applied to $\mathsf{X}$ leaving Bob with system $\mathsf{Y}$. Bob applies an operation $\mathcal{D}^{\mathsf{W}\leftarrow\mathsf{FY}}$ to $\mathsf{FY}$, leaving him with system $\mathsf{W}$. The overall operation from $\mathsf{M}$ to $\mathsf{W}$ is

$$\mathcal{D}^{\mathsf{W}\leftarrow\mathsf{FY}}\mathbf{id}^{\mathsf{F}\leftarrow\mathsf{F}} \otimes \mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\mathcal{E}^{\mathsf{FX}\leftarrow\mathsf{M}}. \tag{44}$$

Fixing $\mathcal{D}^{\mathsf{W}\leftarrow\mathsf{FY}}$ and $\mathcal{E}^{\mathsf{FX}\leftarrow\mathsf{M}}$, (44) is a linear function of $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ which maps any operation $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ to an operation. In fact it satisfies a strictly stronger property, which is that if $\mathcal{N}^{\mathsf{YY}'\leftarrow\mathsf{XX}'}$ is an operation, then it will be mapped to an operation. As shown in [5], any linear map from operations to operations with this property can be written in the form (44).

We define a bipartite operation $\mathcal{Z}^{\mathsf{XW}\leftarrow\mathsf{MY}}$ via

$$\mathcal{Z}^{\mathsf{XW}\leftarrow\mathsf{MY}} := \mathcal{D}^{\mathsf{W}\leftarrow\mathsf{FY}}\mathcal{E}^{\mathsf{FX}\leftarrow\mathsf{M}}. \tag{45}$$

This operation completely determines the map from operations to operations discussed above (see [12]). Evidently this operation is implemented by local operations and one-way quantum communication from Alice to Bob. Any operation of this form is non-signalling from Bob to Alice [1], in the sense that

$$\forall \rho_{\mathsf{Y}} : \ \mathrm{Tr}_{\mathsf{W}}\mathcal{Z}^{\mathsf{XW}\leftarrow\mathsf{MY}}\mathbb{1}_{\mathsf{M}} \otimes \rho_{\mathsf{Y}} = \mathcal{Z}^{\mathsf{X}\leftarrow\mathsf{M}}. \tag{46}$$

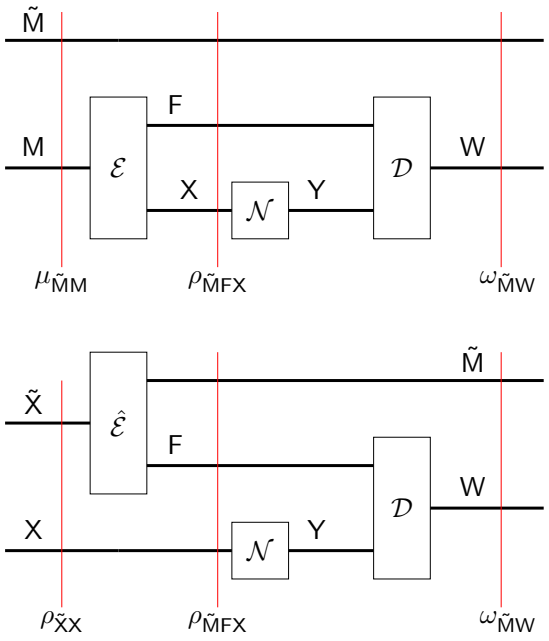



Fig. 1. The systems, operations and states referred to in Section V.

In particular, $\mathcal{Z}^{\mathsf{X}\leftarrow\mathsf{M}} = \mathrm{Tr}_{\mathsf{F}}\mathcal{E}^{\mathsf{F}\mathsf{X}\leftarrow\mathsf{M}}$. Conversely, any bipartite operation which satisfies (46) can be implemented by local operations and quantum communication from Alice to Bob [3]. That is, it can be written in the form (45).

Let $\tilde{\mathsf{M}}$ have the same dimension as system $\mathsf{M}$ and suppose that, initially, Alice has systems $\tilde{\mathsf{M}}\mathsf{M}$ in the state

$$\mu_{\tilde{\mathsf{M}}\mathsf{M}} := \mu_{\mathsf{M}}^{1/2}\Phi^{+}_{\tilde{\mathsf{M}}\mathsf{M}}\mu_{\mathsf{M}}^{1/2} = \mu_{\tilde{\mathsf{M}}}^{1/2}\Phi^{+}_{\tilde{\mathsf{M}}\mathsf{M}}\mu_{\tilde{\mathsf{M}}}^{1/2} \tag{47}$$

where $\mu_{\tilde{\mathsf{M}}} := \mathrm{Tr}_{\mathsf{M}}\mu_{\tilde{\mathsf{M}}\mathsf{M}}$. The 'transpose trick' tells us that $\mu_{\tilde{\mathsf{M}}}^{\mathrm{T}} = \mathbf{id}^{\mathsf{M}\leftarrow\mathsf{M}}\mu_{\mathsf{M}}$. Let

$$\omega_{\tilde{\mathsf{M}}\mathsf{W}} := \mathcal{D}^{\mathsf{W}\leftarrow\mathsf{F}\mathsf{Y}}\mathbf{id}^{\mathsf{F}\leftarrow\mathsf{F}} \otimes \mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\mathcal{E}^{\mathsf{F}\mathsf{X}\leftarrow\mathsf{M}}\mu_{\tilde{\mathsf{M}}\mathsf{M}}. \tag{48}$$

After Alice applies $\mathcal{E}$, the system $\tilde{\mathsf{M}}\mathsf{F}\mathsf{X}$ is in the state $\rho_{\tilde{\mathsf{M}}\mathsf{F}\mathsf{X}} = \mu_{\tilde{\mathsf{M}}}^{1/2}\left(\mathcal{E}^{\mathsf{F}\mathsf{X}\leftarrow\mathsf{M}}\Phi^{+}_{\tilde{\mathsf{M}}\mathsf{M}}\right)\mu_{\tilde{\mathsf{M}}}^{1/2}$. The situation is illustrated in the top half of the figure. Let $\hat{\mathcal{E}}$ be an operation (see previous section) such that $\rho_{\tilde{\mathsf{M}}\mathsf{F}\mathsf{X}} = \hat{\mathcal{E}}^{\tilde{\mathsf{M}}\mathsf{F}\leftarrow\tilde{\mathsf{X}}}\rho_{\tilde{\mathsf{X}}\mathsf{X}}$ where (for the remainder of this article) $\rho_{\tilde{\mathsf{X}}\mathsf{X}}$ is defined to be the state

$$\rho_{\tilde{\mathsf{X}}\mathsf{X}} := \rho_{\mathsf{X}}^{1/2}\Phi^{+}_{\tilde{\mathsf{X}}\mathsf{X}}\rho_{\mathsf{X}}^{1/2}. \tag{49}$$

