International Zurich Seminar on Information and Communication (IZS 2018)

Proceedings

Conference Proceedings

Publication date:
2018-02-21

Permanent link:
https://doi.org/10.3929/ethz-b-000242151

International Zurich Seminar on Information and Communication

February 21 – 23, 2018

Sorell Hotel Zürichberg, Zurich, Switzerland

Proceedings

Acknowledgment of Support

The International Zurich Seminar on Information and Communication (IZS) is organized by the IEEE Switzerland Chapter on Digital Communication Systems, in collaboration with ETH Zurich and technically co-sponsored by the IEEE Information Theory Society. The financial responsibility lies with the ZuSem-Stiftung, Zürich.

Conference Organization

General Co-Chairs

Amos Lapidoth and Stefan M. Moser

Technical Program Committee

Helmut Bölcskei
Yuval Cassuto
Terence H. Chan
Paul Cuff
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Albert Guillén i Fàbregas
Deniz Gunduz
Martin Hänggi
Franz Hlawatsch
Johannes Huber
Ashish Khisti
Tobias Koch
Gerhard Kramer
Frank Kschischang
Hans-Andrea Loeliger
Thomas Mittelholzer
Ron Roth
Igal Sason
Jossy Sayir
Robert Schober
Karthikeyan Shanmugam
Giorgio Taricco
Emre Telatar
Pascal Vontobel
Ligong Wang
Michèle Wigger
Armin Wittneben
Ram Zamir

Organizers of Invited Sessions

Tobias Koch
Igal Sason
Stephan ten Brink
Pascal Vontobel

Local Organization

Silvia Tempel (Secretary)
Michael Lerjen (Web and Publications)
Patrick Strebel (Registration)

Table of Contents

Keynote Lectures

Session 1 Wed 10:00 – 12:00
Topics in Multiterminal Information Theory

Chaired by Anelia Somekh-Baruch (Bar-Ilan University)

*Invited papers are marked by an asterisk.

Session 2 Wed 13:20 – 15:00

Machine Learning for Communications: Theory and Applications

Invited session organizer: Stephan ten Brink (Universität Stuttgart)

Session 3 Wed 15:30 – 17:10

Quantization

Invited session organizer: Tobias Koch (Universidad Carlos III de Madrid)

Session 4 Thu 10:00 – 12:00

Shannon Theory and Secrecy

Chaired by Bernhard Geiger (Graz University of Technology)

Session 5 Thu 13:20 – 15:00

Information Theory and Statistics

Invited session organizer: Igal Sason (Technion – Israel Institute of Technology)

Session 6 Thu 15:30 – 17:10

Coding Theory

Chaired by Iryna Andriyanova (University of Cergy-Pontoise)

Session 7 Fri 10:20 – 12:00

Coded Communication

Chaired by Giuseppe Durisi (Chalmers University of Technology)

Session 8 Fri 13:20 – 14:40

Coding Theory and Applications

Invited session organizer: Pascal Vontobel (Chinese University of Hong Kong)

Session 9 Fri 15:10 – 16:30

Current Trends in Information Theory

Chaired by Albert Guillén i Fàbregas (Universitat Pompeu Fabra)

Recent-Results Posters

Wednesday, February 21

On the Asymptotic Blocklength-Dimension Tradeoff of Composite Hypothesis Testing
Michael Bell, Yuval Kochman (School of CSE, HUJI, Jerusalem, Israel)

Importance Sampling for Random Coding Error Probability Estimation
Josep Font-Segura, Alfonso Martinez (Universitat Pompeu Fabra, Barcelona, Spain)
Albert Guillén i Fàbregas (ICREA and Universitat Pompeu Fabra, Barcelona, Spain, and University of Cambridge, UK)

Thursday, February 22

On Locally Recoverable Fractional Repetition Codes
Yi-Sheng Su (Chang Jung Christian University, Tainan City, Taiwan)

Private Information Retrieval Schemes for Locally Repairable Coded Data
Razane Tajeddine, Oliver W. Gnilke (Aalto University, Espoo, Finland)
Ragnar Freij-Hollanti (Technical University of Munich, Germany)
David Karpuk (Universidad de los Andes, Bogotá, Colombia)
Camilla Hollanti (Aalto University, Espoo, Finland and Technical University of Munich, Germany)
Salim El Rouayheb (Rutgers University, New Jersey, USA)

On the L^1 Flatness Factor of Lattices
Cong Ling, Antonio Campello (Imperial College London, UK)
Ling Liu (Huawei Technologies Shenzhen, China)

Friday, February 23

Coordinated Scheduling for Multi-Cell Non-Orthogonal Multiple Access (NOMA) based Cloud-RAN System
Rupesh Singh Rai, Huiling Zhu, Jiangzhou Wang (University of Kent, UK)

A Dynamic Completion Method for RSS Map Construction
Dimitris Milioris (Nokia Bell Labs, Nozay, France)

A Necessary Condition for Source Broadcasting and Asymmetric Data Transmission

Shraga I. Bross* Hagai Zalach*,
*Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel, *brosss@biu.ac.il, hagzalach@gmail.com*

***Abstract*—We consider the broadcasting of a single Gaussian source over a two-user Gaussian broadcast channel with bandwidth expansion. In addition to the source transmission the encoder sends a message reliably to the "higher quality" user. By following the Khezeli-Chen approach we provide an alternative proof for the outer bound that we've recently obtained for this problem. This approach provides more intuition regarding the optimality of the choice of the auxiliary random variable by means of which the bound is derived.** [1]

I. Introduction and Problem Statement

Consider a communication scenario where an encoder describes a memoryless Gaussian source to a pair of users over a time-discrete additive white Gaussian broadcast channel (BC), and the number of channel uses per source sample is greater than or equal to one. In addition to the source transmission the encoder transmits a private message that should be conveyed reliably just to the "higher quality" user. Our goal is to characterize the set of mean-squared error distortion pairs that are simultaneously achievable given the private message rate. Special case is the classical joint source-channel coding problem of a memoryless Gaussian source over a Gaussian BC considered by Reznic-Feder-Zamir in [1] where the requirement for the reliable transmission of private data is removed.

Formally, the time-k outputs $(Y_{1,k}, Y_{2,k})$ of the Gaussian BC, conditioned on the input x_k, are defined by

$$Y_{i,k} = x_k + Z_{i,k} \quad i = 1,2 \tag{1}$$

where $x_k \in \mathbb{R}$ are the symbols sent by the transmitter, and $Z_{i,k}$ are the time-k additive noise terms at the corresponding outputs. Here $(Z_{1,1},\ldots,Z_{1,n})$ and $(Z_{2,1},\ldots,Z_{2,n})$ are independent memoryless vectors that are independent of $\{x_k\}_{k=1}^n$ with $Z_{i,k} \sim \mathcal{N}(0, N_i)$, $i = 1,2$, $k = 1,\ldots,n$ and it is assumed throughout that $N_2 \geq N_1$.

We denote the encoded memoryless source sequence by $\boldsymbol{S} = (S_1,\ldots,S_m)$, $S_i \sim \mathcal{N}(0,\sigma^2)$ and the channel input by $\boldsymbol{X} = (X_1,\ldots,X_n)$, so the source blocklength is m while the channel blocklength is n, hence the bandwidth expansion ratio $\rho \geq 1$ is defined by $\rho = \frac{n}{m}$. The message W is uniformly distributed over the set $\mathcal{W} = \{1,\ldots,2^{nR}\}$, and the encoder is defined by an encoding function $\varphi^{(m,n)}\colon \mathbb{R}^m \times \mathcal{W} \mapsto \mathbb{R}^n$ so that $\boldsymbol{X} = \varphi^{(m,n)}(\boldsymbol{S}, W)$. The channel input sequence is average-power limited to P, i.e. $\sum_{k=1}^n \mathbb{E}\left[X_k^2\right] \leq nP$, where $\mathbb{E}$ denotes the expectation operator.

The decoder at the first user (the "higher quality" user) consists of two mappings. The first mapping $\phi_W^{(1)}\colon \mathcal{Y}_1^n \rightarrow \{1,\ldots,2^{nR}\}$ is used to decode the message, and we denote by $\hat{W}$ the result of applying it to the received sequence $\boldsymbol{Y}_1$ while the arithmetic average of the probabilities of error associated with the different messages is denoted as $P_{\mathrm{e}}^{(n)}$. The second $\phi_S^{(1)}\colon \mathcal{Y}_1^n \rightarrow \hat{\mathcal{S}}_1^m = (\hat{\mathcal{S}}_{1,1}, \hat{\mathcal{S}}_{1,2},\ldots,\hat{\mathcal{S}}_{1,m})$ is used to reconstruct the source sequence at the "higher quality" user, and we denote by $\hat{\boldsymbol{S}}_1$ the result of applying $\phi_S^{(1)}$ to $\boldsymbol{Y}_1$. The decoder at the second user consists of the mapping $\phi_S^{(2)}\colon \mathcal{Y}_2^n \rightarrow \hat{\mathcal{S}}_2^m = (\hat{\mathcal{S}}_{2,1}, \hat{\mathcal{S}}_{2,2},\ldots,\hat{\mathcal{S}}_{2,m})$ that is used to reconstruct the source sequence at the second user, and we denote by $\hat{\boldsymbol{S}}_2$ the result of applying $\phi_S^{(2)}$ to $\boldsymbol{Y}_2$.

Definition 1: The tuple $(\tilde{R}, D_1, D_2)$ is *achievable* if, for every $\varepsilon > 0$, there exist positive integers m and $n = \rho m$ with corresponding encoder satisfying the average-power constraint, whose rate exceeds $\tilde{R} - \varepsilon$, and decoding/reconstruction mappings $\left(\phi_W^{(1)}, \phi_S^{(1)}\right)$ and $\phi_S^{(2)}$ such that $\lim_{n\to\infty} P_{\mathrm{e}}^{(n)} = 0$, and

$$\overline{\lim_{m\to\infty}} \; \frac{1}{m}\sum_{k=1}^{m} \mathbb{E}\left[(S_k - \hat{S}_{\nu,k})^2\right] \leq D_\nu + \varepsilon, \quad \nu = 1,2. \tag{2}$$

In [2] we derive the following outer bound on the set of attainable distortion pairs for our model, which henceforth will be referred as System Π.

Theorem 1: Let (D_1, D_2) be an achievable distortion pair with a message rate R, let $\alpha \geq 1$ be defined by

$$\frac{D_2}{\sigma^2} = \alpha\left(\frac{N_2}{P+N_2}2^{2R_{\mathrm{d}}}\right)^{\rho}, R_{\mathrm{d}} \triangleq \frac{1}{2}\log\left(1 + \left(2^{2R} - 1\right)\frac{N_1}{N_2}\right)$$

Then

$$D_1 \geq \sup_{\kappa>0} \frac{\sigma^2}{f(\alpha,\kappa,R)} \tag{3}$$

where

$$f(\alpha,\kappa,R) \triangleq \left\{\left(\frac{N_2}{N_1}\left[\alpha + \kappa 2^{-2\rho R_{\mathrm{d}}}\left(1+\frac{P}{N_2}\right)^{\rho}\right]^{\frac{1}{\rho}} - \left(\frac{N_2}{N_1} - 1\right)(\kappa+1)^{\frac{1}{\rho}}\right)^{\rho} - 1\right\}\frac{1}{\kappa}. \tag{4}$$

The proof of Theorem 1, which appears in [2], follows the main steps of the proof of [1, Theorem 1] with the exception that, for the evaluation of the bound when $R > 0$, one needs to assess the amount of interference caused at the "lower quality" user, by the transmission of the private message to the "higher

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

quality" user. To accomplish that we use a rate-disturbance bound in the line of [3].

For the purpose of studying the optimality of the choice of the auxiliary random variable in the proof of [1, Theorem 1], Khezeli and Chen proposed in [4] an alternative approach for proving [1, Theorem 1]. They study a model which is related to the source broadcast model for which they prove a source-channel separation theorem. Then, they leverage this result to derive a necessary condition for the source broadcast problem by means of which they obtain the Reznic-Fedef-Zamir bound and show that the choice of the auxiliary random variable in their proof is indeed optimal.

In this contribution we provide an alternative proof of Theorem 1 which follows the track of the proof in [4, Section IV, and Section VI] with the difference that for our problem we just prove a necessary condition for System $\tilde{\Pi}$—a model that is related to our model but with side-information at the "higher quality" receiver. This necessary condition is then leveraged to derive a necessary condition for our model. Since it is difficult to optimize the latter necessary condition we relax it to an extent which affords the optimization step which (as expected) uses the rate-disturbance bound of [3].

II. Proof of Theorem 1 via a necessary condition for system $\tilde{\Pi}$

Let $p_{Y_1Y_2|X}$ be a discrete memoryless broadcast channel (DMBC) with input alphabet $\mathcal{X}$ and output alphabets $\mathcal{Y}_i, i = 1, 2$. A length-n coding scheme for the DMBC $p_{Y_1Y_2|X}$, with message side-information at Decoder 1, consists of

1) Two private messages W_1 and W_2, such that (W_1, W_2) is uniformly distributed over $\mathcal{W}_1 \times \mathcal{W}_2$,
2) An encoding function $f^{(n)}\colon \mathcal{W}_1 \times \mathcal{W}_2 \to \mathcal{X}^n$,
3) Two decoding functions: $g_1^{(n)}\colon \mathcal{Y}_1^n \times \mathcal{W}_2 \to \mathcal{W}_1$, which maps (Y_1^n, W_2) to $\hat{W}_1$, and $g_2^{(n)}\colon \mathcal{Y}_2^n \to \mathcal{W}_2$, which maps Y_2^n to $\hat{W}_2$.

Definition 2: A rate pair (R_1, R_2) is achievable for the DMBC $p_{Y_1Y_2|X}$, with message W_2 available at Decoder 1, if for every $\varepsilon > 0$ there exists a sequence of encoding functions $f^{(n)}\colon \mathcal{W}_1 \times \mathcal{W}_2 \to \mathcal{X}^n$ with $\frac{1}{n} \log \|\mathcal{W}_i\| \geq R_i - \varepsilon, i = 1, 2$, and decoding functions $g_1^{(n)}\colon \mathcal{Y}_1^n \times \mathcal{W}_2 \to \mathcal{W}_1$ and $g_2^{(n)}\colon \mathcal{Y}_2^n \to \mathcal{W}_2$ such that $\lim_{n\to\infty} \Pr\{(\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)\} = 0$. The capacity region $\mathcal{C}_1(p_{Y_1Y_2|X})$ is the closure of the set of achievable (R_1, R_2) pairs.

Let the region $\mathcal{R}$ be defined as

$$\mathcal{R} = \bigcup_{P_{VXY_1Y_2}} \bigl\{(R_1, R_2) : R_1 \leq I(X; Y_1),\ R_2 \leq I(V; Y_2) \\ R_1 + R_2 \leq I(V; Y_2) + I(X; Y_1|V)\bigr\}, \tag{5}$$

where V is an auxiliary chance variable taking values in $\mathcal{V}$; the union in (5) is over all laws of the form

$$P_{VXY_1Y_2} = P_V(v)\, P_{X|V}(x|v)\, p_{Y_1Y_2|X}(y_1, y_2|x), \tag{6}$$

and it suffices to assume that $\|\mathcal{V}\| \leq \|\mathcal{X}\| + 1$. Further, let $\bar{R}$ denote the closure of $\mathcal{R}$.

In [5, Theorem 3] the authors characterize $\mathcal{C}_1(p_{Y_1Y_2|X})$ as follows: $\mathcal{C}_1(p_{Y_1Y_2|X}) = \bar{R}$.

Following [4] let $\{S_t\}_{t=1}^{\infty}$ in System Π (the model of our problem) be an IID vector Gaussian process, where each S_k is an $\ell \times 1$ zero-mean Gaussian random vector with positive definite covariance matrix Σ_S.

Definition 3: Let ρ be a non-negative number, fix a non-negative rate R, and let $\mathcal{D}_i$, $i = 1, 2$ be a non-empty compact subset of the $\ell \times \ell$ positive semi-definite matrices $\bigl\{D\colon \mathbf{0} \preceq D \preceq \Sigma_S\bigr\}$. We say that $(\rho, R, \mathcal{D}_1, \mathcal{D}_2)$ is achievable for System Π if, for every $\varepsilon > 0$, there exist encoding function $\varphi^{(m,n)}\colon \mathbb{R}^{\ell\times m} \times \mathcal{W} \to \mathcal{X}^n$ as well as decoding and reconstruction functions $\phi_W^{(1)}\colon \mathcal{Y}_1^n \to \{1, \ldots, 2^{nR}\}$ and $\phi_S^{(1)}\colon \mathcal{Y}_1^n \to \mathbb{R}^{\ell\times m}$ at the "higher quality" receiver, and reconstruction function $\phi_S^{(2)}\colon \mathcal{Y}_2^n \to \mathbb{R}^{\ell\times m}$ at the "lower quality" receiver, such that with $\phi_S^{(i)}(Y_i^n) = \hat{S}_i^m$ and with $\|\cdot\|$ denoting the 1-norm,

$$\frac{n}{m} \leq \rho + \varepsilon \tag{7a}$$

$$\min_{D_i\in\mathcal{D}_i} \left\| \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\bigl[(S_t - \hat{S}_{i,t})(S_t - \hat{S}_{i,t})^T - D_i\bigr] \right\| \leq \varepsilon, \\ i = 1, 2 \tag{7b}$$

$$\lim_{n\to\infty} P_{\mathrm{e}}^{(n)} = 0. \tag{7c}$$

The set of achievable $(\rho, R, \mathcal{D}_1, \mathcal{D}_2)$ tuples for System Π is denoted by Γ_{G}.

The system $\tilde{\Pi}$ shown in Figure 1, that we consider next, is the same as System Π except for two differences.

1) Let $\tilde{S} \triangleq \bigl(\tilde{S}_1^T, \tilde{S}_2^T\bigr)^T$ be an $\tilde{\ell} \times 1$ zero-mean Gaussian random vector with positive definite covariance matrix $\Sigma_{\tilde{S}}$, where $\tilde{S}_i$ is an $\tilde{\ell}_i \times 1$ random vector with covariance matrix $\Sigma_{\tilde{S}_i}, i = 1, 2$. The source $\bigl\{\tilde{S}_t\bigr\}_{t=1}^{\infty}$ is defined by $\tilde{S}_t = (\tilde{S}_{1,t}^T, \tilde{S}_{2,t}^T)^T$, where $\bigl\{(\tilde{S}_{1,t}, \tilde{S}_{2,t})\bigr\}_{t=1}^{\infty}$ are IID copies of $(\tilde{S}_1, \tilde{S}_2)$.
2) $\tilde{S}_2^m$ is available at Receiver 1 so that Decoder 1 is defined by a decoding mapping $\tilde{\phi}_W^{(1)}\colon \mathcal{Y}_1^n \times \tilde{S}_2^m \to \{1, \ldots, 2^{nR}\}$ and a reconstruction mapping $\tilde{\phi}_S^{(1)}\colon \mathcal{Y}_1^n \times \tilde{S}_2^m \to \hat{\mathcal{S}}_1^m$.