Note that $\rho_{\tilde{\mathsf{X}}} := \mathrm{Tr}_{\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}} = \mathbf{id}^{\tilde{\mathsf{X}}\leftarrow\mathsf{X}}\rho_{\mathsf{X}}^{\mathrm{T}}$. Then

$$\omega_{\tilde{\mathsf{M}}\mathsf{W}} = \mathcal{D}^{\mathsf{W}\leftarrow\mathsf{F}\mathsf{Y}}\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{M}}\mathsf{F}\mathsf{X}} = \hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}} \tag{50}$$

where

$$\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}} := \mathcal{D}^{\mathsf{W}\leftarrow\mathsf{F}\mathsf{Y}}\hat{\mathcal{E}}^{\tilde{\mathsf{M}}\mathsf{F}\leftarrow\tilde{\mathsf{X}}}. \tag{51}$$

Note the analogy between the expression (50) for the final state of $\tilde{\mathsf{M}}\mathsf{W}$ and the expression (11) for the joint distribution of $M$ and $W$. In terms of the operator representations of $\hat{Z}$ and $Z$, we have

$$\rho_{\mathsf{X}}^{1/2}(\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}\Phi^{+}_{\mathsf{X}\mathsf{Y}\tilde{\mathsf{X}}\tilde{\mathsf{Y}}})\rho_{\mathsf{X}}^{1/2} = \mu_{\tilde{\mathsf{M}}}^{1/2}(\mathcal{Z}^{\mathsf{X}\mathsf{W}\leftarrow\mathsf{M}\mathsf{Y}}\Phi^{+}_{\mathsf{X}\mathsf{Y}\tilde{\mathsf{X}}\tilde{\mathsf{Y}}})\mu_{\tilde{\mathsf{M}}}^{1/2}.$$

## VI. Quantum non-signalling codes

We can view block coding of classical (or quantum) information over a quantum channel as a special case of the scenario described in the previous section. In this case $\mathsf{M}$ and $\mathsf{W}$ are of the same dimension, $k$ (which we call the size of the code). If (as in the classical case) we are concerned with the transmission of a uniformly distributed classical message, then $\mathsf{M}$ stores a uniformly distributed classical message in the computational basis. That is, $\mu_{\mathsf{M}} = \mathbb{1}_{\mathsf{M}}/k$. If $\tilde{\mathsf{M}}$ is measured in the computational basis then we obtain a copy of the message that was sent. The probability of successful transmission is, therefore, the probability of obtaining equal results computational basis measurements are performed on $\tilde{\mathsf{M}}$ and $\mathsf{W}$. The POVM element corresponding to this outcome is

$$\Pi_{\tilde{\mathsf{M}}\mathsf{W}} := \sum_{m} |m\rangle\langle m|_{\tilde{\mathsf{M}}} \otimes |m\rangle\langle m|_{\mathsf{W}},$$

so the success probability of the code is

$$1 - \epsilon = \mathrm{Tr}\Pi_{\tilde{\mathsf{M}}\mathsf{W}}\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}}. \tag{52}$$

In a conventional code, there is no auxiliary forward communication and the bipartite operation is of the form

$$\mathcal{Z}^{\mathsf{X}\mathsf{W}\leftarrow\mathsf{M}\mathsf{Y}} = \mathcal{E}^{\mathsf{X}\leftarrow\mathsf{M}} \otimes \mathcal{D}^{\mathsf{W}\leftarrow\mathsf{Y}} \tag{53}$$

where $\mathcal{E}^{\mathsf{X}\leftarrow\mathsf{M}}$ and $\mathcal{D}^{\mathsf{X}\leftarrow\mathsf{M}}$ are the encoding and decoding operations. The bipartite operation for such codes is not only non-signalling from Bob to Alice, but also from Alice to Bob. We call any forward-assisted quantum code whose bipartite operation is non-signalling in both directions a *quantum non-signalling code* [12]. In terms of the operation $\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}$ this condition is

$$\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}} \otimes \mathbb{1}_{\mathsf{Y}} = \mu_{\tilde{\mathsf{M}}} \otimes \mathcal{Z}^{\mathsf{W}\leftarrow\mathsf{Y}}, \tag{54}$$

and given any operation $\hat{\mathcal{Z}}$ which satisfies this condition the corresponding $\mathcal{Z}$ is non-signalling from Bob to Alice.

**Remark 4.** *In terms of the operator representation of $\mathcal{Z}$, the success probability is a linear functional, the non-signalling and normalising constraints on $\mathcal{Z}$ are affine, while the complete positivity of $\mathcal{Z}$ is equivalent to the operator representation being positive semidefinite. Therefore, maximising the success probability over non-signalling quantum codes is a semidefinite program (see [12]).*

The quantum analog of a channel for which $Y$ and $X$ are independent is for the operation $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ to have the form $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}} = \sigma_{\mathsf{Y}}\mathrm{Tr}_{\mathsf{X}}$. As one would expect, the success probability of a quantum non-signalling code of size $k$ for any such channel is simply $1/k$, that is

$$\forall\sigma_{\mathsf{Y}}: \ \mathrm{Tr}\Pi_{\tilde{\mathsf{M}}\mathsf{W}}\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}} = 1/k. \tag{55}$$