Definition 4: Let $\tilde{\rho}$ be a non-negative number, fix a non-negative rate R, let $\tilde{\mathcal{D}}_1$ be a non-empty compact subset of $\bigl\{\tilde{D}_1\colon \mathbf{0} \preceq \tilde{D}_1 \preceq \Sigma_{\tilde{S}}\bigr\}$, and let $\tilde{\mathcal{D}}_2$ be a non-empty compact subset of $\bigl\{\tilde{D}_2\colon \mathbf{0} \preceq \tilde{D}_2 \preceq \Sigma_{\tilde{S}_2}\bigr\}$. We say that $(\tilde{\rho}, R, \tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2)$ is achievable for System $\tilde{\Pi}$ if, for every $\varepsilon > 0$, there exist encoding function $\tilde{\varphi}^{(m,n)}\colon \mathbb{R}^{\tilde{\ell}_1\times m} \times \mathbb{R}^{\tilde{\ell}_2\times m} \times \mathcal{W} \to \mathcal{X}^n$ as well as decoding and reconstruction functions $\tilde{\phi}_W^{(1)}\colon \mathcal{Y}_1^n \times \mathbb{R}^{\tilde{\ell}_2\times m} \to \{1, \ldots, 2^{nR}\}$ and $\tilde{\phi}_S^{(1)}\colon \mathcal{Y}_1^n \times \mathbb{R}^{\tilde{\ell}_2\times m} \to \mathbb{R}^{\tilde{\ell}\times m}$ at the "higher quality" receiver, and reconstruction function $\tilde{\phi}_S^{(2)}\colon \mathcal{Y}_2^n \to \mathbb{R}^{\tilde{\ell}_2\times m}$ at the "lower quality" receiver, such that

$$\frac{n}{m} \leq \tilde{\rho} + \varepsilon \tag{8a}$$

$$\min_{\tilde{D}_1\in\tilde{\mathcal{D}}_1} \left\| \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\bigl[(\tilde{S}_t - \hat{S}_{1,t})(\tilde{S}_t - \hat{S}_{1,t})^T - \tilde{D}_1\bigr] \right\| \leq \varepsilon \tag{8b}$$

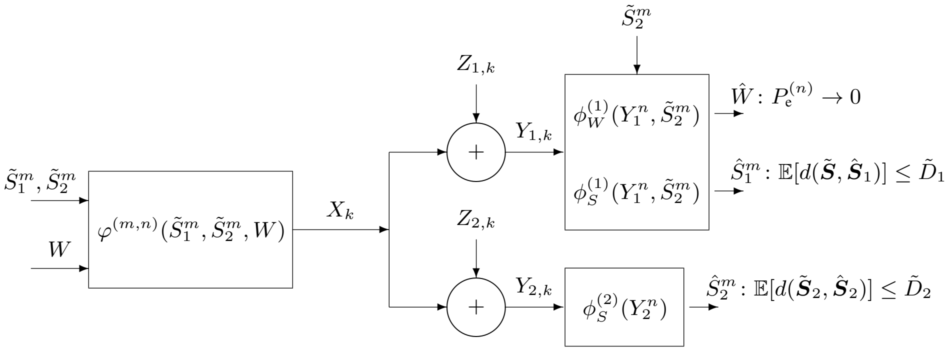

$\tilde{S}_2^m$
$Z_{1,k}$
$\hat{W}: P_e^{(n)} \to 0$
$\phi_W^{(1)}(Y_1^n, \tilde{S}_2^m)$
$Y_{1,k}$
$\tilde{S}_1^m, \tilde{S}_2^m$
$\hat{S}_1^m: \mathbb{E}[d(\tilde{\boldsymbol{S}}, \hat{\boldsymbol{S}}_1)] \le \tilde{D}_1$
$\phi_S^{(1)}(Y_1^n, \tilde{S}_2^m)$
X_k
$\varphi^{(m,n)}(\tilde{S}_1^m, \tilde{S}_2^m, W)$
W
$Z_{2,k}$
$Y_{2,k}$
$\hat{S}_2^m: \mathbb{E}[d(\tilde{\boldsymbol{S}}_2, \hat{\boldsymbol{S}}_2)] \le \tilde{D}_2$
$\phi_S^{(2)}(Y_2^n)$
+
+

Fig. 1. The system $\tilde{\Pi}$.

$$\min_{\tilde{D}_2 \in \tilde{\mathcal{D}}_2} \left\| \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{2,t} - \hat{S}_{2,t})(\tilde{S}_{2,t} - \hat{S}_{2,t})^T - \tilde{D}_2\big] \right\| \le \varepsilon \quad (8c)$$

$$\lim_{n\to\infty} P_e^{(n)} = 0. \quad (8d)$$

The set of achievable $(\tilde{\rho}, R, \tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2)$ tuples for System $\tilde{\Pi}$ is denoted by $\tilde{\Gamma}_G$. Consequently, with $\Sigma_{\tilde{S}_1,\tilde{S}_2} \triangleq \mathbb{E}\big[\tilde{S}_1\tilde{S}_2^T\big]$ and $\Sigma_{\tilde{S}_2,\tilde{S}_1} \triangleq \mathbb{E}\big[\tilde{S}_2\tilde{S}_1^T\big]$,

$$\Sigma_{\tilde{S}} = \begin{bmatrix} \Sigma_{\tilde{S}_1} & \Sigma_{\tilde{S}_1,\tilde{S}_2} \\ \Sigma_{\tilde{S}_2,\tilde{S}_1} & \Sigma_{\tilde{S}_2} \end{bmatrix}$$

Furthermore, for any $\tilde{D}_1 \in \tilde{\mathcal{D}}_1$ we may express $\tilde{D}_1$ as follows

$$\tilde{D}_1 = \begin{bmatrix} \tilde{D}_{1,1} & \tilde{D}_{1,2} \\ \tilde{D}_{2,1} & \tilde{D}_{2,2} \end{bmatrix}$$

where $\tilde{D}_{i,i},\ i = 1,2$ is an $\ell_i \times \ell_i$ matrix.

Assuming that $\tilde{D}_{2,2}$ is invertible, define

$$R_{\tilde{S}_1|\tilde{S}_2}(\tilde{\mathcal{D}}_1) = \min_{\tilde{D}_1 \in \tilde{\mathcal{D}}_1} \frac{1}{2} \log \left(\frac{|\Sigma_{\tilde{S}_1} - \Sigma_{\tilde{S}_1,\tilde{S}_2}\Sigma_{\tilde{S}_2}^{-1}\Sigma_{\tilde{S}_2,\tilde{S}_1}|}{|\tilde{D}_{1,1} - \tilde{D}_{1,2}\tilde{D}_{2,2}^{-1}\tilde{D}_{2,1}|} \right)$$
$$R_{\tilde{S}_2}(\tilde{\mathcal{D}}_2) = \min_{\tilde{D}_2 \in \tilde{\mathcal{D}}_2} \frac{1}{2} \log \left(\frac{|\Sigma_{\tilde{S}_2}|}{|\tilde{D}_2|} \right). \quad (9)$$

We start by deriving a necessary condition for System $\tilde{\Pi}$.

Proposition 1: A necessary condition for the inclusion $(\tilde{\rho}, R, \tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2) \in \tilde{\Gamma}_G$ is that

$$\Big(R_{\tilde{S}_1|\tilde{S}_2}(\tilde{\mathcal{D}}_1) + \tilde{\rho}R, R_{\tilde{S}_2}(\tilde{\mathcal{D}}_2)\Big) \in \tilde{\rho}\,\mathcal{C}_1(p_{Y_1Y_2|X}). \quad (10)$$

Proof: See Section III.

Let $\{Z_t\}_{t=1}^{\infty}$ be an IID vector Gaussian process, independent of $\{S_t\}_{t=1}^{\infty}$, where each component Z_t is an $\ell \times 1$ zero-mean Gaussian random vector with positive definite covariance matrix Σ_Z. Define

$$\tilde{S}_{1,t} \triangleq S_t, \quad \tilde{S}_{2,t} \triangleq S_t + Z_t. \quad (11)$$

Thus, if a tuple $(\rho, R, \mathcal{D}_1, \mathcal{D}_2) \in \Gamma_G$ the conditions (7b) imply that there exists a sequence $\varepsilon_1, \varepsilon_2, \ldots$ converging to zero such that

$$\lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(S_t - \hat{S}_{i,t}^{(\varepsilon_k)})(S_t - \hat{S}_{i,t}^{(\varepsilon_k)})^T\big] = D_i$$

for some $D_i \in \mathcal{D}_i,\ i = 1,2$. Therefore, with the definitions (11)

$$\begin{aligned}
&\lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{1,t} - \hat{S}_{1,t}^{(\varepsilon_k)})(\tilde{S}_{1,t} - \hat{S}_{1,t}^{(\varepsilon_k)})^T\big] \\
&\quad = \lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{1,t} - \hat{S}_{1,t}^{(\varepsilon_k)})(\tilde{S}_{2,t} - \hat{S}_{1,t}^{(\varepsilon_k)})^T\big] \\
&\quad = \lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{2,t} - \hat{S}_{1,t}^{(\varepsilon_k)})(\tilde{S}_{1,t} - \hat{S}_{1,t}^{(\varepsilon_k)})^T\big] \\
&\quad = D_1, \\
&\lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{2,t} - \hat{S}_{1,t}^{(\varepsilon_k)})(\tilde{S}_{2,t} - \hat{S}_{1,t}^{(\varepsilon_k)})^T\big] \\
&\quad = D_1 + \Sigma_Z, \\
&\lim_{k\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\big[(\tilde{S}_{2,t} - \hat{S}_{2,t}^{(\varepsilon_k)})(\tilde{S}_{2,t} - \hat{S}_{2,t}^{(\varepsilon_k)})^T\big] \\
&\quad = \tilde{D}_2 = D_2 + \Sigma_Z.
\end{aligned}$$

This establishes that $(\rho, R, \{\tilde{D}_1\}, \{\tilde{D}_2\}) \in \tilde{\Gamma}_G$ where

$$\tilde{D}_1 = \begin{bmatrix} D_1 & D_1 \\ D_1 & D_2 + \Sigma_Z \end{bmatrix}.$$

Consequently, by Proposition 1,

$$\Big(R_{\tilde{S}_1|\tilde{S}_2}(\tilde{D}_1) + \rho R, R_{\tilde{S}_2}(\tilde{D}_2)\Big) \in \rho\,\mathcal{C}_1(p_{Y_1Y_2|X}). \quad (12)$$

In conclusion, for any $(\rho, R, \mathcal{D}_1, \mathcal{D}_2) \in \Gamma_G$, there exist $D_i \in \mathcal{D}_i,\ i = 1,2$ such that

$$R_{\tilde{S}_1|\tilde{S}_2}(\tilde{D}_1) + \rho R \le \rho I(X;Y_1) \quad (13a)$$
$$R_{\tilde{S}_2}(\tilde{D}_2) \le \rho I(V;Y_2) \quad (13b)$$
$$R_{\tilde{S}_1|\tilde{S}_2}(\tilde{D}_1) + \rho R + R_{\tilde{S}_2}(\tilde{D}_2) \le \rho\big[I(X;Y_1|V) + I(V;Y_2)\big], \quad (13c)$$

where the tuple (V, X, Y_1, Y_2) has a law of the form (6), and using (9) the conditional rate-distortion and rate-distortion functions are expressed by

$$R_{\tilde{S}_1|\tilde{S}_2}(\tilde{D}_1) = \frac{1}{2}\log\Big(\frac{|\Sigma_S - \Sigma_S(\Sigma_S + \Sigma_Z)^{-1}\Sigma_S|}{|D_1 - D_1(D_1 + \Sigma_Z)^{-1}D_1|}\Big)$$

$$= \frac{1}{2}\log\Big(\frac{|\Sigma_S||D_1+\Sigma_Z|}{|D_1||\Sigma_S+\Sigma_Z|}\Big)$$

$$R_{\tilde{S}_2}(\tilde{D}_2) = \frac{1}{2}\log\Big(\frac{|\Sigma_S+\Sigma_Z|}{|D_2+\Sigma_Z|}\Big). \tag{14}$$

Since

$$I(X;Y_1|V) = h(Y_1|V) - h(Y_1|X,V) = h(Y_1|V) - h(Y_1|X) \\ \leq h(Y_1) - h(Y_1|X) = I(X;Y_1),$$

when given $R < I(X;Y_1)$, the set of attainable distortion tuples is defined by the active constraints (13b)–(13c). Nevertheless, unlike [4, Section VI, inequality (66)], it is not clear how to maximize $I(X;Y_1|V)$ subject to the constraint (13b), for the scalar Gaussian BC $p_{Y_1Y_2|X}$, because the precise coupling between V defined by (6) and W is unknown, yet it is conceivable that the rate R imposes a restriction on the set $\tilde{\mathcal{D}}_2$. We suggest the following approach. By (23)

$$nI(X;Y_1|V) = \sum_{k=1}^{n} I(X_k;Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m)$$
$$= \sum_{k=1}^{n}\big[H(Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m) \\ -H(Y_{1,k}|X_kY_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m)\big]$$
$$\overset{(a)}{=} \sum_{k=1}^{n}\big[H(Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m) - H(Y_{1,k}|X^nY_1^{k-1}\tilde{S}_2^m)\big]$$
$$\leq \sum_{k=1}^{n}\big[H(Y_{1,k}|Y_1^{k-1}\tilde{S}_2^m) - H(Y_{1,k}|X^nY_1^{k-1}\tilde{S}_2^m)\big]$$
$$= H(Y_1^n|\tilde{S}_2^m) - H(Y_1^n|X^n\tilde{S}_2^m) = I(X^n;Y_1^n|\tilde{S}_2^m), \tag{15}$$

were (a) follows since $(X^{n\setminus k},Y_{2,k+1}^n) \ominus (X_k,Y_1^{k-1},\tilde{S}_2^m) \ominus Y_{1,k}$ forms a Markov chain.

Also by (22) and (24)

$$mR_{\tilde{S}_2}(\tilde{D}_2) \leq I(\tilde{S}_2^m;\hat{S}_2^m) \leq I(\tilde{S}_2^m;Y_2^n) \leq nI(V;Y_2), \tag{16}$$

Letting $\boldsymbol{U} \triangleq \tilde{S}_2^m$, which according to (11) is independent of W, the combination of (15) and (16) yields

$$\frac{1}{2}\log\Big(\frac{|\Sigma_S+\Sigma_Z|}{|D_2+\Sigma_Z|}\Big) = R_{\tilde{S}_2}(\tilde{D}_2) \leq \frac{1}{m}I(\boldsymbol{U};\boldsymbol{Y}_2) \leq \rho I(V;Y_2) \tag{17a}$$

$$nI(X;Y_1|V) \leq I(\boldsymbol{X};\boldsymbol{Y}_1|\boldsymbol{U}). \tag{17b}$$

Thus, given R, we shall maximize the RHS of (17b) subject to the constraint (17a) on $I(\boldsymbol{Y}_2;\boldsymbol{U})$, for the scalar Gaussian BC $p_{Y_1Y_2|X}$, where now the coupling between W and $\boldsymbol{U}$ is well defined.

Consider the identity

$$I(\boldsymbol{X};\boldsymbol{Y}_1|\boldsymbol{U}) = I(\boldsymbol{X};W|\boldsymbol{U}) + I(\boldsymbol{X};\boldsymbol{Y}_1|\boldsymbol{U},W) \\ -I(\boldsymbol{X};W|\boldsymbol{Y}_1,\boldsymbol{U})$$
$$= I(\boldsymbol{X};W|\boldsymbol{U}) + h(\boldsymbol{Y}_1|\boldsymbol{U},W) - h(\boldsymbol{Z}_1) - h(W|\boldsymbol{U},\boldsymbol{Y}_1) \\ +h(W|\boldsymbol{U},\boldsymbol{X})$$
$$= h(W|\boldsymbol{U}) + h(\boldsymbol{Y}_1|\boldsymbol{U},W) - h(\boldsymbol{Z}_1) - h(W|\boldsymbol{U},\boldsymbol{Y}_1). \tag{18}$$

Now, by Fano's inequality

$$h(W|\boldsymbol{U}) + h(\boldsymbol{Y}_1|\boldsymbol{U},W) - h(\boldsymbol{Z}_1) - h(W|\boldsymbol{U},\boldsymbol{Y}_1) \\ \leq h(W) + h(\boldsymbol{Y}_1|\boldsymbol{U},W) - h(\boldsymbol{Z}_1) + n\delta(\epsilon) \\ = nR + h(\boldsymbol{Y}_1|\boldsymbol{U},W) - \frac{n}{2}\log 2\pi eN_1 + n\delta(\epsilon), \tag{19}$$

where $\delta(\epsilon) \to 0$ as $n \to \infty$. The combination of (13b)–(13c), (17a)–(17b), (18), and (19) suggests that, given R, the conditional rate-distortion function $R_{\tilde{S}_1|\tilde{S}_2}(\tilde{D}_1)$ is upper bounded by the maximum of $\frac{1}{m}[h(\boldsymbol{Y}_1|\boldsymbol{U},W) - n/2\log 2\pi eN_1]$, where the maximization is subject to the l.h.s. constraint (17a) holding with equality for an admissible D_2.

The latter maximization is done in [2], for the case $\ell = 1$ and the choice $\Sigma_Z = \kappa\sigma^2$, and subject to the constraint

$$\frac{1}{m}I(\boldsymbol{U};\boldsymbol{Y}_2) \geq \frac{1}{2}\log\frac{(1+\kappa)\sigma^2}{D_2+\kappa\sigma^2},$$

where an admissible D_2 is of the form $D_2 = \alpha\sigma^2\left(\frac{N_2}{P+N_2}2^{2R_{\mathrm{d}}}\right)^{\rho}$, for some $\alpha \geq 1$. We obtain that

$$h(\boldsymbol{Y}_1|W,\boldsymbol{U}) \leq \frac{n}{2}\log\Bigg\{2\pi e(P+N_2)2^{-2R_{\mathrm{d}}}\left[\frac{D_2/\sigma^2+\kappa}{\kappa+1}\right]^{\frac{1}{\rho}} \\ -2\pi e(N_2-N_1)\Bigg\}.$$

Thus,

$$\frac{1}{2}\log\frac{D_1+\kappa\sigma^2}{D_1(1+\kappa)} \leq \frac{\rho}{2}\log\Bigg\{\Big(\frac{P+N_2}{N_1}\Big)2^{-2R_{\mathrm{d}}}\left[\frac{D_2/\sigma^2+\kappa}{\kappa+1}\right]^{\frac{1}{\rho}} \\ -\Big(\frac{N_2-N_1}{N_1}\Big)\Bigg\}$$

which recovers the bound (3)–(4).

III. Proof of Proposition 1

Consider an arbitrary tuple $(\tilde{\rho},R,\tilde{\mathcal{D}}_1,\tilde{\mathcal{D}}_2) \in \tilde{\Gamma}_{\mathrm{G}}$. Given $\varepsilon > 0$, then there exists an encoding function $\tilde{\varphi}^{(m,n)}\colon \tilde{\mathcal{S}}_1^m \times \tilde{\mathcal{S}}_2^m \times \mathcal{W} \to \mathcal{X}^n$ as well as decoding and reconstruction functions $\tilde{\phi}_W^{(1)}\colon \mathcal{Y}_1^n \times \tilde{S}_2^m \to \{1,\ldots,2^{nR}\}$ and $\tilde{\phi}_S^{(1)}\colon \mathcal{Y}_1^n \times \tilde{S}_2^m \to \hat{\mathcal{S}}_1^m$ at the "higher quality" receiver, and reconstruction function $\tilde{\phi}_S^{(2)}\colon \mathcal{Y}_2^n \to \hat{\mathcal{S}}_2^m$ at the "lower quality" receiver, such that (8b)–(8d) are satisfied. Let Q be a random variable independent of $(W,\tilde{S}_1^m,\tilde{S}_2^m,X^n,Y_1^n,Y_2^n)$ and uniformly distributed over $\{1,\ldots,n\}$. Define $X = X_Q$, $Y_i = Y_{i,Q}, i = 1,2$ and $V = (V_Q,Q)$ where $V_k \triangleq (Y_1^{k-1},Y_{2,k+1}^n,\tilde{S}_2^m)$ so that $V_k \ominus X_k \ominus (Y_{1,k},Y_{2,k})$ forms a Markov chain hence $V \ominus X \ominus (Y_1,Y_2)$ is also a Markov chain. Note that, by Fano's inequality

$$n(R-\eta_n) \leq I(W;Y_1^n\tilde{S}_2^m) = I(W;Y_1^n|\tilde{S}_2^m) \tag{20}$$

where $\lim_{n\to\infty}\eta_n = 0$ and the last equality follows since W is independent of $\tilde{S}_2^m$. Thus,

$$I(\tilde{S}_1^m;\hat{S}_1^m|\tilde{S}_2^m) + n(R-\eta_n) \leq I(\tilde{S}_1^m;\hat{S}_1^m|\tilde{S}_2^m) \\ +I(W;Y_1^n|\tilde{S}_2^m)$$

$$
\begin{aligned}
&\le I(\tilde{S}_1^m;\hat{S}_1^m Y_1^n|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_2^m)\\
&\overset{(a)}{=} I(\tilde{S}_1^m;Y_1^n|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_2^m)\\
&\overset{(b)}{\le} I(\tilde{S}_1^m\tilde{S}_2^m;Y_1^n) + I(W;Y_1^n|\tilde{S}_1^m\tilde{S}_2^m) = I(W\tilde{S}_1^m\tilde{S}_2^m;Y_1^n)\\
&\overset{(c)}{=} I(W\tilde{S}_1^m\tilde{S}_2^m X^n;Y_1^n) \overset{(d)}{=} I(X^n;Y_1^n)\\
&= \sum_{k=1}^n I(X^n;Y_{1,k}|Y_1^{k-1}) \le \sum_{k=1}^n I(X^n Y_1^{k-1};Y_{1,k})\\
&\overset{(e)}{=} \sum_{k=1}^n \big[H(Y_{1,k}) - H(Y_{1,k}|X_k\big] = \sum_{k=1}^n I(X_k;Y_{1,k})\\
&\quad = nI(X_Q;Y_{1,Q}|Q) \overset{(f)}{=} nI(Q,X_Q;Y_{1,Q})\\
&\quad \overset{(f)}{=} nI(X_Q;Y_{1,Q}) = nI(X;Y_1).
\end{aligned} \tag{21}
$$

Here

(a) follows since $\hat{S}_1^m$ is a function of $(Y_1^n,\tilde{S}_2^m)$;

(b) follows since, conditioned on $\tilde{S}_2^m$, W is independent of $\tilde{S}_1^m$, since conditioning cannot increase entropy and because mutual-information is non-negative;

(c) follows since X^n is a function of $(W,\tilde{S}_1^m,\tilde{S}_2^m)$;

(d) follows since $(W,\tilde{S}_1^m,\tilde{S}_2^m) -\!\!\circ\!\!- X^n -\!\!\circ\!\!- Y_1^n$ forms a Markov chain;

(e) follows since $Y_{1,k} -\!\!\circ\!\!- X_k -\!\!\circ\!\!- (X^{n\backslash k},Y_1^{k-1})$ forms a Markov chain; and

(f) follows since Q is independent of $Y_{1,Q}$ and since $Y_{1,Q} -\!\!\circ\!\!- X_{1,Q} -\!\!\circ\!\!- Q$ forms a Markov chain.

Next,

$$
\begin{aligned}
&I(\tilde{S}_2^m;\hat{S}_2^m) \le I(\tilde{S}_2^m;\hat{S}_2^m Y_2^n) \overset{(a)}{=} I(\tilde{S}_2^m;Y_2^n)\\
&= \sum_{k=1}^n I(\tilde{S}_2^m;Y_{2,k}|Y_{2,k+1}^n) \;\le \sum_{k=1}^n I(\tilde{S}_2^m Y_{2,k+1}^n Y_1^{k-1};Y_{2,k})\\
&= \sum_{k=1}^n I(V_k;Y_{2,k}) = nI(V_Q;Y_{2,Q}|Q) \le nI(Q,V_Q;Y_{2,Q})\\
&= nI(V;Y_2),
\end{aligned} \tag{22}
$$

where (a) follows since $\hat{S}_2^m$ is a function of Y_2^n. Finally,

$$
\begin{aligned}
&I(\tilde{S}_1^m;\hat{S}_1^m|\tilde{S}_2^m) + n(R-\eta_n) + I(\tilde{S}_2^m;\hat{S}_2^m)\\
&\le I(\tilde{S}_1^m;\hat{S}_1^m|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;\hat{S}_2^m)\\
&\le I(\tilde{S}_1^m;\hat{S}_1^m Y_1^n|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;\hat{S}_2^m Y_2^n)\\
&\overset{(a)}{=} I(\tilde{S}_1^m;Y_1^n|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_2^n)\\
&\overset{(b)}{\le} I(\tilde{S}_1^m;Y_1^n|\tilde{S}_2^m) + I(W;Y_1^n|\tilde{S}_1^m\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_2^n)\\
&= I(W\tilde{S}_1^m;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_2^n)\\
&\overset{(c)}{=} I(W\tilde{S}_1^m X^n;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_2^n)\\
&\overset{(d)}{=} I(X^n;Y_1^n|\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_2^n)\\
&\overset{(e)}{=} \sum_{k=1}^n \big[I(X_k;Y_{1,k}|Y_1^{k-1}\tilde{S}_2^m) + I(\tilde{S}_2^m;Y_{2,k}|Y_{2,k+1}^n)\big]\\
&\le \sum_{k=1}^n \big[I(X_k Y_{2,k+1}^n;Y_{1,k}|Y_1^{k-1}\tilde{S}_2^m) + I(\tilde{S}_2^m Y_{2,k+1}^n;Y_{2,k})\big]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^n \big[I(X_k;Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m)\\
&\qquad\quad + I(Y_{2,k+1}^n;Y_{1,k}|Y_1^{k-1}\tilde{S}_2^m) + I(\tilde{S}_2^m Y_{2,k+1}^n;Y_{2,k})\big]\\
&\overset{(f)}{=} \sum_{k=1}^n \big[I(X_k;Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m)\\
&\qquad\quad + I(Y_1^{k-1};Y_{2,k}|Y_{2,k+1}^n\tilde{S}_2^m) + I(\tilde{S}_2^m Y_{2,k+1}^n;Y_{2,k})\big]\\
&= \sum_{k=1}^n \big[I(X_k;Y_{1,k}|Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m)\\
&\qquad\qquad + I(Y_1^{k-1}Y_{2,k+1}^n\tilde{S}_2^m;Y_{2,k})\big]\\
&\quad = \sum_{k=1}^n \big[I(X_k;Y_{1,k}|V_k) + I(V_k;Y_{2,k})\big]\\
&\quad = n[I(X_Q;Y_{1,Q}|V_Q,Q) + I(V_Q;Y_{2,Q}|Q)]\\
&\quad \overset{(g)}{=} n[I(X_Q;Y_{1,Q}|V_Q,Q) + I(V_Q,Q;Y_{2,Q})]\\
&\quad = n[I(X;Y_1|V) + I(V;Y_2)].
\end{aligned} \tag{23}
$$

Here

(a) follows since $\hat{S}_1^m$ is a function of $(Y_1^n,\tilde{S}_2^m)$ and $\hat{S}_2^m$ is a function of Y_2^n;

(b) follows since, conditioned on $\tilde{S}_2^m$, W is independent of $\tilde{S}_1^m$, and since conditioning cannot increase entropy;

(c) follows since X^n is a function of $(W,\tilde{S}_1^m,\tilde{S}_2^m)$;

(d) follows since $(W,\tilde{S}_1^m) -\!\!\circ\!\!- (X^n,\tilde{S}_2^m) -\!\!\circ\!\!- Y_1^n$ forms a Markov chain;

(e) follows since $Y_{1,k} -\!\!\circ\!\!- (\tilde{S}_2^m,X_k,Y_1^{k-1}) -\!\!\circ\!\!- X^{n\backslash k}$ forms a Markov chain;

(f) follows by the Csiszár-Körner's identity [6, Lemma 7];

(g) follows since Q is independent of $Y_{2,Q}$.

Since [4, Section IV]

$$
\begin{aligned}
I(\tilde{S}_1^m;\hat{S}_1^m|\tilde{S}_2^m) &\ge mI(\tilde{S}_1;\hat{S}_1^{(\varepsilon)}|\tilde{S}_2)\\
I(\tilde{S}_2^m;\hat{S}_2^m) &\ge mI(\tilde{S}_2;\hat{S}_2^{(\varepsilon)}),
\end{aligned} \tag{24}
$$

the combination of (21), (22), (23), and (24) yields that

$$
\big(I(\tilde{S}_1;\hat{S}_1^{(\varepsilon)}|\tilde{S}_2) + \frac{n}{m}R, I(\tilde{S}_2;\hat{S}_2^{(\varepsilon)})\big) \in \frac{n}{m}\mathcal{C}_1(p_{Y_1Y_2|X}). \tag{25}
$$

In the quadratic Gaussian setting the inclusion (25) translates to (10).

References

[1] Z. Reznic, M. Feder and R. Zamir, “Distortion bounds for broadcasting with bandwidth expansion,” *IEEE Trans. Inform. Theory,* Vol. IT-52, No. 8, pp. 3778-3788, Aug. 2006.

[2] S. I. Bross, and H. Zalach, “Distortion bounds for source broadcasting and asymmetric data transmission with bandwidth expansion,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),* June 2017, pp. 101-105.

[3] B. Bandemer and A. El Gamal, “Communication with disturbance constraints,” *IEEE Trans. Inform. Theory,* vol. 60, no. 8, pp. 4488-4502, Aug. 2014.

[4] K. Khezeli and J. Chen, “A source-channel separation theorem with application to the source broadcast problem,” *IEEE Trans. Inform. Theory,* vol. 62, no. 4, pp. 1764-1781, April 2016.

[5] G. Kramer and S. Shamai (Shitz), “Capacity for classes of broadcast channels with receiver side information,” *IEEE Inf. Theory Workshop,* Lake-Thhoe, CA, USA, Sep. 2007, pp. 313-318.

[6] I. Csiszár and J. Körner, “Broadcast channels with confidential messages,” *IEEE Trans. Inform. Theory,* vol. 24, no. 3 pp. 339-348, May 1978.

Expurgated Bounds for the Asymmetric Broadcast Channel

Ran Averbuch, Nir Weinberger* and Neri Merhav
Department of Electrical Engineering,
Technion – Israel Institute of Technology,
Technion City, Haifa 3200003, Israel,
Email: nir.wein@gmail.com, {rans@campus, merhav@ee}.technion.ac.il

Abstract— **This work contains two main contributions concerning the expurgation of hierarchical ensembles for the asymmetric broadcast channel. The first is an analysis of the optimal maximum likelihood (ML) decoders for the weak and strong user. Two different methods of code expurgation will be used, that will provide two competing error exponents. The second is the derivation of expurgated exponents under the generalized stochastic likelihood decoder (GLD). We prove that the GLD exponents are at least as tight as the maximum between the random coding error exponents derived in an earlier work by Averbuch and Merhav (2017) and one of our ML–based expurgated exponents. By that, we actually prove the existence of hierarchical codebooks that achieve the best of the random coding exponent and the expurgated exponent simultaneously for both users.**

I. Introduction

One of the most elementary system configuation models in multi-user information theory is the broadcast channel (BC). It has been introduced more than four decades ago by Cover [1], and since then, a vast amount of papers and books, analyzing different aspects of the broadcast model, have been published. Although the characterization of the capacity region of the general BC is still an open problem, some special cases have been solved. Most notably, the broadcast channel with degraded message sets, also known as the asymmetric broadcast channel (ABC), was introduced and solved by Körner and Marton [2].

While the capacity region of the ABC has been known for many years, only little is known about its reliability functions. The earliest work on error exponents for the general ABC is of Körner and Sgarro [3]. Later, Kaspi and Merhav [4] have derived tighter lower bounds to the reliability functions of both users by analyzing random coding error exponents of their optimal decoders. Most recently [5], the exact random coding error exponents have been determined for both the *strong user* and the *weak user*, under the ensemble of fixed composition codes.

Even in the single–user case, it is known for many years that the random coding error exponent is not tight (with respect to the reliability function) for relatively low coding rates, and may be improved by expurgation [6], [10]. Specifically, improved bounds are obtained by eliminating codewords that contribute relatively highly to the error probability, and asserting that some upper bound holds for all remaining codewords.

The main objective of this paper is to study expurgation techniques for the hierarchical ensemble used over the ABC. Expurgating a code for the ABC is not a trivial extension of expurgation in the single-user case, because there might be conflicting goals from the viewpoints of the two users. Nonetheless, we were able to define expurgation procedures that guarantee no harm to the performance of either user. This has paved the way to derive tighter lower bounds on the reliability functions of the ABC.

We start by analyzing the optimal maximum likelihood (ML) decoder, and derive some expurgated bounds, that are natural generalizations of the single–user expurgated bound due to Csiszár, Körner and Marton (CKM) [6]. Although our first process of code expurgation is fairly intuitive, there is at least one specific step in our first derivation where exponential tightness might be compromised. This point gives rise to a possible room for improvement upon the results of our first theorem, and indeed, such an improvement is achieved by a second method of expurgation. Here, one starts by expurgating cloud centers, and only afterwards, single codewords. The intuition behind this technique is the following. When the exponential rate of the codewords within a cloud is too high, the weak user can still make a good estimation, merely by relying on the set of cloud centers. The expurgated bounds of our second method, however, are not always tighter than those of the first method, because of other differences in their derivations.

We then expand the scope and consider the generalized likelihood decoder (GLD), which is a more general family of stochastic likelihood decoders. For such decoders, the probability of deciding on a given message is proportional to a general exponential function of the joint empirical distribution of the cloud–center, the codeword and the received channel output vector. The random coding error exponent of the ordinary and the mismatched likelihood decoders for the single–user channel have been derived by Scarlett *et al.* [7]. In a more recent paper by Merhav [8], the expurgated exponent of the GLD has been derived and compared to the classical expurgated bound of [6], showing an explicit improvement at relatively high coding rates. In this paper, we consider GLD's for both the strong and the weak users of an ABC, and derive

* This work was done while N. Weinberger was at the Technion. Currently, he is with the School of Electrical Engineering at Tel–Aviv University.

expurgated exponents under these decoders. These bounds generalize the bound of [8], and prove that they are at least as tight as the maximum between the random coding error exponents of [5] and the expurgated bounds of our first theorem, which are based on the ML decoder. By that, we actually prove the existence of hierarchical codebooks that attain the best of the random coding exponent and the expurgated exponent simultaneously for both users. The main drawback of those error exponents is that they are not easy to calculate since they involve minimizations over relatively cumbersome auxiliary channels, and hence, efficient computation algorithms for the GLD bound are left for further research. From this viewpoint, the exponents of our first theorems are much more attractive.

Due to the space limitation, technical details and proofs are omitted, but can be found in the full version of this paper [11].

II. Notation Conventions

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Sources and channels will be subscripted by the names of the relevant random variables/vectors and their conditionings, whenever applicable, following the standard notation conventions, e.g., Q_X, $Q_{Y|X}$, and so on. When there is no room for ambiguity, these subscripts will be omitted. For a generic joint distribution $Q_{XY} = \{Q_{XY}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$, which will often be abbreviated by Q, information measures will be denoted in the conventional manner, but with a subscript Q, that is, $H_Q(X)$ is the marginal entropy of X, $I_Q(X;Y)$ is the mutual information between X and Y, and so on. The weighted divergence between two conditional distributions (channels), say, $Q_{Z|X}$ and $W = \{W(z|x), x \in \mathcal{X}, z \in \mathcal{Z}\}$, with weighting Q_X is defined as

$$D(Q_{Z|X}||W|Q_X) = \sum\nolimits_{x\in\mathcal{X}} Q_X(x) \sum\nolimits_{z\in\mathcal{Z}} Q_{Z|X}(z|x) \log \frac{Q_{Z|X}(z|x)}{W(z|x)}, \quad (1)$$

where logarithms, here and throughout the sequel, are taken to the natural base. The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$, and the expectation operator with respect to a probability distribution Q will be denoted by $\mathbb{E}_Q\{\cdot\}$. The notation $[x]_+$ will stand for $\max\{0, x\}$.

The type class of Q_U, denoted by $\mathcal{T}(Q_U)$, is the set of all vectors $\boldsymbol{u} \in \mathcal{U}^n$ with $\hat{P}_{\boldsymbol{u}} = Q_U$, where $\hat{P}_{\boldsymbol{u}}$ is the empirical distribution of the sequence $\boldsymbol{u}$. Similarly, $\mathcal{T}(Q_{X|Y}|\boldsymbol{y})$ denotes the conditional type class, induced by the sequence $\boldsymbol{y}$ and the empirical conditional distribution $Q_{X|Y}$.

III. Definitions and Problem Formulation

We consider a memoryless ABC with a finite input alphabet $\mathcal{X}$ and finite output alphabets $\mathcal{Y}$ and $\mathcal{Z}$. Let $W_1 = \{W_1(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ and $W_2 = \{W_2(z|x),\ x \in \mathcal{X},\ z \in \mathcal{Z}\}$ denote the single–letter input–output transition probability matrices, associated with the strong user and the weak user, respectively. When these channels are fed by an input vector $\boldsymbol{x} \in \mathcal{X}^n$, they produce the corresponding output vectors $\boldsymbol{y} \in \mathcal{Y}^n$ and $\boldsymbol{z} \in \mathcal{Z}^n$, according to $W_1(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^n W_1(y_t|x_t)$ and $W_2(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^n W_2(z_t|x_t)$. We are interested in sending one out of M_z *common* messages to both users, and one out of M_y *private* messages to the strong user, that observes $\boldsymbol{y}$. The two messages are chosen under the uniform distribution. Although our results prove the existence of a single sequence of *deterministic* hierarchical constant composition (HCC) codebooks, whose error probabilities are provably bounded, our proof techniques use extensively the following mechanism of random selection of an HCC code for the ABC. Let $\mathcal{U}$ be a finite alphabet, let P_U be a given probability distribution on $\mathcal{U}$, and let $P_{X|U}$ be a given matrix of conditional probabilities of X given U, such that the type–class $\mathcal{T}(P_U)$ and the conditional type–class $\mathcal{T}(P_{X|U}|\boldsymbol{u})$ are non–empty. We first select, independently at random, $M_z = \lceil e^{nR_z} \rceil$ n-vectors ("cloud centers"), $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{M_z-1}$, all under the uniform distribution over the type–class $\mathcal{T}(P_U)$. Next, for each $m = 0, 1, \ldots, M_z - 1$, we select conditionally independently (given $\boldsymbol{u}_m$), $M_y = \lceil e^{nR_y} \rceil$ codewords, $\boldsymbol{x}_{m,0}, \boldsymbol{x}_{m,1}, \ldots, \boldsymbol{x}_{m,(M_y-1)}$, under the uniform distribution across the conditional type–class $\mathcal{T}(P_{X|U}|\boldsymbol{u}_m)$. We denote the sub–code for each cloud by $\mathcal{C}_m(n) = \{\boldsymbol{x}_{m,0}, \boldsymbol{x}_{m,1}, \ldots, \boldsymbol{x}_{m,(M_y-1)}\}$, or just $\mathcal{C}_m$ in short. Thus, the communication rate to the weak user is R_z, while the total communication rate to the strong user is $R_z + R_y$. Once selected, the entire codebook $\mathcal{C}(n) = \cup_{m=0}^{M_z-1} \mathcal{C}_m(n)$, and the collection of cloud centers, $\{\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{M_z-1}\}$, are revealed to the encoder and to both decoders. We denote by $\mathscr{C}$ a sequence of HCC codes, $\{\mathcal{C}(n),\ n = 1, 2, \ldots\}$.