## VII. Quantum hypothesis testing converse

Consider the quantum hypothesis testing problem where the null hypothesis is that the state of $\tilde{\mathsf{X}}\mathsf{Y}$ is $\rho_{\tilde{\mathsf{X}}\mathsf{Y}}$ and the (composite) alternative hypothesis is that that state of $\tilde{\mathsf{X}}\mathsf{Y}$ is of the form $\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}}$ where $\rho_{\tilde{\mathsf{X}}} = \mathrm{Tr}_{\mathsf{Y}}\rho_{\tilde{\mathsf{X}}\mathsf{Y}}$ and $\sigma_{\mathsf{Y}}$ is any state. We can specific a quantum hypothesis test by giving the POVM element $T_{\tilde{\mathsf{X}}\mathsf{Y}}$ corresponding to acceptance of the null hypothesis. Let

$$\beta^*_\epsilon(\rho_{\tilde{\mathsf{X}}\mathsf{Y}}) := \min_{0\leq T_{\tilde{\mathsf{X}}\mathsf{Y}}\leq \mathbb{1}} \max_{\sigma_{\mathsf{Y}}} \mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}} \tag{56}$$

$$\text{subject to } \mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}\mathsf{Y}} \geq 1-\epsilon. \tag{57}$$

For any two states $\rho_0$ and $\rho_1$ of the same system we define

$$\beta_\epsilon(\rho_0\|\rho_1) := \min\{\mathrm{Tr} T\rho_1 : \mathrm{Tr} T\rho_0 \geq 1-\epsilon, 0 \leq T \leq \mathbb{1}\}.$$

By von Neumann's minimax theorem

$$\beta^*_\epsilon(\rho_{\tilde{\mathsf{X}}\mathsf{Y}}) = \max_{\sigma_{\mathsf{Y}}} \beta_\epsilon(\rho_{\tilde{\mathsf{X}}\mathsf{Y}}\|\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}}). \tag{58}$$

We now give the quantum generalisation of Proposition 2.

**Proposition 5.** *There is a quantum non-signalling code of size $k$ with input state $\rho_{\mathsf{X}}$ and error probability $\epsilon$ for operation $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ if and only if there is a quantum hypothesis test $T_{\tilde{\mathsf{X}}\mathsf{Y}}$ satisfying*

$$\mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}} = 1-\epsilon, \text{ and} \tag{59}$$

$$\forall \sigma_{\mathsf{Y}} : \ \mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}} = 1/k \tag{60}$$

*where $\rho_{\tilde{\mathsf{X}}\mathsf{X}} = \rho_{\mathsf{X}}^{1/2}\Phi^+_{\tilde{\mathsf{X}}\mathsf{X}}\rho_{\mathsf{X}}^{1/2}$.*

*Proof.* First the converse part: Suppose that there is a non-signalling code $\mathcal{Z}$ with properties stated in (5). Consider the test obtained by applying the operation $\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}$ to system $\tilde{\mathsf{X}}\mathsf{Y}$, measuring both $\tilde{\mathsf{M}}$ and $\mathsf{W}$ in their computational bases, and accepting (the null hypothesis) when the two results are equal. By (52) and (55) this test has the required properties.

For the direct part, let $T_{\tilde{\mathsf{X}}\mathsf{Y}}$ be a test satisfying (59) and and (60), and let

$$\begin{aligned}\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}} :& A_{\tilde{\mathsf{X}}\mathsf{Y}} \mapsto \frac{1}{k}\Pi_{\tilde{\mathsf{M}}\mathsf{W}}\mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}A_{\tilde{\mathsf{X}}\mathsf{Y}} \\ &+ \frac{1}{k(k-1)}\left(\mathbb{1}-\Pi_{\tilde{\mathsf{M}}\mathsf{W}}\right)\mathrm{Tr}(\mathbb{1}-T_{\tilde{\mathsf{X}}\mathsf{Y}})A_{\tilde{\mathsf{X}}\mathsf{Y}}\end{aligned} \tag{61}$$

where $\Pi_{\tilde{\mathsf{M}}\mathsf{W}} := \sum_m |m\rangle\langle m|_{\tilde{\mathsf{M}}} \otimes |m\rangle\langle m|_{\mathsf{W}}$. It is easy to check that this is non-signalling from Bob to Alice, and the property (60) ensures that this $\hat{\mathcal{Z}}^{\tilde{\mathsf{M}}\mathsf{W}\leftarrow\tilde{\mathsf{X}}\mathsf{Y}}$ satisfies (54). That it has the desired error probability follows from (52), (59) and $\Pi_{\tilde{\mathsf{M}}\mathsf{W}}(\mathbb{1}-\Pi_{\tilde{\mathsf{M}}\mathsf{W}}) = 0$. □

**Corollary 6.** *If there is a non-signalling code of size $k$ and average input state $\rho_{\mathsf{X}}$ and error probability $\epsilon$ for $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ then*

$$k \leq \min_{\sigma_{\mathsf{Y}}} \beta_\epsilon(\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}}\|\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}})^{-1}. \tag{62}$$

*If there is a non-signalling code of size $k$ and error probability $\epsilon$ for $\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}$ then*

$$k \leq \max_{\rho_{\mathsf{X}}} \min_{\sigma_{\mathsf{Y}}} \beta_\epsilon(\mathcal{N}^{\mathsf{Y}\leftarrow\mathsf{X}}\rho_{\tilde{\mathsf{X}}\mathsf{X}}\|\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}})^{-1}. \tag{63}$$

This converse applies to entanglement-assisted codes because they are non-signalling. For memoryless channels, analysing the large block length limit of the upper bound on rate that it gives recovers (see [11]) the known, single-letter formula for the entanglement-assisted classical capacity of a quantum channel [2].

As noted in [11], if we are dealing with codes of the form (53), then the hypothesis test constructed in the direct part of (5) can be implemented by local measurements and classical post-processing of the results (to compare the outcomes). This means that we can obtain a better converse for such codes by restricting the optimisation over hypothesis tests to those which can be implemented in this way. In [11] it was shown that if we restrict to those which can implemented by local operations and one-way classical communication from Alice to Bob then the converse obtained is equivalent to the one obtained in [8].

In Corollary 6 we do not have a quantum analog of Theorem 3 because the implication is only one way. If we could show that one can restrict to quantum tests satisfying $\mathrm{Tr} T_{\tilde{\mathsf{X}}\mathsf{Y}}\rho_{\tilde{\mathsf{X}}} \otimes \sigma_{\mathsf{Y}} = \beta$ for all $\sigma_{\mathsf{Y}}$ without changing the minimax type-II error probability then we could add the other direction of implication to Corollary 6. Whether this is true is open at the time of writing.

## Acknowledgment

My thanks to Debbie Leung and Andreas Winter for many useful discussions on this topic.

## References

[1] D. Beckman, D. Gottesman, M. A. Nielsen, and J. Preskill, "Causal and localizable quantum operations," *Phys. Rev. A*, vol. 64, p. 052309, Oct. 2001.

[2] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, 2637–2655, Oct. 2002.

[3] T. Eggeling, D. Schlingemann, R. F. Werner, "Semicausal operations are semilocalizable," *(EPL) Europhysics Letters*, vol. 57, p. 782, 2002.

[4] M. Hayashi, *Quantum Information: An Introduction*, Springer-Verlag 2006.

[5] G. Chiribella, G. M. D'Ariano, and P. Perinotti, "Transforming quantum opeartions: Quantum supermaps," *(EPL) Europhysics Letters*, vol. 83, no. 3, p. 30004, 2008.

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

[7] T. S. Cubitt, D. Leung, W. Matthews, and A. Winter, "Zero-error channel capacity and simulation assisted by non-local correlations," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5509–5523, Aug. 2011.

[8] L. Wang, R. Renner, "One-Shot Classical-Quantum Capacity and Hypothesis Testing," *Phys. Rev. Lett.* **108**, 200501, May 2012.

[9] W. Matthews, "A linear program for the finite block length converse of Polyanskiy-Poor-Verdú via nonsignalling codes" *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7036–7044, Dec. 2012.

[10] Y. Polyanskiy, "Saddle Point in the Minimax Converse for Channel Coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2576–2595, May. 2013.

[11] W. Matthews, S. Wehner, "Finite blocklength converse bounds for quantum channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7317–7329, Nov. 2014.

[12] W. Matthews, D. Leung, "On the power of PPT-preserving and non-signalling codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 8, pp. 4486–4499, Aug. 2015.

# Combining Detection with Other Tasks of Information Processing

Neri Merhav, Department of EE, Technion, Haifa 3200003, Israel

Classical detection theory, based on the Neyman–Pearson theorem provides the optimal rule for deciding between two hypotheses concerning the distribution or density of a given observation or sequence of observations. It tells us that best trade-off between the two kinds of probability of error is achieved by the likelihood ratio test (LRT). In certain situations, however, this decision between the two hypotheses might be only one of the tasks to be carried out. For example, consider a scenario where under hypothesis $\mathcal{H}_0$, the sequence of observations that we receive is just pure noise (or useless/irrelevant for any other reason), which contains no useful information that may interest us, whereas under hypothesis $\mathcal{H}_1$, the data that we have at hand has emerged from a desirable information source, and in this case, further processing is called for, such as lossless or lossy data compression, parameter estimation channel decoding encryption, further classification, etc.

The straightforward approach to this problem would be to first apply Neyman–Pearson hypothesis testing, and then, if hypothesis $\mathcal{H}_1$ is accepted, perform the corresponding task using the best strategy available. This approach *separates* between optimal decision and the optimality of the subsequent task. A more sophisticated approach, however, is to solve the two problems jointly, namely, to devise a decision rule that takes into account also the cost of the subsequent task (in case it is to be carried out), and on the other hand, optimize the strategy of the following task, taking into account that the data belongs to the decision region of $\mathcal{H}_1$.

In this talk, I will present a unified approach of optimally combining the detection problem with the second information processing task, which is based on a simple extension of the Neyman–Pearson lemma. It minimizes the relevant cost of the second task subject to constraints on the false alarm and misdetection probabilities. We then apply this generalized Neyman–Pearson lemma to three different problems: (i) joint detection and source coding [1], (ii) detection of codeword vs. pure noise, followed by channel decoding in case a signal was detected [2], and (iii) combined channel detection and channel decoding [3].

In all three problems, we derive the optimal solution and assess the asymptotic performance. It turns out that in problems (ii) and (iii) there is an asymptotic *separation principle* in the sense that the same error exponents are achieved by separating the detection from the second task. This is not the case, however, in problem (i).

## References

[1] N. Merhav, "Asymptotically optimal decision rules for joint detection and source coding," *IEEE Trans. Inform. Theory*, vol. 60, no. 11, pp. 6787–6795, November 2014.

[2] N. Weinberger and N. Merhav, "Codeword or noise? Exact random coding exponents for joint detection and decoding," *IEEE Trans. Inform. Theory*, vol. 60, no. 9, pp. 5077–5094, September 2014.

[3] N. Weinberger and N. Merhav, "Channel detection in coded communication," submitted to *IEEE Trans. Inform. Theory*, September 2015. `arXiv:1509.01806`.

**International Zurich Seminar on Communications (IZS), March 2 – 4, 2016**

# Hypothesis Testing and Quasi-Perfect Codes

Gonzalo Vazquez-Vilar
Universidad Carlos III de Madrid
gvazquez@ieee.org

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

Sergio Verdú
Princeton University
verdu@princeton.edu

***Abstract*—Hypothesis testing lower bounds to the channel coding error probability are studied. For a family of symmetric channels, block lengths and coding rates, the error probability of the best code is shown to coincide with that of a binary hypothesis test with certain parameters. The points in which they coincide, are precisely the points at which perfect or quasi-perfect codes exist. General conditions are given for a code to attain minimum error probability.**

## I. Introduction

Consider the channel coding problem of transmitting a set of messages over a binary symmetric channel (BSC). The sphere-packing bound [1, Eq. (5.8.19)] establishes a lower bound on the block error probability of a code with a given rate and blocklength. This bound follows from counting the maximum number of non-overlapping Hamming spheres that can be packed in the output space. In certain cases the sphere-packing bound is achievable. A binary code is said to be *perfect* if non-overlapping Hamming spheres of radius $t$ centered on the codewords exactly fill out the space. Perfect codes are a subset of the class of quasi-perfect codes. A *quasi-perfect* code is defined as a code in which Hamming spheres of radius $t$ centered on the codewords are non-overlapping and Hamming spheres of radius $t+1$ cover the space, possibly with overlaps. Since quasi-perfect codes attain the sphere-packing bound for a BSC, they achieve the minimum error probability among all the codes with the same block length and rate [1, Sec. 5.8]. However, these codes are rare. For each rate $R$, $0 < R < 1$, there exists a block length beyond which neither perfect nor quasi-perfect codes exist [2], [3].