For any of the following described decoding rules, denote by $[\hat{m}(\boldsymbol{y}), \hat{i}(\boldsymbol{y})]$ the decoded pair of the strong user, and by $\tilde{m}(\boldsymbol{z})$ the decoded cloud of the weak user. The ML decoder for the strong user is given by

$$[\hat{m}(\boldsymbol{y}), \hat{i}(\boldsymbol{y})] = \operatorname*{arg\,max}_{0\le m\le M_z-1, 0\le i\le M_y-1} W_1(\boldsymbol{y}|\boldsymbol{x}_{mi}), \quad (2)$$

and the optimal ML decoder for the weak user is given by

$$\tilde{m}(\boldsymbol{z}) = \operatorname*{arg\,max}_{0\le m\le M_z-1} \left\{ \frac{1}{M_y} \sum_{\boldsymbol{x}\in\mathcal{C}_m} W_2(\boldsymbol{z}|\boldsymbol{x}) \right\}. \quad (3)$$

The likelihood decoder is a stochastic decoder, that chooses the decoded message according to the posterior probability mass function, induced by the channel output (either $\boldsymbol{y}$ or $\boldsymbol{z}$). For the strong user, the ordinary likelihood decoder randomly selects the estimated message $(\hat{m}, \hat{i})$ according to the following posterior distribution

$$P(m, i|\boldsymbol{y}) = \frac{W_1(\boldsymbol{y}|\boldsymbol{x}_{mi})}{\sum_{m'=0}^{M_z-1} \sum_{i'=0}^{M_y-1} W_1(\boldsymbol{y}|\boldsymbol{x}_{m'i'})}.$$

The generalized likelihood decoder (GLD) for the strong user is defined by the conditional probability

$$P(m, i|\boldsymbol{y}) = \frac{\exp\{ng(\hat{P}_{\boldsymbol{u}_m \boldsymbol{x}_{mi} \boldsymbol{y}})\}}{\sum_{m'=0}^{M_z-1} \sum_{i'=0}^{M_y-1} \exp\{ng(\hat{P}_{\boldsymbol{u}_{m'} \boldsymbol{x}_{m'i'} \boldsymbol{y}})\}},$$

where $\hat{P}_{\boldsymbol{u}_m \boldsymbol{x}_{mi} \boldsymbol{y}}$ is the empirical distribution of $(\boldsymbol{u}_m, \boldsymbol{x}_{mi}, \boldsymbol{y})$, and $g(\cdot)$ is a given continuous, real valued functional of this empirical distribution. In the same manner, the ordinary likelihood decoder for the weak user randomly selects the estimated cloud $\tilde{m}$ according to

$$P\left(m|\boldsymbol{z}\right) = \frac{\sum_{i=0}^{M_y-1} W_2(\boldsymbol{z}|\boldsymbol{x}_{mi})}{\sum_{m'=0}^{M_z-1}\sum_{i'=0}^{M_y-1} W_2(\boldsymbol{z}|\boldsymbol{x}_{m'i'})},$$

while the GLD for the weak user is defined by

$$P\left(m|\boldsymbol{z}\right) = \frac{\sum_{i=0}^{M_y-1} \exp\{ng(\hat{P}_{\boldsymbol{u}_m \boldsymbol{x}_{mi} \boldsymbol{z}})\}}{\sum_{m'=0}^{M_z-1}\sum_{i'=0}^{M_y-1} \exp\{ng(\hat{P}_{\boldsymbol{u}_{m'} \boldsymbol{x}_{m'i'} \boldsymbol{z}})\}}.$$

Exactly as the universal decoders derived in [5], generalized decoders may also depend on the cloud–centers, which may be helpful, since all of the codewords in each sub–code are highly correlated via their cloud–center. One of the most important properties of the GLD is the following. The union bound, which is used in the first steps of the derivations for both users, actually provides an exact expression for the probability of error, unlike in the analyses of the ML decoders, where the union bound harms the exponential tightness, at least for relatively high rates. The generalized likelihood decoders cover several important special cases, such as the ordinary likelihood decoder and the mismatched likelihood decoder. For the strong user, the deterministic ML and the maximum mutual information decoders [9] can be obtained from the GLD by limiting operations.

Let $\boldsymbol{Y} \in \mathcal{Y}^n$ and $\boldsymbol{Z} \in \mathcal{Z}^n$ be the random channel outputs resulting from the transmission of $\boldsymbol{x}_{mi}$. For a given code $\mathcal{C}(n)$, define the error probabilities as

$$P_{\mathrm{e}|mi}(\mathcal{C}(n)) = \Pr\left\{[\hat{m}(\boldsymbol{Y}), \hat{i}(\boldsymbol{Y})] \neq (m,i) \middle| \boldsymbol{x}_{mi} \text{ sent}\right\}, \quad (4)$$

and

$$P_{\mathrm{e}|m}(\mathcal{C}(n)) = \frac{1}{M_y}\sum_{i=0}^{M_y-1} \Pr\left\{\tilde{m}(\boldsymbol{Z}) \neq m | \boldsymbol{x}_{mi} \text{ sent}\right\}, \quad (5)$$

where in both definitions, $\Pr\{\cdot\}$ designates the probability measure associated with the randomness of the channel outputs given its input, and the randomness of the stochastic decoders. Moreover, the error probabilities are defined to be zero whenever the blocklength is such that no code can be generated. Our main objective is to prove the existence of sequences of HCC codes $\mathscr{C} = \{\mathcal{C}(n)\}_{n=1}^{\infty}$ and obtain the tightest possible single–letter expressions that lower bound the following limits

$$E_{\mathrm{su}}(\mathscr{C}) \triangleq \liminf_{n\to\infty} \left[-\tfrac{1}{n}\log \max_{m,i} P_{\mathrm{e}|mi}(\mathcal{C}(n))\right], \quad (6)$$

$$E_{\mathrm{wu}}(\mathscr{C}) \triangleq \liminf_{n\to\infty} \left[-\tfrac{1}{n}\log \max_{m} P_{\mathrm{e}|m}(\mathcal{C}(n))\right], \quad (7)$$

both for the ML decoder and the GLD.

In a recent paper [5], exact random coding error exponents have been derived for both users of the ABC. We may expect to improve these error exponents, at least when one of the coding rates is low, by code expurgation. In this paper, we derive expurgated exponents for the ABC under ML decoding in two different methods. In addition, we discuss the GLD, that enables us to achieve the best between the random coding bound and one of the ML–based expurgated bounds.

IV. Main Results

A. Maximum Likelihood Decoding

For maximum likelihood decoding, we distinguish between two different methods of expurgation for the HCC ensemble. The first method is based on the following technique of expurgation: we randomly draw a HCC codebook, and then simultaneously expurgate both bad clouds and bad codewords within the remaining clouds. The resulting expurgated bounds are given in Theorem 1. In order to state our first theorem, we start with the following definitions. We define the sets $\mathcal{S} \triangleq \{Q_{UXX'} : Q_{UX'} = Q_{UX} = P_{UX}\}$ and $\mathcal{P} \triangleq \{Q_{UU'XX'} : Q_{U'X'} = Q_{UX} = P_{UX}\}$, and the averaged Chernoff distance function by

$$D_s(Q_{XX'}) \triangleq -\mathbb{E}_Q \log\left[\sum\nolimits_{y\in\mathcal{Y}} W^{1-s}(y|X)\cdot W^s(y|X')\right].$$

For the weak user, define an error exponent function as

$$E_{\mathrm{wu}}^{\mathrm{ML1}}(R_y,R_z) \triangleq \max_{0\leq t\leq 1} \min_{\substack{Q_{UU'XX'}\in\mathcal{P}\\ I_Q(UX;U'X')\leq 2R_y+R_z}} [I_Q(UX;U'X') + D_t(Q_{XX'})] - R_y - R_z. \quad (8)$$

Next, for the strong user we define the following error exponent functions

$$E_{\mathrm{su\text{-}1}}^{\mathrm{ML1}}(R_y,s) \triangleq \min_{\{Q_{UXX'}\in\mathcal{S}:\ I_Q(X;X'|U)\leq R_y\}} [I_Q(X;X'|U) + D_s(Q_{XX'})] - R_y, \quad (9)$$

$$E_{\mathrm{su\text{-}2}}^{\mathrm{ML1}}(R_y,R_z,s) \triangleq \min_{\substack{Q_{UU'XX'}\in\mathcal{P}\\ I_Q(UX;U'X')\leq R_y+R_z}} [I_Q(UX;U'X') + D_s(Q_{XX'})] - R_y - R_z, \quad (10)$$

$$E_{\mathrm{su}}^{\mathrm{ML1}}(R_y,R_z) \triangleq \max_{0\leq s\leq 1} \min\left\{E_{\mathrm{su\text{-}1}}^{\mathrm{ML1}}(R_y,s), E_{\mathrm{su\text{-}2}}^{\mathrm{ML1}}(R_y,R_z,s)\right\}.$$

Theorem 1. There exists a sequence $\mathscr{C}$ of HCC codes, with a rate pair (R_y, R_z) for which both

$$E_{\mathrm{su}}(\mathscr{C}) \geq E_{\mathrm{su}}^{\mathrm{ML1}}(R_y,R_z) \text{ and } E_{\mathrm{wu}}(\mathscr{C}) \geq E_{\mathrm{wu}}^{\mathrm{ML1}}(R_y,R_z). \quad (11)$$

The second method is somewhat different, and the idea behind it is the following. At the first step, we expurgate sub-codes, merely according to their cloud–centers. Then, at the second step, we fix the set of cloud–centers of the remaining clouds from the first step, and then expurgate specific codewords, as well as clouds, according to some collective behavior of their codewords. The resulting expurgated bounds are given in Theorem 2, and as can be seen below, the expressions are more complicated than those of Theorem 1, at least for the weak user.

In order to state our second theorem, we need a few definitions. For a given marginal Q_{UZ}, let $\mathcal{S}(Q_{UZ})$ denote the set of conditional distributions $\{Q_{X|UZ}\}$ such that $\sum_z Q_{UZ}(u,z)Q_{X|UZ}(x|u,z) = P_{UX}(u,x)$ for every

$(u,x) \in \mathcal{U} \times \mathcal{X}$, where $P_{UX} = P_U \times P_{X|U}$. We denote $\bar{t} = 1-t$. For the weak user, define

$$\begin{aligned}\hat{D}_t(R_y, Q_{UU'}) &\triangleq \min_{Q_{Z|UU'}} \min_{Q_{X|UZ} \in \mathcal{S}(Q_{UZ})} \min_{Q_{X'|U'Z} \in \mathcal{S}(Q_{U'Z})} \\ &\{\bar{t} \cdot D(Q_{Z|UX} \| W_{Z|X} | Q_{UX}) + \bar{t} \cdot I_Q(Z; U'|U) \\ &+ t \cdot D(Q_{Z|U'X'} \| W_{Z|X'} | Q_{U'X'}) + t \cdot I_Q(Z; U|U') \\ &+ t \cdot [I_Q(X; Z|U) - R_y]_+ + \bar{t} \cdot [I_Q(X'; Z|U') - R_y]_+\}.\end{aligned}$$

We define the set $\mathcal{Q} \triangleq \{Q_{UU'} : \ Q_U = Q_{U'} = P_U\}$ and an error exponent function

$$\begin{aligned}E_{\text{wu}}^{\text{ML2}}(R_y, R_z) \triangleq \max_{0 \le t \le 1} \min_{\{Q_{UU'} \in \mathcal{Q}: \ I_Q(U;U') \le R_z\}} &[I_Q(U; U') \\ &+ \hat{D}_t(R_y, Q_{UU'})] - R_z. \qquad (12)\end{aligned}$$

Next, for the strong user we define the following error exponent functions

$$\begin{aligned}E_{\text{su-1}}^{\text{ML2}}(R_y, s) &\triangleq \min_{\{Q_{UXX'} \in \mathcal{S}: \ I_Q(X; X'|U) \le R_y\}} [I_Q(X; X'|U) \\ &\qquad + D_s(Q_{XX'})] - R_y, \qquad (13) \\ E_{\text{su-2}}^{\text{ML2}}(R_y, R_z, s) &\triangleq \min_{\{Q_{UU'XX'} \in \mathcal{P}: \ I_Q(U;U') \le R_z\}} [I_Q(UX; U'X') \\ &\qquad + D_s(Q_{XX'})] - R_y - R_z, \qquad (14) \\ E_{\text{su}}^{\text{ML2}}(R_y, R_z) &\triangleq \max_{0 \le s \le 1} \min\left\{E_{\text{su-1}}^{\text{ML2}}(R_y, s), E_{\text{su-2}}^{\text{ML2}}(R_y, R_z, s)\right\}.\end{aligned}$$

Theorem 2. There exists a sequence $\mathscr{C}$ of HCC codes, with a rate pair (R_y, R_z) for which both

$$E_{\text{su}}(\mathscr{C}) \ge E_{\text{su}}^{\text{ML2}}(R_y, R_z) \text{ and } E_{\text{wu}}(\mathscr{C}) \ge E_{\text{wu}}^{\text{ML2}}(R_y, R_z). \quad (15)$$

Discussion: First, all of the expressions in Theorems 1 and 2 generalize the well–known CKM expurgated bound [6]. For example, it can be easily recovered from $E_{\text{wu}}^{\text{ML1}}(R_y, R_z)$, when degenerating the hierarchical codebook by choosing $R_y = 0$, as well as $P_{X|U}(x|u) = \delta(x-u)$ ($\mathcal{X} = \mathcal{U}$).

Concerning the strong user, each bound is given by the minimum between two different expressions. The first expression is related to error events within the cloud of the true codeword. In fact, we have that $E_{\text{su-1}}^{\text{ML1}}(R_y, s) = E_{\text{su-1}}^{\text{ML2}}(R_y, s)$, where the difference is given by the second components, $E_{\text{su-2}}^{\text{ML1}}(R_y, R_z, s)$ and $E_{\text{su-2}}^{\text{ML2}}(R_y, R_z, s)$, for which the expurgation method cause a change in the final expressions. Although the objectives in (10) and (14) are the same, the constraints are different, and are not subsets of each other.

Concerning the weak user, the situation is much more complicated, because of the structure of the optimal decoder. The derivation in the proof of Theorem 1 contains the following inequality that may harm the tightness of the bound:

$$\begin{aligned}&\left[\sum\nolimits_{x \in \mathcal{C}} W_2(\boldsymbol{z}|\boldsymbol{x})\right]^{1-t} \cdot \left[\sum\nolimits_{x' \in \mathcal{C}'} W_2(\boldsymbol{z}|\boldsymbol{x}')\right]^t \\ &\le \sum\nolimits_{x \in \mathcal{C}} \sum\nolimits_{x' \in \mathcal{C}'} W_2^{1-t}(\boldsymbol{z}|\boldsymbol{x}) \cdot W_2^t(\boldsymbol{z}|\boldsymbol{x}'). \qquad (16)\end{aligned}$$

Because of this passage, the bound of Theorem 1 is inferior to the bound of Theorem 2, at relatively high values of R_y. Specifically, the expression given in Theorem 2 reaches a plateau at high R_y, while the expression of Theorem 1 reaches zero. In this regime, there is no loss in the exponent of the weak user if its decoder treats the satellites codewords as noise. In this event, the satellite-rate is immaterial, and the exponent of the weak user only depends on R_z. One should note that the improvement at high rates is obtained by expressions which are more complicated to compute. However, the resulting exponent of Theorem 1 still outperforms the result of Theorem 2, at least for relatively low R_y values (see Fig. 1).

We next provide some numerical results (Fig. 1), comparing our expurgated bounds for the weak user, as given by Theorems 1 and 2. Let W_1 and W_2 be two binary symmetric channels (BSC) with crossover parameters $p_y = 0.0005$ and $p_z = 0.001$, respectively. Let $\mathcal{U}$ be binary as well and let P_U be uniformly distributed over $\{0, 1\}$. Also, let $P_{X|U}$ be a BSC with crossover parameter $p_{x|u} = 0.15$.

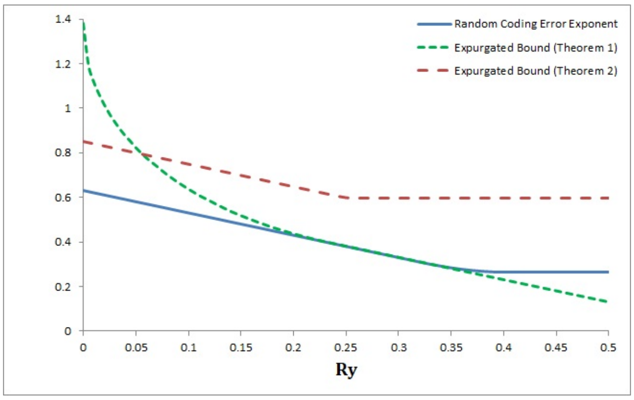


Fig. 1. Expurgated bounds for the weak user ($R_z = 0$).

B. Generalized Stochastic Likelihood Decoding

As was already mentioned earlier, the GLD enables us to make a tighter derivation for the probability of error, and therefore, the resulting expurgated bounds are strictly tighter, at least at relatively high rates. The drawback of the expressions of Theorem 3 is that they are quite cumbersome, at least when compared to those of Theorems 1 or 2. In order to characterize the expurgated bounds of the GLD, we define first a few quantities. Let

$$\begin{aligned}&\phi(R_y, Q_{UY}) \\ &\triangleq \max_{\{Q_{X|UY}: \ I_Q(X;Y|U) \le R_y\}} [g(Q) - I_Q(X; Y|U)] + R_y, \\ &\psi(R_y, R_z, Q_Y) \\ &\triangleq \max_{\substack{Q_{UX|Y}: \ I_Q(U;Y) \le R_z \\ I_Q(UX;Y) \le R_z + R_y}} [g(Q) - I_Q(UX; Y)] + R_z + R_y.\end{aligned}$$

Also, define

$$\begin{aligned}&\Upsilon(Q_{UXX'}, R_y, R_z) \triangleq \min_{Q_{Y|UXX'}} \Big(D(Q_{Y|UX} \| W_{Y|X} | Q_{UX}) \\ &+ I_Q(X'; Y|UX) + [\max\{g(Q_{UXY}), \phi(R_y, Q_{UY}), \\ &\psi(R_y, R_z, Q_Y)\} - g(Q_{UX'Y})]_+\Big), \qquad (17)\end{aligned}$$

$$\Omega(Q_{UU'XX'}, R_y, R_z) \triangleq \min_{Q_{Y|UU'XX'}} \Big(D(Q_{Y|UX} \| W_{Y|X} | Q_{UX}) + I_Q(U'X'; Y|UX) + [\max\{g(Q_{UXY}), \phi(R_y, Q_{UY}), \psi(R_y, R_z, Q_Y)\} - g(Q_{U'X'Y})]_+ \Big). \tag{18}$$

We define the following error exponent functions. For the weak user,

$$E^{\text{GLD}}_{\text{wu}}(R_y, R_z) \triangleq \min_{\substack{Q_{UU'XX'} \in \mathcal{P}:\ I_Q(U;U') < R_z \\ I_Q(UX;U'X') < 2R_y + R_z}} [I_Q(UX; U'X') + \Omega(Q, R_y, R_z)] - R_y - R_z, \tag{19}$$

and for the strong user

$$E^{\text{GLD}}_{\text{su-1}}(R_y, R_z) \triangleq \min_{\{Q_{UXX'} \in \mathcal{S}:\ I_Q(X;X'|U) < R_y\}} [I_Q(X; X'|U) + \Upsilon(Q, R_y, R_z)] - R_y, \tag{20}$$

$$E^{\text{GLD}}_{\text{su-2}}(R_y, R_z) \triangleq \min_{\substack{Q_{UU'XX'} \in \mathcal{P}:\ I_Q(UX;U') < R_z \\ I_Q(UX;U'X') < R_y + R_z}} [I_Q(UX; U'X') + \Omega(Q, R_y, R_z)] - R_y - R_z, \tag{21}$$

$$E^{\text{GLD}}_{\text{su}}(R_y, R_z) \triangleq \min\{E^{\text{GLD}}_{\text{su-1}}(R_y, R_z), E^{\text{GLD}}_{\text{su-2}}(R_y, R_z)\}. \tag{22}$$

Theorem 3. There exists a sequence $\mathscr{C}$ of HCC codes, with a rate pair (R_y, R_z) for which both

$$E_{\text{su}}(\mathscr{C}) \geq E^{\text{GLD}}_{\text{su}}(R_y, R_z) \text{ and } E_{\text{wu}}(\mathscr{C}) \geq E^{\text{GLD}}_{\text{wu}}(R_y, R_z). \tag{23}$$

Discussion

• An expurgated bound for the GLD in the single user regime has been derived by Merhav [8]. It should be noticed that the resulting expressions of Theorem 3, as well as some parts of its proof (in [11, Section 7]) are nontrivial generalizations of the single–user case.