A generalization of the definition of perfect and quasi-perfect codes beyond the Hamming space was proposed by Hamada in [4]. Using a variation of the Fano metric, Hamada derived a lower bound to the channel coding error probability. This bound is achievable by perfect and quasi-perfect codes (defined with respect to the new metric), whenever they exist. This result applies for a class of symmetric discrete memoryless channels.

Binary hypothesis testing has been shown instrumental in the derivation of converse bounds (see e.g. [5], [6]), one prominent recent example being the the meta-converse bound by Polyanskiy *et al.* [7, Th. 27]. Particularized for the BSC, the meta-converse bound recovers the sphere-packing bound [1, Eq. (5.8.19)] (see [7, Sec. III.H] for details). As a result, when perfect or quasi-perfect codes exist, the the meta-converse bound gives the minimum error probability in the BSC.

In this work, we generalize the definitions of perfect and quasi-perfect codes for a class of symmetric channels and we establish a connection between hypothesis testing lower bounds and perfect or quasi-perfect codes. The results of this paper are general enough to recover Hamada's condition for achieving minimum error probability [4, Th. 3].

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the Spanish Ministry of Economy and Competitiveness under grants TEC2012-38800-C03-03, TEC2013-41718-R and FPDI-2013-18602, by the US National Science Foundation under Grant CCF-1016625, and by the Center for Science of Information, an NSF Science and Technology Center under Grant CCF-0939370.

## II. Generalized Quasi-Perfect Codes

Consider the one-shot channel coding problem, where an equiprobable message $v \in \{1, \dots, M\}$ is to be transmitted over a random transformation $P_{Y|X}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ with $\mathcal{X}$ and $\mathcal{Y}$ discrete alphabets. A channel code $\mathcal{C}$ is defined as the set of $M$ codewords $\mathcal{C} = \{x_1, \dots, x_M\}$ assigned to each of the messages. We assume that the maximum likelihood (ML) rule is used to choose the decoded message $\hat{v} \in \{1, \dots, M\}$. The error probability is given by

$$\epsilon(\mathcal{C}) = \Pr[\hat{V} \neq V] \tag{1}$$

$$= 1 - \frac{1}{M} \sum_{y} \max_{x \in \mathcal{C}} P_{Y|X}(y|x). \tag{2}$$

*Definition 1:* A discrete channel is *symmetric* if the rows of the transition matrix of the channel (with inputs as rows and outputs as columns), i. e., $P_{Y|X}(\cdot|x)$, are permutations of each other.

This definition of symmetric channels coincides with that of uniformly dispersive channels of Massey [8, Sec. 4.2] and is less restrictive than those of Cover and Thomas [9] and Gallager [1]. The definition in [9, Sec. 7.2] additionally requires that the columns of the channel transition matrix be permutations of each other, i.e., uniformly focusing according to [8, Sec. 4.2]. The definition in [1, p. 94] requires the channel transition matrix to be partitioned in submatrices such that each submatrix fulfills the condition in [9, Sec. 7.2]. Relations among these definitions are investigated in [10, Sec. VI.B].

We define $\mathcal{S}_x(\theta)$ to be the set of output sequences $y$ with a likelihood given input $x$ of at least $\theta \in [0, 1]$., i. e.,

$$\mathcal{S}_x(\theta) \triangleq \Big\{ y \in \mathcal{Y} \,\Big|\, P_{Y|X}(y|x) \geq \theta \Big\}. \tag{3}$$

We denote the interior and the shell of $\mathcal{S}_x(\theta)$, respectively, as

$$\mathcal{S}_x^{\bullet}(\theta) \triangleq \Big\{ y \in \mathcal{Y} \,\Big|\, P_{Y|X}(y|x) > \theta \Big\}, \tag{4}$$

$$\mathcal{S}_x^{\circ}(\theta) \triangleq \Big\{ y \in \mathcal{Y} \,\Big|\, P_{Y|X}(y|x) = \theta \Big\}. \tag{5}$$

Although we are not assuming that the input and output alphabets are identical and $P_{Y|X}(y|x)$ (or the related Fano metric $\sim \log P_{Y|X}(y|x)$) do not fulfill the properties of a mathematical distance in general, we refer to $\mathcal{S}_x(\theta)$ as a sphere of radius $\theta$ centered on $x$. For specific channels, such as the binary symmetric channel, $\log P_{Y|X}(y|x)$ is an affine function of the Hamming distance between $x$ and $y$ and hence $\mathcal{S}_x(\theta)$ becomes a sphere with respect to that distance.

*Proposition 1:* Let $P_{Y|X}(y|x)$ be a symmetric channel defined over input and output alphabets $\mathcal{X}, \mathcal{Y}$. Then, cardinalities (or "volumes") $|\mathcal{S}_x(\theta)|, |\mathcal{S}_x^{\bullet}(\theta)|, |\mathcal{S}_x^{\circ}(\theta)|$ are independent of $x$.

Then, for any symmetric channel, we define $S(\theta) \triangleq |\mathcal{S}_x(\theta)|$, $S_{\bullet}(\theta) \triangleq |\mathcal{S}_x^{\bullet}(\theta)|$, $S_{\circ}(\theta) \triangleq |\mathcal{S}_x^{\circ}(\theta)|$. Obviously, $S(\theta) = S_{\bullet}(\theta) + S_{\circ}(\theta)$.

*Definition 2:* A code is *perfect* if there exists $\theta \in [0,1]$ such that

$$\bigcup_{x \in \mathcal{C}} \mathcal{S}_x(\theta) = \mathcal{Y}, \tag{6}$$

where the union is disjoint. More generally, a code is *quasi-perfect* if there exists $\theta \in [0,1]$ such that (6) is satisfied and the codeword-centered spheres $\{\mathcal{S}_x^{\bullet}(\theta), x \in \mathcal{C}\}$ are disjoint.

This definition of perfect codes coincides with that in [4, Def. 1] when the channel fulfills the Properties 1-4 in [4]. Definition 2 applies however to any symmetric channel according to 1 (which corresponds to Property 4 in [4]). Also, the definition of quasi-perfect code in Definition 2 includes both perfect and quasi-perfect codes from [4, Def. 1].

## III. The Meta-Converse Bound

Let $\hat{H} \in \{0,1\}$ be the random variable associated to the output of a binary hypothesis test discriminating between distributions $P$ (hypothesis 0) and $Q$ (hypothesis 1). Then, the test can be described by the conditional distribution $P_{\hat{H}|Y}$. Let $\pi_{j|i}$ denote the probability of deciding $j$ when $i$ is the true hypothesis. More precisely, we define

$$\pi_{0|1} \triangleq \sum_y Q(y) P_{\hat{H}|Y}(0|y), \tag{7}$$

$$\pi_{1|0} \triangleq \sum_y P(y) P_{\hat{H}|Y}(1|y). \tag{8}$$

Let $\alpha_\beta\big(P,Q\big)$ denote the minimum error probability $\pi_{1|0}$ among all tests $T \triangleq P_{\hat{H}|Y}$ with $\pi_{0|1}$ at most $\beta$, that is

$$\alpha_\beta\big(P,Q\big) \triangleq \inf_{T:\pi_{0|1} \le \beta} \pi_{1|0}. \tag{9}$$

In [11], Neyman and Pearson derived the explicit form of a (possibly randomized) test $T$ achieving the optimum trade-off (9), given by

$$T_{\text{NP}}(0|y) = \begin{cases} 1, & \text{if } \frac{P(y)}{Q(y)} > \gamma, \\ p, & \text{if } \frac{P(y)}{Q(y)} = \gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\gamma \ge 0$ and $p \in [0,1]$ are parameters chosen such that $\pi_{0|1} = \beta$.

Let $P_X^{\mathcal{C}}$ denote the channel input distribution induced by the codebook $\mathcal{C} = \{x_1, \ldots, x_M\}$, i. e.,

$$P_X^{\mathcal{C}}(x) \triangleq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{x = x_m\}, \tag{11}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

It has been shown in [12, Th. 1] that the exact error probability $\epsilon(\mathcal{C})$ in (2) can be expressed as the best type-0 error probability of an induced binary hypothesis test discriminating between the original distribution $P_X^{\mathcal{C}} \times P_{Y|X}$ and an alternative product distribution $P_X^{\mathcal{C}} \times Q_Y$ with type-1-error equal to $\frac{1}{M}$, i. e.,

$$\epsilon(\mathcal{C}) = \max_{Q_Y} \Big\{ \alpha_{\frac{1}{M}} \Big( P_X^{\mathcal{C}} \times P_{Y|X}, P_X^{\mathcal{C}} \times Q_Y \Big) \Big\}. \tag{12}$$

The right hand side of Eq. (12) is precisely the meta-converse bound [7, Th. 26] after optimization over the auxiliary distribution $Q_Y$. By choosing the auxiliary output distribution $\bar{Q}_Y(y) = |\mathcal{Y}|^{-1}$ and minimizing over all distributions defined over the input alphabet $\mathcal{X}$, identity (12) can be weakened to obtain

$$\epsilon(\mathcal{C}) \ge \inf_{P_X} \Big\{ \alpha_{\frac{1}{M}} \big( P_X \times P_{Y|X}, P_X \times \bar{Q}_Y \big) \Big\}. \tag{13}$$

For the class of symmetric channels considered in Definition 1, we resort to the Neyman-Pearson lemma to find an alternative expression for right-hand side of (13). This expression will be then shown to coincide with the exact error probability $\epsilon(\mathcal{C})$ when $\mathcal{C}$ is a quasi-perfect code according to Definition 2.

## IV. Optimal Code Structure

We particularize the Neyman-Pearson test (10) with $P \leftarrow P_X \times P_{Y|X}$ and $Q \leftarrow P_X \times \bar{Q}_Y$,

$$T_{\text{NP}}(0|x,y) = \begin{cases} 1, & \text{if } y \in \mathcal{S}_x^{\bullet}(\theta), \\ p, & \text{if } y \in \mathcal{S}_x^{\circ}(\theta), \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $\theta = \gamma |\mathcal{Y}|^{-1}$ and $p \in [0,1]$ are parameters that allow to balance $\pi_{1|0}$ and $\pi_{0|1}$. We proceed to analyze the two error types.

Substituting (14) in (7) we obtain

$$\pi_{0|1} = \sum_{x,y} P_X(x) \bar{Q}_Y(y) T_{\text{NP}}(0|x,y) \tag{15}$$

$$= |\mathcal{Y}|^{-1} \sum_x P_X(x) \Big( \big|\mathcal{S}_x^{\bullet}(\theta)\big| + p \big|\mathcal{S}_x^{\circ}(\theta)\big| \Big) \tag{16}$$

$$= |\mathcal{Y}|^{-1} \Big( S_{\bullet}(\theta) + p S_{\circ}(\theta) \Big). \tag{17}$$

Given the constraint on $\pi_{0|1}$ imposed by (13), and the structure of the Neyman-Pearson test, the parameters $p, \theta \in [0,1]$ are chosen such that $\pi_{0|1} = \frac{1}{M}$, i.e.,

$$S_\bullet(\theta) + pS_\circ(\theta) = \frac{|\mathcal{Y}|}{M}. \tag{18}$$

Substituting (14) in (8) we obtain

$$\pi_{1|0} = 1 - \sum_{x,y} P_X(x) P_{Y|X}(y|x) T_{\text{NP}}(0|x,y) \tag{19}$$

$$= 1 - \sum_{x} P_X(x) \Bigg( \sum_{y \in \mathcal{S}_x^\bullet(\theta)} P_{Y|X}(y|x) + p \sum_{y \in \mathcal{S}_x^\circ(\theta)} P_{Y|X}(y|x) \Bigg). \tag{20}$$

For an arbitrary $x$, let $P_{Y|X}(y_i|x)$, $i = 1, \ldots, |\mathcal{Y}|$, denote the output likelihoods indexed in decreasing order. Given the symmetry condition in Definition 1, the vector $\big(P_{Y|X}(y_1|x), \ldots, P_{Y|X}(y_{|\mathcal{Y}|}|x)\big)$ does not depend on the specific value of $x$. Then, for any $x$, we define $\psi_i \triangleq P_{Y|X}(y_i|x)$, $i = 1, \ldots, |\mathcal{Y}|$, and rewrite (20) as

$$\pi_{1|0} = 1 - \left( \sum_{i=1}^{S_\bullet(\theta)} \psi_i + p \sum_{i=1}^{S_\circ(\theta)} \psi_{i+S_\bullet(\theta)} \right). \tag{21}$$

Using (18) and (21), it follows that the lower bound (13) can be rewritten as

$$\epsilon(\mathcal{C}) \geq 1 - \left( \sum_{i=1}^{S_\bullet(\theta)} \psi_i + p \sum_{i=1}^{S_\circ(\theta)} \psi_{i+S_\bullet(\theta)} \right), \tag{22}$$

where $p, \theta \in [0,1]$ are such that $S_\bullet(\theta) + pS_\circ(\theta) = \frac{|\mathcal{Y}|}{M}$.