• The expression of (19) has the same structure as (8), except that here the functional $\Omega(Q, R_y, R_z)$ replaces the expected Chernoff distance, and an additional constraint ($I_Q(U; U') < R_z$) has been added. We prove in [11, Appendix A] that at least for the choice $g(Q) = \mathbb{E}_Q \log W_2(Z|X)$, $E^{\text{GLD}}_{\text{wu}}(R_y, R_z)$ is at least as tight as $E^{\text{ML1}}_{\text{wu}}(R_y, R_z)$.

• One of the main advantages of the GLD, is the fact that the derivation of its probability of error may be exponentially tighter than the derivations in the proofs of Theorems 1 or 2. As a consequence, we show in [11, Appendix B] that $E^{\text{GLD}}_{\text{wu}}(R_y, R_z)$ cannot be smaller than the random coding error exponent of the weak user at any pair of rates, by examining the former for the suboptimal universal metric $g(Q) = I_Q(UX; Z)$. We conclude that $E^{\text{GLD}}_{\text{wu}}(R_y, R_z)$ is at least as tight as the maximum between $E^{\text{ML1}}_{\text{wu}}(R_y, R_z)$ and the random coding exponent, $E^{\text{RC}}_{\text{wu}}(R_y, R_z)$.

• The same can be proved for the strong user, i.e., that $E^{\text{GLD}}_{\text{su}}(R_y, R_z)$ is at least as tight as the maximum between $E^{\text{ML1}}_{\text{su}}(R_y, R_z)$ and the random coding error exponent, $E^{\text{RC}}_{\text{su}}(R_y, R_z)$. We conclude, that there exist a HCC codebook, for which one user works in the "expurgated region", while the other user works in the "random coding region". For example, it may be the case when the channel to the strong user is quite clean, while the channel to the weak user is very noisy.

• We were not able to determine whether the bound of Theorem 3 is at least as tight as the maximum between the bounds of the first two theorems, although we conjecture that it is indeed the case when choosing one of the decoding metrics $g(Q) = \beta \mathbb{E}_Q \log W_2(Z|X)$ or $g(Q) = \beta I_Q(UX; Z)$, and letting $\beta \to \infty$.

V. Proof Sketch of Theorem 1

We start by proving that for any $s, t \in [0, 1]$

$$P_{\text{e}|mi}(\mathcal{C}(n)) \leq \sum_{Q \in \mathcal{S}} N^{\text{IN}}_{mi}(Q, \mathcal{C}(n)) \cdot e^{-nD_s(Q_{XX'})} + \sum_{Q \in \mathcal{P}} N^{\text{OUT}}_{mi}(Q, \mathcal{C}(n)) \cdot e^{-nD_s(Q_{XX'})}, \tag{24}$$

$$P_{\text{e}|m}(\mathcal{C}(n)) \leq \frac{1}{M_y} \sum_{Q \in \mathcal{P}} \hat{N}_m(Q, \mathcal{C}(n)) \cdot e^{-nD_t(Q_{XX'})}, \tag{25}$$

where $N^{\text{IN}}_{mi}(Q, \mathcal{C}(n))$, $N^{\text{OUT}}_{mi}(Q, \mathcal{C}(n))$ and $\hat{N}_m(Q, \mathcal{C}(n))$ are suitable type–class enumerators[1]. By using the method of types and Markov's inequality, it is proved that for every $\epsilon > 0$ and all sufficiently large n, there exists a code $\mathcal{C}(n)$ with a rate pair (R_y, R_z), that satisfies, for every (m, i) and every Q,

$$N^{\text{IN}}_{mi}(Q, \mathcal{C}(n)) \leq N^{\text{IN}}_{\epsilon}(Q), \tag{26a}$$

$$N^{\text{OUT}}_{mi}(Q, \mathcal{C}(n)) \leq N^{\text{OUT}}_{\epsilon}(Q), \tag{26b}$$

$$\hat{N}_m(Q, \mathcal{C}(n)) \leq \hat{N}_{\epsilon}(Q). \tag{26c}$$

Upon substituting these *deterministic* upper bounds back into (24) and (25), we conclude the bounds given in Theorem 1.

References

[1] T. M. Cover, "Broadcast channels," *IEEE Trans. on Inform. Theory*, vol. 18, no. 1, pp. 2–14, January 1972.

[2] J. Körner and K. Marton, "General broadcast channels with degraded message sets," *IEEE Trans. on Inform. Theory*, vol. IT-23, pp. 60-64, January 1977.

[3] J. Körner and A. Sgarro, "Universally attainable error exponents for broadcast channels with degraded message sets," *IEEE Trans. on Inform. Theory*, vol. 26, no.6, pp. 670-679, November 1980.

[4] Y. Kaspi and N. Merhav, "Error exponents for broadcast channels with degraded message sets," *IEEE Trans. on Inform. Theory*, vol. 57, no. 1, pp. 101–123, January 2011.

[5] R. Averbuch and N. Merhav, "Exact random coding exponents and universal decoders for the asymmetric broadcast channel," submitted for publication, February 2017. Available on–line at `https://arxiv.org/pdf/1702.08003.pdf`

[6] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," in *Proc. ISIT '77*, p. 107 (abstract), Cornell University, Ithaca, New York, 1977.

[7] J. Scarlett, A. Martinèz, and A. Guillén i Fàbregas, "The likelihood decoder: error exponents and mismatch," *Proc. 2015 IEEE International Symposium on Information Theory (ISIT 2015)*, pp. 86–90, Hong Kong, June 2015.

[8] N. Merhav, "The generalized stochastic likelihood decoder: random coding and expurgated bounds," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 5039–5051, August 2017. See also a correction at *IEEE Trans. on Inform. Theory*, vol. 63, no. 10, pp. 6827–6829, Oct. 2017.

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

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

[11] R. Averbuch, N. Weinberger, and N. Merhav, "Expurgated bounds for the asymmetric broadcast channel," available on–line at `https://arxiv.org/pdf/1711.10299.pdf`

[1] Meaning, the number of codewords that fall in a given type–class.

The Arbitrarily Varying Broadcast Channel with Degraded Message Sets with Causal Side Information at the Encoder

Uzi Pereg and Yossef Steinberg
Department of Electrical Engineering
Technion, Haifa 32000, Israel.
Email: uzipereg@campus.technion.ac.il, ysteinbe@ee.technion.ac.il

***Abstract*—In this work, we study the arbitrarily varying broadcast channel (AVBC), when state information is available at the transmitter in a causal manner. We establish inner and outer bounds on both the random code capacity region and the deterministic code capacity region with degraded message sets. The capacity region is then determined for a class of channels satisfying a condition on the mutual informations between the strategy variables and the channel outputs. As an example, we consider the arbitrarily varying binary symmetric broadcast channel with correlated noises. We show cases where the condition holds, hence the capacity region is determined, and other cases where there is a gap between the bounds.**

***Index Terms*—Arbitrarily varying channel, broadcast channel, degraded message sets, causal state information, Shannon strategies, side information, minimax theorem, deterministic code, random code, symmetrizability.**

I. Introduction

The arbitrarily varying channel (AVC) was first introduced by Blackwell *et al.* [3] to describe a communication channel with unknown statistics, that may change over time. It is often described as communication in the presence of an adversary, or a *jammer*, attempting to disrupt communication.

The arbitrarily varying broadcast channel (AVBC) without side information (SI) was first considered by Jahn [8], who derived an inner bound on the random code capacity region, namely the capacity region achieved by encoder and decoders with a random experiment, shared between the three parties. As indicated by Jahn, the arbitrarily varying broadcast channel inherits some of the properties of its single user counterpart. In particular, the random code capacity region is not necessarily achievable using deterministic codes [3]. Furthermore, Jahn showed that the deterministic code capacity region either coincides with the random code capacity region or else, it has an empty interior [8]. This phenomenon is an analogue of Ahlswede's dichotomy property [1]. Then, in order to apply Jahn's inner bound, one has to verify whether the capacity region has non-empty interior or not. As observed in [7], this can be resolved using the results of Ericson [6] and Csiszár and Narayan [5]. Specifically, a necessary and sufficient condition for the capacity region to have a non-empty interior is that both users marginal channels are non-symmetrizable.

Various models of interest involve SI available at the encoder. In [12], the arbitrarily varying degraded broadcast channel with non-causal SI is addressed, using Ahlswede's Robustification and Elimination Techniques [2]. The single user AVC with causal SI was addressed by Csiszár and Körner [4], while their approach is independent of Ahlswede's work. A straightforward application of Ahlswede's Robustification Technique (RT) would violate the causality requirement.

In this work, we study the AVBC with causal SI available at the encoder. We extend Ahlswede's Robustification and Elimination Techniques [1, 2], originally used in the setting of *non*-causal SI. In particular, we derive a modified version of Ahlswede's RT for the setting of causal SI. In a recent paper by the authors [10], a similar proof technique is applied to the arbitrarily varying *degraded* broadcast channel with causal SI. Here, we generalize those results, and consider a *general* broadcast channel with degraded message sets with causal SI.

We establish inner and outer bounds on the random code and deterministic code capacity regions. Furthermore, we give conditions on the AVBC under which the bounds coincide, and the capacity region is determined. As an example, we consider the arbitrarily varying binary symmetric broadcast channel with correlated noises. We show that in some cases, the conditions hold and the capacity region is determined. Whereas, in other cases, there is a gap between the bounds. A full manuscript with proofs can be found in [9].

II. Definitions and Previous Results

A. Channel Description

A state-dependent broadcast channel (BC) consists of finite input, state and outputs alphabets $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{Y}_1 \times \mathcal{Y}_2$, resepectively, and a collection of probability functions $W_{Y_1,Y_2|X,S}$. The channel is memoryless without feedback, and therefore

$$W_{Y_1^n,Y_2^n|X^n,S^n}(y_1^n,y_2^n|x^n,s^n) = \prod_{i=1}^{n} W_{Y_1,Y_2|X,S}(y_{1,i},y_{2,i}|x_i,s_i)\ . \quad (1)$$

The marginals $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ correspond to User 1 and User 2 respectively. When causal SI is available, the channel

This research was supported by the Israel Science Foundation (grant No. 1285/16).

input at time i may depend on the sequence of past and present states s^i. The AVBC is a BC $W_{Y_1,Y_2|X,S}$ with a state sequence $S^n \sim q(s^n)$, with an unknown joint probability mass function (pmf) $q(s^n)$, not necessarily independent nor stationary. We denote the AVBC with causal SI by $\mathcal{B} = \{W_{Y_1,Y_2|X,S}\}$.

To analyze the AVBC, we consider the *compound BC*, defined as a BC with a discrete memoryless state, where the state distribution $q(s)$ belongs to a given set $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ is the space of all state distributions. We denote the compound BC with causal SI by $\mathcal{B}^{\mathcal{Q}}$.

B. Coding

Deterministic and random codes are defined below, where a deterministic code is also referred to simply as 'code'.

Definition 1. A $(2^{nR_0}, 2^{nR_1}, n)$ code for the AVBC $\mathcal{B}$ with degraded message sets with causal SI consists of the following; two message sets $[1:2^{nR_0}]$ and $[1:2^{nR_1}]$, assuming 2^{nR_0} and 2^{nR_1} are integers, a sequence of n encoding functions $f_i : [1:2^{nR_0}] \times [1:2^{nR_1}] \times \mathcal{S}^i \to \mathcal{X}$, $i \in [1:n]$, and two decoding functions, $g_1 : \mathcal{Y}_1^n \to [1:2^{nR_0}] \times [1:2^{nR_1}]$ and $g_2 : \mathcal{Y}_2^n \to [1:2^{nR_0}]$. At time $i \in [1:n]$, given a pair of messages $(m_0, m_1) \in [1:2^{nR_0}] \times [1:2^{nR_1}]$ and a sequence s^i, the encoder sends $x_i = f_i(m_0, m_1, s^i)$. Decoder 1 receives y_1^n, and finds an estimate for the message pair $(\hat{m}_0, \hat{m}_1) = g_1(y_1^n)$. Decoder 2 only estimates the common message with $\widetilde{m}_0 = g_2(y_2^n)$. We denote the code by $\mathscr{C} = (f^n, g_1, g_2)$.

Define the conditional probability of error given $s^n \in \mathcal{S}^n$,

$$P_{e|s^n}^{(n)}(\mathscr{C}) = \frac{1}{2^{n(R_0+R_1)}} \sum_{m_0=1}^{2^{nR_0}} \sum_{m_1=1}^{2^{nR_1}} \sum_{\mathcal{D}(m_0,m_1)^c} W_{Y_1^n,Y_2^n|X^n,S^n}(y_1^n, y_2^n | f^n(m_0, m_1, s^n), s^n)\,, \quad (2)$$

where $\mathcal{D}(m_0, m_1) \triangleq \big\{ (y_1^n, y_2^n) : g_1(y_1^n) = (m_0, m_1), g_2(y_2^n) = m_0 \big\}$. Now, define the average probability of error for a distribution $q(s^n)$, by $P_e^{(n)}(q, \mathscr{C}) = \sum_{s^n \in \mathcal{S}^n} q(s^n) P_{e|s^n}^{(n)}(\mathscr{C})$. We say that $\mathscr{C}$ is a $(2^{nR_0}, 2^{nR_1}, n, \varepsilon)$ code if $P_e^{(n)}(q, \mathscr{C}) \leq \varepsilon$, for all $q(s^n) \in \mathcal{P}(\mathcal{S}^n)$. Achievable rate pairs and the capacity region $\mathbb{C}(\mathcal{B})$ are defined as usual.

Definition 2. A $(2^{nR_0}, 2^{nR_1}, n)$ random code for the AVBC with causal SI consists of a collection of $(2^{nR_0}, 2^{nR_1}, n)$ codes $\{\mathscr{C}_\gamma = (f_\gamma^n, g_{1,\gamma}, g_{2,\gamma})\}_{\gamma \in \Gamma}$, along with a probability distribution $\mu(\gamma)$ over the code collection Γ. We denote such a code by $\mathscr{C}^\Gamma$. Similarly, a $(2^{nR_0}, 2^{nR_1}, n, \varepsilon)$ random code satisfies $P_e^{(n)}(q, \mathscr{C}^\Gamma) = \sum_{\gamma \in \Gamma} \mu(\gamma) P_e^{(n)}(q, \mathscr{C}_\gamma) \leq \varepsilon$, for all $q(s^n) \in \mathcal{P}(\mathcal{S}^n)$. The random code capacity region is denoted by $\mathbb{C}^\star(\mathcal{B})$, where the superscript ' $\star$ ' stands for random code achievability. The definitions above are naturally extended to the compound BC, by limiting the requirements to i.i.d. state distributions in $\mathcal{Q}$. The corresponding capacity regions are denoted by $\mathbb{C}(\mathcal{B}^{\mathcal{Q}})$ and $\mathbb{C}^\star(\mathcal{B}^{\mathcal{Q}})$. Next, define superposition coding using Shannon strategies [11].

Definition 3. A $(2^{nR_0}, 2^{nR_1}, n)$ Shannon strategy code consists of two strategy sequences, $u_0^n : [1:2^{nR_0}] \to \mathcal{U}_0^n$ and $u_1^n : [1:2^{nR_0}] \times [1:2^{nR_1}] \to \mathcal{U}_1^n$, an encoding function $\xi : \mathcal{U}_0 \times \mathcal{U}_1 \times \mathcal{S} \to \mathcal{X}$, and decoding functions $g_1(y_1^n)$ and $g_2(y_2^n)$. The codeword is given by $x^n = \xi^n(u_0^n(m_0), u_1^n(m_0, m_1), s^n) \triangleq \big[\xi(u_{0,i}^n(m_0), u_{1,i}^n(m_0, m_1), s_i) \big]_{i=1}^n$.

C. In the Absence of Side Information

Denote the AVBC without SI by $\mathcal{B}_0$, and let

$$\mathsf{R}_{0,in}^\star \triangleq \bigcup_{p(x,u)} \bigcap_{q(s)} \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq I_q(U; Y_2)\,, \\ & R_1 & \leq I_q(X; Y_1|U)\,, \\ & R_0 + R_1 & \leq I_q(X; Y_1) \end{array} \right\} \quad (3)$$

Theorem 1 (Jahn's Inner Bound [8]). The random code capacity region of an AVBC $\mathcal{B}_0$ with degraded message sets without SI is inner bounded by $\mathsf{R}_{0,in}^\star$. That is, $\mathbb{C}^\star(\mathcal{B}_0) \supseteq \mathsf{R}_{0,in}^\star$.

Theorem 2 (Dichotomy [8]). The capacity region of $\mathcal{B}_0$ either coincides with the random code capacity region or else, its interior is empty, *i.e.* $\mathbb{C}(\mathcal{B}_0) = \mathbb{C}^\star(\mathcal{B}_0)$ or else, $\text{int}\big(\mathbb{C}(\mathcal{B}_0)\big) = \emptyset$.

A necessary and sufficient condition for $\text{int}\big(\mathbb{C}(\mathcal{B}_0)\big) \neq \emptyset$ is given in terms of the following. A DMC $W_{Y|X,S}$ is said to be *symmetrizable* if for some conditional distribution $J(s|x)$,

$$\sum_{s \in \mathcal{S}} W_{Y|X,S}(y|x_1, s) J(s|x_2) = \sum_{s \in \mathcal{S}} W_{Y|X,S}(y|x_2, s) J(s|x_1)\,, \quad (4)$$

for all $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, $\text{int}\big(\mathbb{C}(\mathcal{B}_0)\big) \neq \emptyset$ if and only if $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ are *not* symmetrizable [6, 5, 7].

III. Main Results

We present our results below.

A. The Compound BC with Causal SI

Define

$$\mathsf{R}_{in}(\mathcal{B}^{\mathcal{Q}}) \triangleq \bigcup_{\substack{p(u_0,u_1), q \in \mathcal{Q} \\ \xi(u_0,u_1,s)}} \bigcap \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq I_q(U_0; Y_2)\,, \\ & R_1 & \leq I_q(U_1; Y_1|U_0)\,, \\ & R_0 + R_1 & \leq I_q(U_0, U_1; Y_1) \end{array} \right\} \quad (5)$$

$$\mathsf{R}_{out}(\mathcal{B}^{\mathcal{Q}}) \triangleq \bigcap_{q \in \mathcal{Q}} \bigcup_{\substack{p(u_0,u_1), \\ \xi(u_0,u_1,s)}} \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq I_q(U_0; Y_2)\,, \\ & R_1 & \leq I_q(U_1; Y_1|U_0)\,, \\ & R_0 + R_1 & \leq I_q(U_0, U_1; Y_1) \end{array} \right\} \quad (6)$$

s.t. $X = \xi(U_0, U_1, S)$, where U_0 and U_1 are auxiliary random variables, independent of S, and the union is over the pmf $p(u_0, u_1)$ and the set of all functions $\xi : \mathcal{U}_0 \times \mathcal{U}_1 \times \mathcal{S} \to \mathcal{X}$.