The next result shows that for a quasi-perfect code $\mathcal{C}$, (22) holds with equality. That is, when they exist, quasi-perfect codes attain the minimum error probability.

*Theorem 1:* Let $P_{Y|X}$ be a symmetric channel according to Definition 1 and let $\mathcal{C}$ be a quasi-perfect code according to Definition 2. Then,

$$\epsilon(\mathcal{C}) = 1 - \left( \sum_{i=1}^{S_\bullet(\theta)} \psi_i + p \sum_{i=1}^{S_\circ(\theta)} \psi_{i+S_\bullet(\theta)} \right), \tag{23}$$

where $p, \theta \in [0,1]$ are such that $S_\bullet(\theta) + pS_\circ(\theta) = \frac{|\mathcal{Y}|}{M}$.

*Proof:* Before showing that (23) holds with equality for arbitrary quasi-perfect codes, we include the (simpler) proof for the particular case of perfect codes.

*a) Perfect codes:* Consider a perfect code $\mathcal{C}$ according to Definition 2. Then, the spheres $\mathcal{S}_x(\theta)$ centered at the codewords are disjoint and their union covers the output space, thus, we have that $MS(\theta) = |\mathcal{Y}|$. These spheres are precisely the ML decision regions for each of the codewords. Then, the error probability (2) can be written as

$$\epsilon(\mathcal{C}) = 1 - \frac{1}{M} \sum_{m=1}^{M} \sum_{y \in \mathcal{S}_{x_m}(\theta)} P_{Y|X}(y|x_m). \tag{24}$$

For symmetric channels, the set $\big\{ P_{Y|X}(y|x_m) \,\big|\, y \in \mathcal{S}_{x_m}(\theta) \big\}$ does not depend on the specific codeword $x_m$. This set coincides with $\{\psi_1, \ldots, \psi_{S(\theta)}\}$, which are, by definition, the $S(\theta)$ largest elements in $\{\psi_1, \ldots, \psi_{|\mathcal{Y}|}\}$. Then, we rewrite (24) as

$$\epsilon(\mathcal{C}) = 1 - \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{S(\theta)} \psi_i \tag{25}$$

$$= 1 - \sum_{i=1}^{S(\theta)} \psi_i. \tag{26}$$

Since $MS(\theta) = |\mathcal{Y}|$, according to (18), we must have $p = 1$, and (26) coincides with the right-hand side of (23).

*b) Quasi-perfect codes:* Consider now a quasi-perfect code $\mathcal{C}$ according to Definition 2. The spheres $\mathcal{S}_x^\bullet(\theta)$ centered at the codewords are disjoint. However, in general, the sets $\mathcal{S}_x^\circ(\theta)$ centered at each of the codewords do overlap. These overlaps correspond to ML decoding ties, and can be resolved arbitrarily without affecting the error probability.

Let $\{\mathcal{P}_m\}$, $m = 1, \ldots, M$, be any partition of the output space such that $\mathcal{P}_m \subseteq \mathcal{S}_{x_m}(\theta)$, $m = 1, \ldots, M$. Let $P_m^\circ \triangleq |\mathcal{P}_m \cap \mathcal{S}_{x_m}^\circ(\theta)|$. Following similar steps as in (25), we obtain

$$\epsilon(\mathcal{C}) = 1 - \frac{1}{M} \sum_{m=1}^{M} \left( \sum_{i=1}^{S_\bullet(\theta)} \psi_i + \sum_{i=1}^{P_m^\circ} \psi_{i+S_\bullet(\theta)} \right) \tag{27}$$

$$= 1 - \left( \sum_{i=1}^{S_\bullet(\theta)} \psi_i + \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{P_m^\circ} \psi_{i+S_\bullet(\theta)} \right). \tag{28}$$

Since the total number of sequences in the output space is $|\mathcal{Y}|$, then it must hold that $MS_\bullet(\theta) + \sum_{m=1}^{M} P_m^\circ = |\mathcal{Y}|$. Using (18) we obtain

$$pS_\circ(\theta) = \frac{1}{M} \sum_{m=1}^{M} P_m^\circ. \tag{29}$$

From the definition of $\mathcal{S}_{x_m}^\circ$, it follows that $\psi_i = \theta$ for $S_\bullet(\theta) + 1 \leq i \leq S_\bullet(\theta) + S_\circ(\theta)$. Since by definition, $P_m^\circ \leq S_\circ(\theta)$, we have that

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{P_m^\circ} \psi_{i+S_\bullet(\theta)} = \frac{\theta}{M} \sum_{m=1}^{M} P_m^\circ \tag{30}$$

$$= \theta p S_\circ(\theta) \tag{31}$$

$$= p \sum_{i=1}^{S_\circ(\theta)} \psi_{i+S_\bullet(\theta)}, \tag{32}$$

where (31) follows from (29). As a result, the right-hand side of (23) and (28) coincide. ∎

Eq. (12) shows that the meta-converse bound, after optimization over the auxiliary distribution $Q_Y$, coincides with the exact error probability $\epsilon(\mathcal{C})$ of any code $\mathcal{C}$ (see [12] for details). Theorem 1 shows that, for certain symmetric channels, the relaxation (13) also coincides with the minimum error probability for quasi-perfect codes, whenever they exist.