Lemma 3. 1) If $(R_0, R_1) \in \mathsf{R}_{in}(\mathcal{B}^{\mathcal{Q}})$, then for some $a > 0$ and sufficiently large n, there exists a $(2^{nR_0}, 2^{nR_1}, n, e^{-an})$ Shannon strategy code over the compound BC $\mathcal{B}^{\mathcal{Q}}$ with causal SI.

2) $\mathbb{C}(\mathcal{B}^{\mathcal{Q}}) = \mathsf{R}_{out}(\mathcal{B}^{\mathcal{Q}})$ if $\text{int}\big(\mathbb{C}(\mathcal{B}^{\mathcal{Q}})\big) \neq \emptyset$.

Lemma 3 is partially proved in Section V. The full proof is available in [9].

Remark 1. A BC $\mathcal{B}^q$ with random parameters, governed by a memoryless state $S \sim q(s)$, for a given $q \in \mathcal{P}(\mathcal{S})$, is a special case of the above. By Theorem 3, the capacity region of $\mathcal{B}^q$ with degraded message sets with causal SI is given by $\mathsf{R}_{out}(\mathcal{B}^{\mathcal{Q}})$, with $\mathcal{Q} = \{q\}$.

B. The AVBC with Causal SI

We give inner and outer bounds, on the random code and deterministic code capacity regions, and conditions under which the bounds coincide. Define

$$\mathsf{R}^{\star}_{in} \triangleq \mathsf{R}_{in}(\mathcal{B}^{\mathcal{P}(\mathcal{S})}) \, , \; \mathsf{R}^{\star}_{out} \triangleq \mathsf{R}_{out}(\mathcal{B}^{\mathcal{P}(\mathcal{S})}) . \tag{7}$$

We define a condition in terms of the following.

Definition 4. We say that a function $\xi(u_0, u_1, s)$ and a set $\mathcal{D} \subseteq \mathcal{P}(\mathcal{U}_0 \times \mathcal{U}_1)$ achieve $\mathsf{R}^{\star}_{in}$ and $\mathsf{R}^{\star}_{out}$ if

$$\mathsf{R}^{\star}_{in} = \bigcup_{p \in \mathcal{D}} \bigcap_{q(s)} \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq I_q(U_0; Y_2) \, , \\ & R_1 & \leq I_q(U_1; Y_1 | U_0) \, , \\ & R_0 + R_1 & \leq I_q(U_0, U_1; Y_1) \end{array} \right\} , \tag{8a}$$

$$\mathsf{R}^{\star}_{out} = \bigcap_{q(s)} \bigcup_{p \in \mathcal{D}} \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq I_q(U_0; Y_2) \, , \\ & R_1 & \leq I_q(U_1; Y_1 | U_0) \, , \\ & R_0 + R_1 & \leq I_q(U_0, U_1; Y_1) \end{array} \right\} , \tag{8b}$$

s.t. $X = \xi(U_0, U_1, S)$. That is, the unions in (5) and (6), taking $\mathcal{Q} = \mathcal{P}(\mathcal{S})$, can be restricted to the particular function $\xi(u_0, u_1, s)$ and set $\mathcal{D}$.

Given a function $\xi(u_0, u_1, s)$, if a set $\mathcal{D}$ achieves $\mathsf{R}^{\star}_{in}$ and $\mathsf{R}^{\star}_{out}$, then every set $\mathcal{D}' \supseteq \mathcal{D}$ achieves those regions. Yet, the condition below may hold with $\mathcal{D}$, but not $\mathcal{D}'$.

Definition 5. Define the condition $\mathscr{T}$; for some $\xi(u_0, u_1, s)$ and $\mathcal{D}^{\star}$ that achieve $\mathsf{R}^{\star}_{in}$ and $\mathsf{R}^{\star}_{out}$, there exists $q^* \in \mathcal{P}(\mathcal{S})$ which minimizes $I_q(U_0; Y_2)$, $I_q(U_1; Y_1|U_0)$, and $I_q(U_0, U_1; Y_1)$, for all $p(u_0, u_1) \in \mathcal{D}^{\star}$.

Intuitively, when $\mathscr{T}$ holds, $q^*(s)$ is the worst jamming strategy for both users simultaneously.

Theorem 4. 1) The random code capacity region of $\mathcal{B}$ with degraded message sets with causal SI is bounded by

$$\mathsf{R}^{\star}_{in} \subseteq \mathbb{C}^{\star}(\mathcal{B}) \subseteq \mathsf{R}^{\star}_{out} \, . \tag{9}$$

2) If Condition $\mathscr{T}$ holds, $\mathbb{C}^{\star}(\mathcal{B}) = \mathsf{R}^{\star}_{in} = \mathsf{R}^{\star}_{out}$.

Theorem 4 is partially proved in Section VI. The full proof is available in [9]. We move to the deterministic code capacity region, which demonstrates a dichotomy property.

Theorem 5. The capacity region of an AVBC $\mathcal{B}$ with degraded message sets with causal SI either coincides with the random code capacity region or else, it has an empty interior. That is, $\mathbb{C}(\mathcal{B}) = \mathbb{C}^{\star}(\mathcal{B})$ or else, $\text{int}\big(\mathbb{C}(\mathcal{B})\big) = \emptyset$.

Theorem 5 is proved in [9]. Theorem 4 and Theorem 5 yield the following corollary.

Corollary 6.

$$\mathbb{C}(\mathcal{B}) \supseteq \mathsf{R}^{\star}_{in} \, , \text{ if } \text{int}\big(\mathbb{C}(\mathcal{B})\big) \neq \emptyset \, , \tag{10}$$

$$\mathbb{C}(\mathcal{B}) \subseteq \mathsf{R}^{\star}_{out} \, . \tag{11}$$

A sufficient condition for $\text{int}\big(\mathbb{C}(\mathcal{B})\big) \neq \emptyset$ is given by the following. Let $U = (U_0, U_1)$, hence $\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1$. For every pair of functions $\xi : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$ and $\xi' : \mathcal{U}_0 \times \mathcal{S} \to \mathcal{X}$, define $V^{\xi}_{Y_1|U,S}(y_1|u,s) = W_{Y_1|X,S}(y_1|\xi(u,s),s)$ and $V^{\xi'}_{Y_2|U_0,S}(y_2|u_0,s) = W_{Y_2|X,S}(y_2|\xi'(u_0,s),s)$.

Corollary 7. If $V^{\xi}_{Y_1|U,S}$ and $V^{\xi'}_{Y_2|U_0,S}$ are non-symmetrizable for some $\xi : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$ and $\xi' : \mathcal{U}_0 \times \mathcal{S} \to \mathcal{X}$, and Condition $\mathscr{T}$ holds, then $\mathbb{C}(\mathcal{B}) = \mathsf{R}^{\star}_{in} = \mathsf{R}^{\star}_{out}$.

IV. Example

Consider an arbitrarily varying binary symmetric broadcast channel (BSBC) with correlated noises,

$$Y_1 = X + Z_S \mod 2 \, , \; Y_2 = X + N_S \mod 2 \, ,$$

where X, Y_1, Y_2, S, Z_S, N_S are binary, and

$$Z_s \sim \text{Bernoulli}(\theta_s) \, , \; N_s \sim \text{Bernoulli}(\varepsilon_s) \, , \text{ for } s \in \{0,1\} \, ,$$

where S, Z_0, Z_1, N_0, N_1 are independent, with $\theta_0 \leq \varepsilon_0 \leq \frac{1}{2}$ and $\frac{1}{2} \leq \varepsilon_1 \leq \theta_1$. Although Y_2 is degraded given $S = s$, we note that this channel is *not* degraded in the sense defined in [10]. We have the following results. First, without SI, the capacity region is $\mathbb{C}(\mathcal{B}_0) = \{(0,0)\}$. For the setting where causal SI is available at the encoder, we consider two cases.

Case 1: Suppose that $\theta_0 \leq 1 - \theta_1 \leq \varepsilon_0 \leq 1 - \varepsilon_1 \leq \frac{1}{2}$, *i.e.* $S = 1$ is a noisier channel state than $S = 0$, for both users. Then, our derivation shows that Condition $\mathscr{T}$ holds, and the capacity region with causal SI is given by

$$\mathbb{C}(\mathcal{B}) = \bigcup_{0 \leq \beta \leq 1} \left\{ \begin{array}{lrl} (R_0, R_1) : & R_0 & \leq 1 - h(\beta * \varepsilon_1) \, , \\ & R_1 & \leq h(\beta * \theta_1) - h(\theta_1) \end{array} \right\} , \tag{12}$$

where $h(\cdot)$ is the binary entropy function and $\alpha * \beta \triangleq (1 - \alpha)\beta + \alpha(1 - \beta)$. The derivation of the results is given in [9].

Figure 1 provides a graphical interpretation. Let $\mathbb{C}(\mathcal{B}^q)$ denote the capacity region of a BSBC $W_{Y_1,Y_2|X,S}$ with causal SI governed by an i.i.d. state $S \sim \text{Bernoulli}(q)$, for a given $0 \leq q \leq 1$. Condition $\mathscr{T}$ implies that there exists $0 \leq q^* \leq 1$ such that $\mathbb{C}(\mathcal{B}^{q^*}) \subseteq \mathbb{C}(\mathcal{B}^q)$ for all $0 \leq q \leq 1$, hence, $\mathbb{C}(\mathcal{B}) = \mathbb{C}(\mathcal{B}^{q^*})$. Indeed, looking at Figure 1, it appears that the regions $\mathbb{C}(\mathcal{B}^q)$, for $0 \leq q \leq 1$, form a well ordered set, hence $\mathbb{C}(\mathcal{B}) = \mathbb{C}(\mathcal{B}^{q^*})$ with $q^* = 1$.

Case 2: Suppose that $\theta_0 \leq 1 - \theta_1 \leq 1 - \varepsilon_1 \leq \varepsilon_0 \leq \frac{1}{2}$, *i.e.* $S = 1$ is noisier for User 1, while $S = 0$ is noisier for User 2. Figure 2(a) demonstrates the gap between the bounds in this case. In Figure 2(b), the dashed and dotted lines depict $\mathbb{C}(\mathcal{B}^{q=0})$ and $\mathbb{C}(\mathcal{B}^{q=1})$ respectively. The colored lines depict $\mathbb{C}(\mathcal{B}^q)$ for $0 < q < 1$. It appears that the intersection $\mathsf{R}^{\star}_{out} = \bigcap_{0 \leq q \leq 1} \mathbb{C}(\mathcal{B}^q)$ reduces to $\mathbb{C}(\mathcal{B}^{q=0}) \cap \mathbb{C}(\mathcal{B}^{q=1})$.

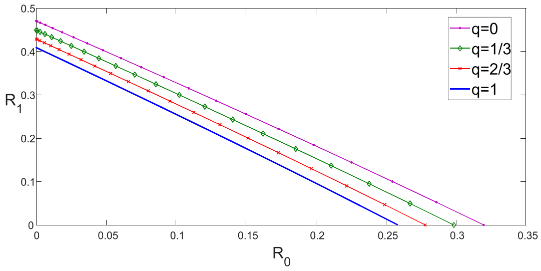
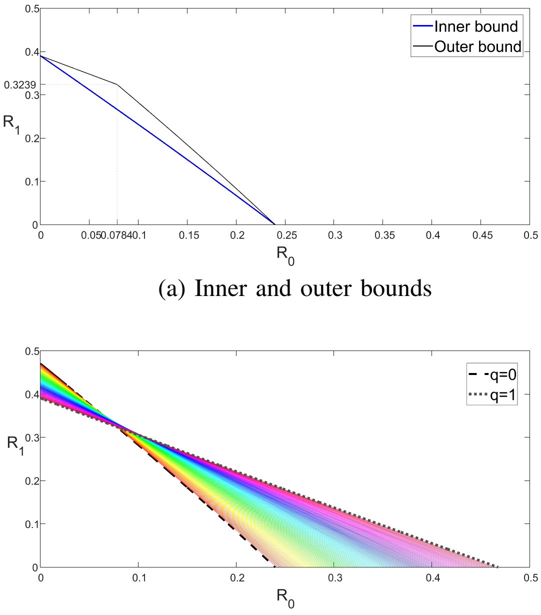


Fig. 1. The capacity region of the arbitrarily varying binary symmetric broadcast channel with correlated noises, in case 1. The area below the lowest curve is the capacity region of the AVBC $\mathcal{B}$ with causal SI, with $\theta_0 = 0.12$, $\theta_1 = 0.85$, $\varepsilon_0 = 0.18$ and $\varepsilon_1 = 0.78$. The curves depict $\mathbb{C}(\mathcal{B}^q)$ for $q = 0, 1/3, 2/3, 1$. The capacity region is $\mathbb{C}(\mathcal{B}) = \mathbb{C}(\mathcal{B}^{q=1})$.



(a) Inner and outer bounds

(b) The regions $\mathbb{C}(\mathcal{B}^q)$, $0 \leq q \leq 1$

Fig. 2. The inner and outer bounds on the capacity region of the arbitrarily varying binary symmetric broadcast channel with correlated noises, in case 2, with $\theta_0 = 0.12$, $\theta_1 = 0.85$, $\varepsilon_0 = 0.22$ and $\varepsilon_1 = 0.88$.

V. Proof of Lemma 3

Consider part 1. We use superposition coding with Shannon strategies, and decode using joint typicality with respect to a state type which is "close" to some $q \in \mathcal{Q}$. Let $\delta > 0$. Basic method of types concepts, such as a δ-typical set $\mathcal{A}^\delta(P_X)$, are defined as in [4]. Also, define a set of state types,

$$\hat{\mathcal{Q}}_n = \left\{ \hat{P}_{s^n} \;:\; s^n \in \mathcal{A}^{\delta_1}(q) \,,\, \text{for some } q \in \mathcal{Q} \right\} \,, \tag{13}$$

i.e. the set of types that are δ_1-close to some $q(s)$ in $\mathcal{Q}$, with $\delta_1 \triangleq \frac{\delta}{2 \cdot |\mathcal{S}|}$. Now, a code for the compound BC with causal SI is constructed as follows.

Codebook Generation: Fix P_{U_0,U_1} and $\xi(u_0, u_1, s)$. Generate 2^{nR_0} independent sequences at random, $u_0^n(m_0) \sim \prod_{i=1}^n P_{U_0}(u_{0,i})$ for $m_0 \in [1 : 2^{nR_0}]$. Then, for every $m_0 \in [1 : 2^{nR_0}]$, generate 2^{nR_1} sequences at random, $u_1^n(m_0, m_1) \sim \prod_{i=1}^n P_{U_1|U_0}(u_{1,i}|u_{0,i}(m_0))$ for $m_1 \in [1 : 2^{nR_1}]$, conditionally independent given $u_0^n(m_0)$.

Encoding: To send (m_0, m_1), transmit $x_i = \xi(u_{0,i}(m_0), u_{1,i}(m_0, m_1), s_i)$ at time $i \in [1 : n]$.

Decoding: Let $P^q_{Y_1,Y_2|U_0,U_1}(y_1, y_2|u_0, u_1) = \sum_{s\in\mathcal{S}} q(s) \cdot W_{Y_1,Y_2|X,S}\,(y_1, y_2|\xi(u_0, u_1, s), s)$. Observing y_2^n, decoder 2 finds a unique $\widetilde{m}_0 \in [1 : 2^{nR_0}]$ such that $(u_0^n(\widetilde{m}_0), y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^q_{Y_2|U_0})$ for some $q \in \hat{\mathcal{Q}}_n$. If there is none, or more than one such message, declare an error. Similarly, decoder 1 finds a unique pair $(\hat{m}_0, \hat{m}_1)$ such that $(u_0^n(\hat{m}_0), u_1^n(\hat{m}_0, \hat{m}_1), y_1^n) \in \mathcal{A}^\delta(P_{U_0,U_1} P^q_{Y_1|U_0,U_1})$ for some $q \in \hat{\mathcal{Q}}_n$. If there is none, or more than one, declare an error.

Analysis of Probability of Error: Assume w.l.o.g. that the users sent $(M_0, M_1) = (1, 1)$. Let $q(s) \in \mathcal{Q}$ denote the *actual* state distribution chosen by the jammer. The error event for decoder 2 is the union of the following events.

$$\begin{aligned}
\mathcal{E}_{2,1} =& \{(U_0^n(1), Y_2^n) \notin \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0}) \,,\, \text{for all } q' \in \hat{\mathcal{Q}}_n\} \\
\mathcal{E}_{2,2} =& \{(U_0^n(m_0), Y^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0}) \,, \\
& \quad \text{for some } m_0 \neq 1,\, q' \in \hat{\mathcal{Q}}_n\}
\end{aligned} \tag{14}$$

We now claim that $\mathcal{E}_{2,1}$ implies that

$$(U_0^n(1), Y_2^n) \notin \mathcal{A}^{\delta/2}(P_{U_0} P^{q''}_{Y_2|U_0}) \,,\, \text{for all } q'' \in \mathcal{Q} \,. \tag{15}$$

Indeed, assume to the contrary that $\mathcal{E}_{2,1}$ holds but $(U_0^n(1), Y_2^n) \in \mathcal{A}^{\delta/2}(P_{U_0} P^{q''}_{Y_2|U_0})$ for some $q'' \in \mathcal{Q}$. Then, for sufficiently large n, there exists a type $q'(s)$ such that $|q'(s) - q''(s)| \leq \delta_1$. Thus, $q' \in \hat{\mathcal{Q}}_n$ (see (13)), and $|P^{q'}_{Y_2|U_0}(y_2|u_0) - P^{q''}_{Y_2|U_0}(y_2|u_0)| \leq \frac{\delta}{2}$. Hence, $(U_0^n(1), Y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0})$, which contradicts $\mathcal{E}_{2,1}$. Thus,

$$\Pr(\mathcal{E}_{2,1}) \leq \Pr\left((U_0^n(1), Y_2^n) \notin \mathcal{A}^{\delta/2}(P_{U_0} P^q_{Y_2|U_0}) \right) \,. \tag{16}$$

The last expression tends to zero exponentially as $n \to \infty$ by the law of large numbers and Chernoff's bound. Now,

$$\begin{aligned}
&\Pr(\mathcal{E}_{2,2}) \leq (n+1)^{|\mathcal{S}|} \cdot 2^{nR_0} \\
&\cdot \sup_{m_0 \neq 1, q' \in \hat{\mathcal{Q}}_n} \Pr\left\{ (U_0^n(m_0), Y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0}) \right\} .
\end{aligned} \tag{17}$$

Since $U_0^n(m_0)$ is independent of Y_2^n for every $m_0 \neq 1$,

$$\begin{aligned}
&\Pr\left((U_0^n(m_0), Y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0}) \right) \\
&= \sum_{u_0^n \in \mathcal{U}_2^n} P_{U_0^n}(u_0^n) \cdot \sum_{y_2^n \,:\, (u_0^n, y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0})} P^q_{Y_2^n}(y_2^n)
\end{aligned} \tag{18}$$

If $(u_0^n, y_2^n) \in \mathcal{A}^\delta(P_{U_0} P^{q'}_{Y_2|U_0})$, then by [4, Lemmas 2.6–2.7], for some arbitrarily small $\varepsilon_1 > 0$,

$$P^q_{Y_2^n}(y_2^n) \leq 2^{-nH(\hat{P}_{y_2^n})} \leq 2^{-n(H_{q'}(Y_2) - \varepsilon_1)} \,. \tag{19}$$

Therefore, by (17)−(19) and [4, Lemma 2.13],

$$\Pr(\mathcal{E}_{2,2}) \leq (n+1)^{|\mathcal{S}|} \cdot \sup_{q' \in \mathcal{Q}} 2^{-n[I_{q'}(U_0;Y_2) - R_0 - \varepsilon_2]} \,, \tag{20}$$

which tends to zero exponentially as $n \to \infty$, provided that $R_0 < \inf_{q' \in \mathcal{Q}} I_{q'}(U_0; Y_2) - \varepsilon_2$. It remains to bound the probabilty of error for User 1. Using the standard techniques in conjuction with the arguements above, we find that the probability of error decays exponentially, provided that $R_1 < \inf_{q' \in \mathcal{Q}} I_{q'}(U_1;Y_1|U_0) - \varepsilon_3$ and $R_0 + R_1 < \inf_{q' \in \mathcal{Q}} I_{q'}(U_0,U_1;Y_1) - \varepsilon_4$.

Achievability for Part 2 is proved using a similar coding scheme, with the addition of a codeword *suffix*. At time $i = n+1$, having completed the transmission of the messages, the type of the state sequence s^n is known to the encoder. If $\text{int}\big(\mathbb{C}(\mathcal{B}^{\mathcal{Q}})\big) \neq \emptyset$, the type of s^n can be reliably communicated to both receivers as a suffix of negligible length. The receivers first estimate the type, and then use joint typicality with respect to the estimated type. We note that this does not agree with the definition of a Shannon strategy code, since the transmission of the type depends upon previous states. The details and the converse proof for part 2 are omitted, due to lack of space. □

VI. Proof of Theorem 4

Let $(R_0, R_1) \in \mathsf{R}^{\star}_{in}$. To prove the inner bound, we begin with Ahlswede's RT, stated below. Let $\varphi : \mathcal{S}^n \to [0,1]$ be a given function. If, for some fixed $\alpha_n \in (0,1)$, and for all $q(s^n) = \prod_{i=1}^n q(s_i)$, with $q \in \mathcal{P}(\mathcal{S})$,

$$\sum_{s^n \in \mathcal{S}^n} q(s^n)\varphi(s^n) \leq \alpha_n \ , \tag{21}$$

then, $\frac{1}{n!} \sum\limits_{\pi \in \Pi_n} \varphi(\pi s^n) \leq (n+1)^{|\mathcal{S}|}\alpha_n$, for all $s^n \in \mathcal{S}^n$, where Π_n is the set of all n-tuple permutations $\pi : \mathcal{S}^n \to \mathcal{S}^n$. Let $\mathscr{C}$ be a Shannon strategy code (see Definition 3) as in part 1 of Lemma 3. Hence, (21) holds with $\varphi(s^n) = P^{(n)}_{e|s^n}(\mathscr{C})$ and $\alpha_n = e^{-an}$. Thus, by Ahlswede's RT, for every $s^n \in \mathcal{S}^n$,

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} P^{(n)}_{e|\pi s^n}(\mathscr{C}) \leq (n+1)^{|\mathcal{S}|} e^{-an} \leq e^{-\theta n} \tag{22}$$

for large n, with some $\theta > 0$. On the other hand, For every $\pi \in \Pi_n$, $P^{(n)}_{e|\pi s^n}(\mathscr{C}) = \frac{1}{2^{n(R_0+R_1)}} \sum_{m_0,m_1} e(m_0, m_1, \pi s^n)$, where

$$\begin{aligned} & e(m_0, m_1, \pi s^n) \\ & \overset{(a)}{=} \sum_{(\pi y_1^n, \pi y_2^n) \notin \mathcal{D}} W^n(y_1^n, y_2^n | \pi^{-1} \xi^n(u_0^n, u_1^n, \pi s^n), s^n) \\ & \overset{(b)}{=} \sum_{(\pi y_1^n, \pi y_2^n) \notin \mathcal{D}} W^n(y_1^n, y_2^n | \xi^n(\pi^{-1} u_0^n, \pi^{-1} u_1^n, s^n), s^n) \end{aligned} \tag{23}$$

where we have used the short notations $W^n \equiv W_{Y_1^n, Y_2^n | X^n, S^n}$, $\mathcal{D} \equiv \mathcal{D}(m_0, m_1)$, $u_0^n \equiv u_0^n(m_0)$, $u_1^n \equiv u_1^n(m_0, m_1)$; in (a) we change the summation order and use the fact the channel is memoryless, and (b) follows since for a Shannon strategy code, $x_i = \xi(u_{0,i}, u_{1,i}, s_i)$. Then, consider the random code $\mathscr{C}^{\Pi}$, specified by $f^n_\pi(m_0, m_1, s^n) = \xi^n(\pi^{-1} u_0^n(m_0), \pi^{-1} u_1^n(m_0, m_1), s^n)$ and $g_{k,\pi}(y_k^n) = g_k(\pi y_k^n)$, $k = 1, 2$, for $\pi \in \Pi_n$, with $\mu(\pi) = \frac{1}{n!}$. Such permutations can be implemented without knowing s^n, hence this coding scheme does not violate the causality requirement. From (23), we see that $P^{(n)}_{e|s^n}(\mathscr{C}^{\Pi}) = \sum_{\pi \in \Pi_n} \mu(\pi) P^{(n)}_{e|\pi s^n}(\mathscr{C})$, for all $s^n \in \mathcal{S}^n$, hence by (22), we have that the probability of error is bounded by $P^{(n)}_e(q, \mathscr{C}^{\Pi}) \leq e^{-\theta n}$, for every $q(s^n) \in \mathcal{P}(\mathcal{S}^n)$.

The outer bound follows from part 2 of Lemma 3, since the random code capacity region of the AVBC is included within the random code capacity region of the compound BC.

The proof of part 2 of the theorem is straightforward. □

References

[1] R. Ahlswede. "Elimination of correlation in random codes for arbitrarily varying channels". *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 44.2 (June 1978), pp. 159–175.

[2] R. Ahlswede. "Arbitrarily varying channels with states sequence known to the sender". *IEEE Trans. Inform. Theory* 32.5 (Sept. 1986), pp. 621–629.

[3] D. Blackwell, L. Breiman, and A. J. Thomasian. "The capacities of certain channel classes under random coding". *Ann. Math. Statist.* 31.3 (Sept. 1960), pp. 558–567.

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

[5] I. Csiszár and P. Narayan. "The capacity of the arbitrarily varying channel revisited: positivity, constraints". *IEEE Trans. Inform. Theory* 34.2 (Mar. 1988), pp. 181–193.

[6] T. Ericson. "Exponential error bounds for random codes in the arbitrarily varying channel". *IEEE Trans. Inform. Theory* 31.1 (Jan. 1985), pp. 42–48.

[7] E. Hof and S. I. Bross. "On the deterministic-code capacity of the two-user discrete memoryless Arbitrarily Varying General Broadcast channel with degraded message sets". *IEEE Trans. Inform. Theory* 52.11 (Nov. 2006), pp. 5023–5044.

[8] J. H. Jahn. "Coding of arbitrarily varying multiuser channels". *IEEE Trans. Inform. Theory* 27.2 (Mar. 1981), pp. 212–226.

[9] U. Pereg and Y. Steinberg. "The Arbitrarily Varying Broadcast Channel with Degraded Message Sets with Causal Side Information at the Encoder". `arXiv:1709.04770` (Sept. 2017). URL: https://arxiv.org/pdf/1709.04770.pdf.

[10] U. Pereg and Y. Steinberg. "The arbitrarily varying degraded broadcast channel with causal Side information at the encoder". *Proc. IEEE Int'l Symp. Inform. Theory (ISIT'2017)*. Aachen, Germany.

[11] C. E. Shannon. "Channels with side Information at the transmitter". *IBM J. Res. Dev.* 2.4 (Oct. 1958), pp. 289–293.

[12] A. Winshtok and Y. Steinberg. "The arbitrarily varying degraded broadcast channel with states known at the encoder". *Proc. IEEE Int'l Symp. Inform. Theory (ISIT'2006)*. Seattle, Washington, July 2006, pp. 2156–2160.

Distributed Hypothesis Testing Over a Noisy Channel

[1]Sadaf Salehkalaibar and [2]Michèle Wigger
[1]ECE Department, College of Engineering, University of Tehran, Tehran, Iran, s.saleh@ut.ac.ir
[2]LTCI, Telecom ParisTech, Université Paris-Saclay, 75013 Paris, France, michele.wigger@telecom-paristech.fr

***Abstract*—A coding and testing scheme is presented for the distributed hypothesis testing problem over a noisy channel. The coding scheme combines the Shimokawa-Han-Amari hypothesis testing scheme with Borade's unequal error protection (UEP) channel coding. The type-II error exponent of our scheme consists of three competing error exponents: two of them coincide with the exponents found by Shimokawa-Han-Amari for distributed hypothesis testing over a noiseless link (with the rate be replaced by the mutual information between channel input and output), and the third includes Borade's miss-detection exponent for UEP over a noisy channel. Depending on the problem setup, any of the three exponents can be active. When testing against conditional independence, only the two Shimokawa-Han-Amari exponents are active, and the scheme achieves the optimal type-II error exponent found by Sreekuma and Gündüz.**

I. Introduction

Consider a distributed hypothesis testing problem where a sensor describes its collected information to a remote decision center over a noisy channel. The decision center decides on a binary hypothesis ($\mathcal{H} = 0$ or $\mathcal{H} = 1$) that determines the joint probability distribution underlying its own observation and the information observed at the sensor. The goal of the communication is to maximize the type-II error (deciding $\hat{\mathcal{H}} = 0$ when $\mathcal{H} = 1$) exponent under a constrained type-I error (deciding $\hat{\mathcal{H}} = 1$ when $\mathcal{H} = 0$).

The special case of this problem where communication takes place over a noiseless link was studied in [1]–[4]. These works present achievable type-II error exponents for general joint probability distributions underlying the two hypotheses and the optimal type-II error exponent for the special case called "testing against conditional independence" [4]. Distributed hypothesis testing problems over noiseless networks with multiple sensors or decision centers or with relays have been considered in [4]–[8]. The work most closely related to this paper is by Sreekumar and Gündüz [9]. It proves that the optimal type-II error exponent for "testing against conditional independence" over a noisy channel, coincides with the optimal type-II error exponent of the same test over a noiseless link of rate equal to the capacity of the noisy channel. Their result is based on a joint hypothesis-testing and channel-coding scheme, see also [9, Remark 6] for a discussion on this.

In this work, we propose a coding scheme for distributed hypothesis testing over a noisy channel with general probability distributions. The coding and testing scheme applies separate hypothesis testing and channel coding by combining the Shimokawa-Han-Amari (SHA) hypothesis-testing scheme [3] with Borade's unequal error protection (UEP) channel coding [12]. The idea is to reinforce the protection of the message that the SHA scheme produces to indicate that the transmitter decides on the alternative hypothesis $\mathcal{H} = 1$. Our analysis in general shows three competing error exponents, two of them coincide with the two competing error exponents obtained for testing over a noiseless link [3] when the communication rate is replaced by the mutual information between input and output of the channel. The third error exponent depends again on this mutual information, and on Borade's *miss-detection exponent* [12] for channel coding with UEP. In the special case of "testing against conditional independence", recover the optimal exponent by Sreekuma and Gündüz [9]. In this case, our third error exponent is never active and the overall type-II error exponent depends on the noisy channel only through its capacity.

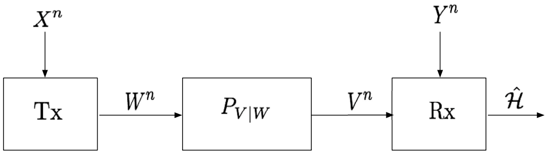


Fig. 1. Hypothesis testing over a noisy channel

Notation: We mostly follow the notation in [10]. Moreover, we use $\mathrm{tp}(\cdot)$ to denote the *joint type* of a tuple. For a joint type π_{AB} over alphabets $\mathcal{A} \times \mathcal{B}$, we denote by $I_{\pi_{AB}}(A; B)$ the mutual information of a pair of random variables (A, B) with probability mass function (pmf) π_{AB}. Similarly for entropy, conditional entropy, and conditional mutual information. When it is unambiguous, we may abbreviate π_{AB} by π. We also abbreviate *independent and identically distributed* by i.i.d.

II. System Model

Consider the distributed hypothesis testing problem in Fig. 1, where a transmitter observes source sequence X^n and a receiver source sequence Y^n. Under the null hypothesis:

$$\mathcal{H} = 0\colon (X^n, Y^n) \quad \text{i.i.d.} \sim P_{XY}, \tag{1}$$

and under the alternative hypothesis:

$$\mathcal{H} = 1\colon (X^n, Y^n) \quad \text{i.i.d.} \sim Q_{XY}. \tag{2}$$

for two given pmfs P_{XY} and Q_{XY}. The transmitter can communicate with the receiver over n uses of a discrete memory channel $(\mathcal{W}, \mathcal{V}, P_{V|W})$ where $\mathcal{W}$ denotes the finite channel input alphabet and $\mathcal{V}$ the finite channel output alphabet. Specifically, the transmitter feeds inputs

$$W^n = f^{(n)}(X^n) \tag{3}$$

to the channel, where $f^{(n)}$ denotes the chosen (possibly stochastic) encoding function

$$f^{(n)} : \mathcal{X}^n \to \mathcal{W}^n. \tag{4}$$

Based on the sequence of channel outputs V^n and the source sequence Y^n, the receiver decides on the hypothesis $\mathcal{H}$. That means, it produces the guess

$$\hat{\mathcal{H}} = g^{(n)}(V^n, Y^n), \tag{5}$$

by means of a decoding function

$$g^{(n)} \colon \mathcal{V}^n \times \mathcal{Y}^n \to \{0, 1\}. \tag{6}$$

Definition 1: For each $\epsilon \in (0, 1)$, an exponent θ is said ϵ-achievable, if for each sufficiently large blocklength n, there exist encoding and decoding functions $(f^{(n)}, g^{(n)})$ such that the corresponding type-I and type-II error probabilities at the receiver

$$\alpha_n \triangleq \Pr[\hat{\mathcal{H}} = 1 | \mathcal{H} = 0], \tag{7}$$

$$\beta_n \triangleq \Pr[\hat{\mathcal{H}} = 0 | \mathcal{H} = 1], \tag{8}$$

satisfy

$$\alpha_n \leq \epsilon, \tag{9}$$

and

$$-\overline{\lim_{n\to\infty}} \frac{1}{n} \log \beta_n \geq \theta. \tag{10}$$

The goal is to maximize the type-II error exponent θ.

III. Coding and Testing Scheme

We describe a coding and testing scheme for the general distributed hypothesis testing problem over a noisy channel. The analysis of the scheme is postponed to Section V.

Preparations: Choose a large positive integer n, an auxiliary distribution P_T over $\mathcal{W}$, a conditional channel input distribution $P_{W|T}$, and a conditional source distribution $P_{S|X}$ over a finite auxiliary alphabet $\mathcal{S}$ so that

$$I(S; X) < I(S; Y) + I(V; W|T), \tag{11}$$

where the mutual informations in (11) are calculated according to the following joint distribution

$$P_{SXYWVT} = P_{S|X} \cdot P_{XY} \cdot P_T \cdot P_{W|T} \cdot P_{V|W}. \tag{12}$$

Then choose a sufficiently small $\mu > 0$ and nonnegative rates (R, R') so that

$$R + R' = I(X; S) + \mu \tag{13}$$

$$R < I(V; W|T) \tag{14}$$

$$R' < I(S; Y). \tag{15}$$

Code Construction: Construct a random codebook

$$\mathcal{C}_S = \big\{S^n(m, \ell) \colon m \in \{1, ..., \lfloor 2^{nR} \rfloor\}, \ell \in \{1, ..., \lfloor 2^{nR'} \rfloor\}\big\},$$

by independently drawing all codewords i.i.d. according to $P_S(s) = \sum_{x\in\mathcal{X}} P_X(x) P_{S|X}(s|x)$.

Generate a sequence T^n i.i.d. according to P_T. Construct a random codebook

$$\mathcal{C}_W = \big\{W^n(m) : m \in \{1, ..., \lfloor 2^{nR} \rfloor\}\big\}$$

superpositioned on T^n where each codeword is drawn independently according to $P_{W|T}$ conditioned on T^n. Reveal the realizations of the codebooks and the sequence T^n to all terminals.

Transmitter: Given that it observes the source sequence $X^n = x^n$, the transmitter looks for a pair (m, ℓ) that satisfies

$$(s^n(m, \ell), x^n) \in \mathcal{T}^n_{\mu/2}(P_{SX}). \tag{16}$$

If successful, it picks one of these pairs uniformly at random and sends the codeword $w^n(m)$ over the channel. Otherwise it sends the sequence of inputs t^n over the channel.

Receiver: Assume that $V^n = v^n$ and $Y^n = y^n$ and that the "time-sharing sequence" $T^n = t^n$. The receiver first looks for an index $m' \in \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ so that

$$(w^n(m'), v^n, t^n) \in \mathcal{T}^n_{\mu}(P_{WVT}). \tag{17}$$

If it is not successful, it declares $\hat{\mathcal{H}} = 1$. Otherwise, it randomly picks one of the indices ℓ' that satisfy

$$H_{\text{tp}(s^n(m',\ell'),y^n)}(S|Y) = \min_{\tilde{\ell}\in\{1,\ldots,\lfloor 2^{nR'} \rfloor\}} H_{\text{tp}(s^n(m',\tilde{\ell}),y^n)}(S|Y), \tag{18}$$

and checks whether

$$(s^n(m', \ell'), y^n) \in \mathcal{T}^n_{\mu}(P_{SY}). \tag{19}$$

If successful, it declares $\hat{\mathcal{H}} = 0$. Otherwise, it declares $\hat{\mathcal{H}} = 1$.

IV. An Achievable Error Exponent

The coding and testing scheme described in the previous section allows to establish the following theorem.

Theorem 1: Every error exponent $\theta \geq 0$ that satisfies the following condition (33) is achievable:

$$\theta \leq \max_{\substack{P_{S|X}, P_{TW}: \\ I(S;X|Y) \leq I(W;V|T)}} \min\big\{\theta_1, \theta_2, \theta_3\big\}, \tag{20}$$

where

$$\theta_1 = \min_{\substack{\tilde{P}_{SXY}: \\ \tilde{P}_{SX} = P_{SX} \\ \tilde{P}_{SY} = P_{SY}}} D(\tilde{P}_{SXY} || Q_{XY} P_{S|X}), \tag{21}$$

$$\theta_2 = \min_{\substack{\tilde{P}_{SXY}: \\ \tilde{P}_{SX} = P_{SX} \\ \tilde{P}_Y = P_Y \\ H(S|Y) \leq H_{\tilde{P}}(S|Y)}} \Big[D(\tilde{P}_{SXY} || P_{S|X} Q_{XY}) + I(V; W|T) - I(S; X|Y)\Big], \tag{22}$$

$$\theta_3 = D(P_Y || Q_Y) + I(V; W|T) - I(S; X|Y)$$

$$+\sum_{t\in\mathcal{W}} P_T(t)\cdot D(P_{V|T=t}||P_{V|W=t}), \tag{23}$$

and all expressions are calculated with respect to the joint distribution in (12).

Proof: Based on the scheme in Section V. ■

Lemma 1: It suffices to consider the auxiliary random variable S over an alphabet $\mathcal{S}$ that is of size $|\mathcal{S}| = |\mathcal{X}| + 2$. For the specical case of $P_Y = Q_Y$, it suffices to consider $|\mathcal{S}| = |\mathcal{X}| + 1$.

Proof: Based on Carathéodory's theorem. Omitted. ■

Our coding and testing scheme combines the SHA hypothesis testing scheme for a noiseless link [3] with Borade's UEP channel coding that protects the 0-message (which indicates that the transmitter decides on $\mathcal{H} = 1$) better than the other messages [11], [12]. In fact, since here we are only interested in the type-II error exponent, the receiver should decide on $\mathcal{H} = 0$ only if the transmitter also shares this opinion.