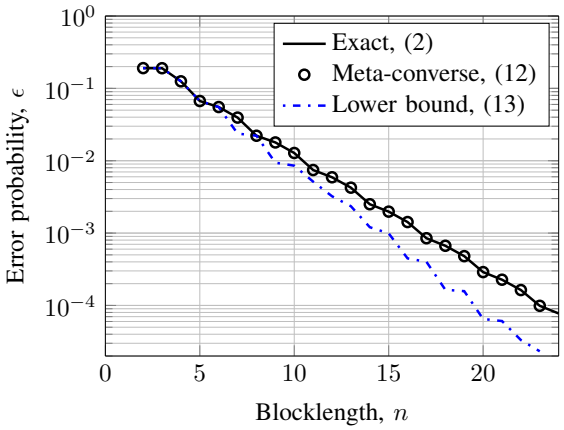



Fig. 1. Error probability for the BSC with parameters $\delta = 0.1$, $M = 4$.

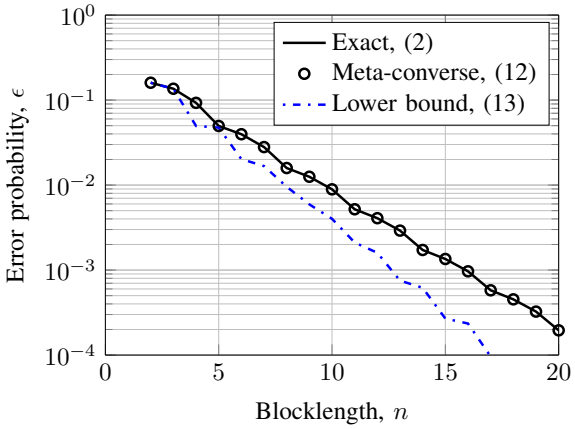



Fig. 2. Error probability for the BSC with parameters $\delta = 0.1$, $M = 3$.

Theorem 1 recovers [4, Th. 3] in the same generality. The hypothesis testing approach reported in this work is conceptually different to that in [4] and allows further extensions. For example, in this work we have restricted ourselves $Q_Y = \bar{Q}_Y$, although different $Q_Y$ are obviously possible.

*Example: BSC*

Figures 1 and 2 depict the minimal error probability for the transmission of $M$ messages over $n$ channel uses of a BSC with cross-over probability $\delta = 0.1$. We plot the exact error probability (2) and the meta-converse bound (12) computed for the best code [13], compared with the lower bound in (13).

From Fig. 1 we can see that the three curves coincide for $M = 4$ and $n = 2, 3, 4, 5, 6, 8$. According to Theorem 1, a quasi-perfect code can be built for these values of $n$ as follows. The output sequences belonging to the decision regions of each of the codewords must have the $\lceil \frac{2^n}{M} \rceil$ or $\lfloor \frac{2^n}{M} \rfloor$ largest likelihoods in $\{\psi_i\}$. For instance, for $M = 4$ and $n = 4$, this implies that the decision regions must include 1 output sequence at Hamming distance 0 to the closest codeword, and 3 output sequences at distance 1. This distance spectrum is achievable, for example, by the code $\mathcal{C} = \{0000, 0001, 1110, 1111\}$, that therefore attains the smallest error probability. Note that this code is not optimum in terms of minimum distance (see [13, Sec. IV] for details).

Similarly, Fig. 2 shows the three curves for $M = 3$. We can see that they coincide for $M = 3$ and $n = 2, 3, 5$. For $n = 4$ the decision regions of a quasi-perfect code should include 1 output sequence at Hamming distance 0 of the corresponding codeword, 4 output sequences at distance 1, and at most 1 output sequence at distance 2. However, there exists no configuration of the codewords such that three of these sets are packed in the output space. Therefore, there exists a strictly positive gap between (12) and (13) and the bound in (13) is not achievable.

*Example: BEC*

Since the binary erasure channel (BEC) is symmetric, quasi-perfect codes according to Definition 2 attain the minimum error probability. Unfortunately, these codes might not exist in general. To see this, consider a BEC with erasure probability $0 < \delta < \frac{1}{2}$. For any input $\boldsymbol{x} \in \mathcal{X}^n$, the all-erasures sequence is the least probable of the $2^n$ output sequences with non-zero probability. Therefore, for values of $\theta$ such that $S(\theta) < 2^n$, the all-erasures sequence does not belong to any set $\mathcal{S}_{\boldsymbol{x}}(\theta)$, $\boldsymbol{x} \in \mathcal{X}^n$. Since for any perfect code $S(\theta) \approx \frac{3^n}{M}$ (see (18)), even moderate values of $M$ imply that (6) does not hold, and neither perfect nor quasi-perfect codes exist.

## References


[1] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, Inc., 1968.

[2] C. Shannon, R. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. II," *Information and Control*, vol. 10, no. 5, pp. 522 – 552, 1967.

[3] T. Baicheva, I. Bouyukliev, S. Dodunekov, and V. Fack, "Binary and ternary linear quasi-perfect codes with small dimensions," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4335–4339, Sept 2008.

[4] M. Hamada, "A sufficient condition for a code to achieve the minimum decoding error probability–generalization of perfect and quasi-perfect codes," *IEICE Trans. on Fund. of Electronics, Comm. and Comp. Sciences*, vol. E83-A, no. 10, pp. 1870–1877, Oct. 2000.

[5] C. Shannon, R. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol. 10, no. 1, pp. 65 – 103, 1967.

[6] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 405–417, 1974.

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

[8] J. L. Massey. (1998) Applied Digital Information Theory I, Lecture Notes. [Online]. Available: http://www.isi.ee.ethz.ch/research.html

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. NJ: Wiley-Interscience, 2006.

[10] Y. Polyanskiy, "Saddle point in the minimax converse for channel coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2576–2595, May 2013.

[11] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Soc. Lond. A*, vol. 231, no. 694-706, p. 289, 1933.

[12] G. Vazquez-Vilar, A. Tauste Campo, A. Guillén i Fàbregas, and A. Martinez, "Bayesian M-ary hypothesis testing: The meta-converse and Verdú-Han bounds are tight," *preprint arXiv:1411.3292v2*, 2015.

[13] P.-N. Chen, H.-Y. Lin, and S. Moser, "Optimal ultrasmall block-codes for binary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7346–7378, Nov 2013.

# Author Index