The expressions in Theorem 1 show three competing error exponents. In (21) and (22), we recognize the two competing error exponents of the SHA scheme for the noiseless setup: θ_1 is the exponent associated to the event that the receiver reconstructs the correct binned codeword and θ_2 is associated to the event that either the binning or the noisy channel introduces a decoding error. The exponent θ_3 in (23) is new and can be associated to the event that the specially protected 0-message is wrongly decoded. We remark in particular that θ_3 contains the term

$$E_{\text{miss}} := \sum_{t\in\mathcal{W}} P_T(t)\cdot D(P_{V|T=t}||P_{V|W=t}), \tag{24}$$

which represents the largest possible *miss-detection exponent* for a single specially protected message at a given rate $I(W;V|T)$ [12, Th. 34].

Which of the three exponents $\theta_1, \theta_2, \theta_3$ is smallest depends on the source and channel parameters and the choice of $P_{S|X}$ and P_W. Notice that the third error exponent θ_3 is inactive for channels with large miss-detection exponent (24), such as binary symmetric channels with small cross-over probability, or for sources where

$$\min_{\substack{\tilde{P}_{SXY}:\\ \tilde{P}_{SX}=P_{SX}\\ \tilde{P}_Y=P_Y}} D(\tilde{P}_{SXY}||P_{S|X}Q_{XY}) = D(P_Y||Q_Y), \tag{25}$$

This is the case for example when "testing against conditional independence" [4] where both terms are 0.

Corollary 1 (Lemma 5 in [9]): Consider the "testing against independence" setup where

$$Y = (\bar{Y}, Z), \tag{26}$$

and $Q_{X\bar{Y}Z}$ decomposes as

$$Q_{X\bar{Y}Z} = P_{XZ}\cdot P_{\bar{Y}|Z}. \tag{27}$$

Error exponent $\theta \geq 0$ is achievable if,

$$\theta \leq \max_{\substack{P_{S|X},\ P_W:\\ I(S;X|Z)\leq I(W;V)}} I(S;\bar{Y}|Z), \tag{28}$$

where mutual informations are calculated with respect to the joint law $P_{X\bar{Y}Z}P_{S|X}P_W P_{V|W}$.

Proof: Fix independent random variables T and W and a random variable S so that

$$I(S;X|Z) \leq I(W;V|T) = I(W;V). \tag{29}$$

Then, Theorem 1 specializes to:

$$\begin{aligned}
\theta_1 &= \min_{\substack{\tilde{P}_{SX\bar{Y}Z}:\\ \tilde{P}_{SX}=P_{SX}\\ \tilde{P}_{S\bar{Y}Z}=P_{S\bar{Y}Z}}} D(\tilde{P}_{SX\bar{Y}Z}||Q_{X\bar{Y}Z}P_{S|X})\\
&= \min_{\substack{\tilde{P}_{SX\bar{Y}Z}:\\ \tilde{P}_{SX}=P_{SX}\\ \tilde{P}_{S\bar{Y}Z}=P_{S\bar{Y}Z}}} D(\tilde{P}_{SX\bar{Y}Z}||P_{XZ}P_{\bar{Y}|Z}P_{S|X})\\
&= D(P_{S\bar{Y}Z}||P_Z P_{\bar{Y}|Z}P_{S|Z})\\
&= I(S;\bar{Y}|Z).
\end{aligned}$$

Moreover, exponents θ_2 and θ_3 cannot be smaller than $I(S;\bar{Y}|Z)$ because of the nonnegativity of the KL-divergence and the mutual information and because

$$\begin{aligned}
&I(V;W) - I(S;X) + I(S;\bar{Y},Z)\\
&\quad = I(V;W) - I(S;X|Z) + I(S;\bar{Y}|Z)\\
&\quad \geq I(S;\bar{Y}|Z),
\end{aligned} \tag{30}$$

where the inequality holds by (29). ■

Notice that the error exponent in Corollary 1 is optimal [9].

We now present an example and evaluate the largest type-II error exponents attained by our scheme. We also show that depending on the choice of the model parameters, a different error exponent θ_1, θ_2, or θ_3 is active.

Example 1: Let under the null hypothesis

$$\begin{aligned}
\mathcal{H} = 0: \qquad & X \sim \text{Bern}(p_0), \quad Y = X \oplus N_0,\\
& N_0 \sim \text{Bern}(q_0),
\end{aligned} \tag{31}$$

for N_0 independent of X. Under the alternative hypothesis:

$$\mathcal{H} = 1: \qquad X \sim \text{Bern}(p_1), \quad Y \sim \text{Bern}(p_0 \star q_0), \tag{32}$$

with X and Y independent. Assume that $P_{V|W}$ is a binary symmetric channel (BSC) with cross-over probability $r \in [0, 1/2]$.

For this example, $P_Y = Q_Y$ and Theorem 1 simplifies to:

$$\theta \leq \max_{\substack{P_{S|X}, P_{TW}:\\ I(S;X|Y)\leq I(W;V|T)}} \min\{\theta_1, \theta_2, \theta_3\}, \tag{33}$$

where

$$\theta_1 \leq D(P_X||Q_X) + I(S;Y), \tag{34}$$

$$\theta_2 \leq D(P_X||Q_X) + I(V;W|T) + I(S;Y) - I(S;X), \tag{35}$$

$$\theta_3 \leq \sum_{t\in\mathcal{W}} P_T(t) D(P_{V|T=t}||P_{V|W=t}) + I(V;W|T) + I(S;Y) - I(S;X). \quad (36)$$

Depending on the parameters of the setup and the choice of the auxiliary distributions, either of the exponents θ_1, θ_2, or θ_3 is active. For example, when the cross-over probability of the BSC is large, $r \geq 0.4325$,

$$D(P_X||Q_X) \geq \sum_{t\in\mathcal{W}} P_T(t) D(P_{V|T=t}||P_{V|W=t}) + I(V;W|T), \quad (37)$$

and irrespective of the choice of the random variables S, T, W the exponent θ_3 is smaller than θ_1 and θ_2. It is then optimal to choose S constant and (T, W) so as to maximize the sum $\sum_{t\in\mathcal{W}} P_T(t)D(P_{V|T=t}||P_{V|W=t})+I(V;W|T)$. In particular, for a scenario with parameters $p_0 = 0.1, q_0 = 0.25, p_1 = 0.2$ and $r = \frac{4}{9}$ one obtains numerically that the optimal error exponent achieved by our scheme is $\theta = 0.0358$.

In contrast, when the cross-over probability of the BSC is small, the miss-detection exponent (24) is large and the exponent θ_3 is never active irrespective of the choice of the auxiliary random variable S. The overall exponent is then determined by the smaller of θ_1 and θ_2, and in particular by a choice S, X, W that makes the two equal. In this case, for a scenario with parameters $p_0 = 0.2, q_0 = 0.3, p_1 = 0.4$, and $r = 0.1$, the largest exponent achieved by our scheme is $\theta = 0.19$.

V. Proof of Theorem 1

The proof of the theorem is based on the scheme in Section III. Before analyzing this scheme, notice that by the functional representation lemma, there exists a function γ over appropriate domains and for each time $t \in \{1, \ldots, n\}$ a random variable ϕ_t over a finite alphabet Φ so that the time-t channel input and output satisfy:

$$V_t = \xi(W_t, \phi_t). \quad (38)$$

Let $\mathcal{P}^n$ be the set of all types over the product alphabets $\mathcal{S}^n \times \mathcal{S}^n \times \mathcal{W}^n \times \mathcal{W}^n \times \mathcal{W}^n \times \mathcal{V}^n \times \Phi^n \times \mathcal{X}^n \times \mathcal{Y}^n$, and let $\mathcal{P}_\mu^n$ be the subset of types $\pi_{SS'TWW'V\phi XY} \in \mathcal{P}^n$ that simultaneously satisfy the following conditions:

$$|\pi_{SX} - P_{SX}| \leq \mu/2, \quad (39a)$$

$$|\pi_{S'Y} - P_{SY}| \leq \mu, \quad (39b)$$

$$|\pi_{TW'V} - P_{TWV}| \leq \mu, \quad (39c)$$

$$\pi_{V|\phi TW} = \mathbb{1}\{V = \xi(\phi, T, W)\}, \quad (39d)$$

$$H_{\pi_{S'Y}}(S|Y) \leq H_{\pi_{SY}}(S|Y). \quad (39e)$$

We first analyze the type-I error probability averaged over the random code construction. Let (M, L) be the indices of the codeword chosen at the transmitter, if they exist, and define the following events:

$$\mathcal{E}_{\text{Tx}} \colon \{\nexists(m,\ell) \colon (S^n(m,\ell), X^n) \in \mathcal{T}^n_{\mu/2}(P_{SX})\} \quad (40)$$

$$\mathcal{E}_{\text{Rx}}^{(1)} \colon \{(S^n(M,L), Y^n) \notin \mathcal{T}^n_\mu(P_{SY})\} \quad (41)$$

$$\mathcal{E}_{\text{Rx}}^{(2)} \colon \{\exists m' \neq M \colon (T^n, W^n(m'), V^n) \in \mathcal{T}^n_\mu(P_{TW}P_{V|W})\} \quad (42)$$

$$\mathcal{E}_{\text{Rx}}^{(3)} \colon \{\exists \ell' \neq L \colon H_{\text{tp}(s^n(M,\ell'),y^n)}(S|Y) = \min_{\tilde{\ell}} H_{\text{tp}(s^n(M,\tilde{\ell}),y^n)}(S|Y)\}. \quad (43)$$

With these definitions, we obtain for all sufficiently small values of μ and sufficiently large blocklengths n:

$$\alpha_n \leq \Pr[\mathcal{E}_{\text{Tx}}] + \Pr[\mathcal{E}_{\text{Rx}}^{(1)}|\mathcal{E}_{\text{Tx}}^c] + \Pr[\mathcal{E}_{\text{Rx}}^{(2)}|\mathcal{E}_{\text{Tx}}^c, \mathcal{E}_{\text{Rx}}^{(1)c}] + \Pr[\mathcal{E}_{\text{Rx}}^{(3)}|\mathcal{E}_{\text{Rx}}^{(1)c}, \mathcal{E}_{\text{Tx}}^c] \quad (44)$$

$$\leq \epsilon/4 + \epsilon/4 + \epsilon/4 + \epsilon/4 = \epsilon, \quad (45)$$

where the first summand of (44) can be upper bounded by means of the covering lemma [10] and the rate constraint (15); the second by means of the Markov lemma [10]; the third by means of the packing lemma [10] and the rate constraint (14); and the fourth by following similar steps as in analysis of the type-I error probability in [5, Appendix H].

Now, consider the type-II error probability. Let $\mathcal{P}^n_{\mu,0}$ be the subset of types $\pi_{S'TW'\phi VXY}$ over the alphabets $\mathcal{S}^n \times \mathcal{W}^n \times \mathcal{W}^n \times \Phi^n \times \mathcal{V}^n \times \mathcal{X}^n \times \mathcal{Y}^n$ that satisfy (39b), (39c), and

$$\pi_{V|\phi T} = \mathbb{1}\{V = \xi(T, \phi)\}. \quad (46)$$

Define for each pair $(m, m') \in \{1, \ldots, \lfloor 2^{nR} \rfloor\}^2$ and $(\ell, \ell') \in \{1, \ldots, \lfloor 2^{nR'} \rfloor\}^2$ the set:

$$\mathcal{A}(m, m', \ell, \ell') := \Big\{(\varphi^n, x^n, y^n) : \text{tp}\big(S^n(m,\ell), S^n(m',\ell'), W^n(m), W^n(m'), \varphi^n, \xi^n(W^n(m), \varphi^n), x^n, y^n\big) \in \mathcal{P}^n_\mu\Big\};$$

and for each $m' \in \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ and $\ell' \in \{1, \ldots, \lfloor 2^{nR'} \rfloor\}$ the set:

$$\mathcal{A}(0, m', \ell') := \Big\{(\varphi^n, x^n, y^n) : \text{tp}\big(S^n(m',\ell'), T^n, W^n(m'), \varphi^n, \xi^n(T^n, \varphi^n), x^n, y^n\big) \in \mathcal{P}^n_{\mu,0}\Big\}. \quad (47)$$

By $\xi^n(W^n(m), \varphi^n)$, here we mean the component-wise application of the function $\xi(.,.)$ defined in (38) to the n-length sequences $W^n(m)$ and φ^n.

Define the region $\mathcal{A}_{\text{Rx},n} \subseteq \Phi^n \times \mathcal{X}^n \times \mathcal{Y}^n$

$$\mathcal{A}_{\text{Rx},n} \triangleq \bigcup_{m,m'} \bigcup_{\ell,\ell'} \mathcal{A}(m, m', \ell, \ell') \cup \bigcup_{m',\ell'} \mathcal{A}(0, m', \ell'), \quad (48)$$

where m and m' take value in $\{1, \ldots, \lfloor 2^{nR} \rfloor\}$ and ℓ and ℓ' in $\{1, \ldots, \lfloor 2^{nR'} \rfloor\}$. Notice that $\mathcal{A}_{\text{Rx},n}$ is deterministic for a given codebook, but random in the analysis here.

Since $\mathcal{A}_{\text{Rx},n}$ includes the acceptance region at the receiver, the average (over the random codebooks) type-II error probability is upper bounded as:

$$\mathbb{E}_{\mathcal{C}}[\beta_n] \leq \Pr\left[(\phi^n, X^n, Y^n) \in \mathcal{A}_{\text{Rx},n} | \mathcal{H} = 1\right]. \quad (49)$$

We can then write:

$$\mathbb{E}_{\mathcal{C}}[\beta_n]$$

$$\leq \Pr\Big[(\phi^n, X^n, Y^n) \in$$

$$\bigcup_{m,m'} \bigcup_{\ell,\ell'} \mathcal{A}(m,m',\ell,\ell') \cup \bigcup_{m',\ell'} \mathcal{A}(0,m',\ell')|\mathcal{H}=1\Big]$$
$$\le \Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{(m,\ell)} \mathcal{A}(m,m,\ell,\ell)|\mathcal{H}=1\Big]$$
$$+\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{(m,\ell)\neq(m',\ell')} \mathcal{A}(m,m',\ell,\ell')|\mathcal{H}=1\Big]$$
$$+\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{m',\ell'} \mathcal{A}(0,m',\ell')|\mathcal{H}=1\Big]. \quad (50)$$

In a similar way as in [5], it can be shown that for sufficiently large blocklengths n, the first probability in (50) is upper bounded as:

$$\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{m} \bigcup_{\ell} \mathcal{A}(m,m,\ell,\ell)|\mathcal{H}=1\Big] \le 2^{-n\theta_{1,\mu}}, \quad (51)$$

where

$$\theta_{1,\mu} := min_{\pi_{SXY}\in\mathcal{P}^n_\mu} D(\pi_{SXY}||P_{S|X}Q_{XY}) - \delta(\mu) \quad (52)$$

for a function $\delta(\mu)$ that goes to zero as $\mu\to 0$. Moreover, for sufficiently large n, the second probability in (50) is upper bounded as:

$$\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{(m,\ell)\neq(m',\ell')} \mathcal{A}(m,m',\ell,\ell')\Big] \le 2^{-n\theta_{2,\mu}}, \quad (53)$$

where

$$\theta_{2,\mu} := \min_{\pi_{SS'WW'VXY}\in\mathcal{P}^n_\mu} D(\pi_{SXY}||P_{S|X}Q_{XY})$$
$$+ I(S;Y) + I(V;W|T) - I(S;X) - \delta'(\mu), \quad (54)$$

for a function $\delta'(\mu)$ that goes to zero as $\mu\to 0$.

The last term in (50) is upper bounded for sufficiently large blocklength n:

$$\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{m',\ell'} \mathcal{A}(0,m',\ell')|\mathcal{H}=1\Big]$$
$$\le \sum_{m',\ell'} \Pr\Big[(\phi^n,X^n,Y^n)\in \mathcal{A}(0,m',\ell')|\mathcal{H}=1\Big]$$
$$\le \sum_{m',\ell'} \sum_{\pi_{S'TW'\phi VXY}\in\mathcal{P}^n_{\mu,0}}$$
$$\Pr\Big[\text{tp}\Big(S^n(m',\ell'),T^n,W^n(m'),\phi^n,\xi^n(T^n,\phi^n),X^n,Y^n\Big) = \pi_{S'TW'\phi VXY}\Big|\mathcal{H}=1\Big]$$
$$\le \sum_{m',\ell'} \sum_{\pi_{S'TW'\phi VXY}\in\mathcal{P}^n_{\mu,0}} 2^{-nD\left(\pi_{S'TW'\phi VXY}\middle|\middle|P_S P_{TW} P_\phi \pi_{V|\phi T} Q_{XY}\right)}$$

where the last inequality holds by the way the random codebooks are generated and because given $\mathcal{H}=1$, the sources X^n,Y^n are i.i.d. $\sim Q_{XY}$. Define now

$$\tilde{\theta}_{3,\mu} := \min_{\substack{\pi_{S'TW'\phi VXY}\\ \in\mathcal{P}^n_{\mu,0}}} D(\pi_{S'TW'\phi VXY}||P_S P_{TW} P_\phi \pi_{V|\phi T} Q_{XY})$$
$$- R - R' - \mu \quad (55)$$

and notice that there exist functions $\delta''(\mu)$ that $\to 0$ as $\mu\to 0$ and so that the following inequalities hold:

$$\tilde{\theta}_{3,\mu} \overset{(a)}{\ge} \min_{\pi_{S'TW'\phi VXY}\in\mathcal{P}^n_{\mu,0}} \Big[D(\pi_{TW'\phi V}||P_{TW}P_\phi\pi_{V|\phi T})$$
$$+ D(\pi_{XY}||Q_{XY}) + \mathbb{E}_{\pi_{XY}}[D(\pi_{S'|XY}||P_S)]\Big]$$
$$- I(S;X) - 2\mu$$
$$\overset{(b)}{\ge} \min_{\pi_{S'TW'VXY}\in\mathcal{P}^n_{\mu,0}} \Big[D(\pi_{TW'V}||P_{TW}P_{V|W=T})\Big]$$
$$+ D(P_Y||Q_Y) + I(S;Y) - I(S;X) - \delta''(\mu)$$
$$\overset{(c)}{\ge} \mathbb{E}_{P_{TW}}[D(P_{V|W}||P_{V|W=T})]$$
$$+ D(P_Y||Q_Y) + I(S;Y) - I(S;X) - \delta''(\mu)$$
$$= D(P_Y||Q_Y) + I(V;W|T) + I(S;Y) - I(S;X)$$
$$+ \sum_t P_T(t)\cdot D(P_{V|T=t}||P_{V|W=t}) - \delta''(\mu)$$
$$:= \theta_{3,\mu}. \quad (56)$$

All three inequalities are based on the data processing inequality for KL-divergences; (a) also uses (13); and (b) and (c) also use the continuity of KL-divergences and that all types in $\mathcal{P}^n_{0,\mu}$ satisfy (39b), (39c), and (46). Thus, for sufficiently large n:

$$\Pr\Big[(\phi^n,X^n,Y^n)\in \bigcup_{m',\ell'} \mathcal{A}(0,m',\ell')|\mathcal{H}=1\Big] \le 2^{-n\theta_{3,\mu}}. (57)$$

Combining (50), (51), (53), and (57), taking $\mu\to 0$ and $n\to\infty$, the proof can be established by standard arguments.

References

[1] A. Ahlswede and I. Csiszar, “Hypothesis testing with communication constraints,” *IEEE Trans. on Info. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.

[2] T. S. Han, “Hypothesis testing with multiterminal data compression,” *IEEE Trans. on Info. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.

[3] H. Shimokawa, T. Han and S. I. Amari, “Error bound for hypothesis testing with data compression,” in *Proc. IEEE Int. Symp. on Info. Theory*, Jul. 1994, p. 114.

[4] M. S. Rahman and A. B. Wagner, “On the Optimality of binning for distributed hypothesis testing,” *IEEE Trans. on Info. Theory*, vol. 58, no. 10, pp. 6282–6303, Oct. 2012.

[5] S. Salehkalaibar, M. Wigger and L. Wang, “Hypothesis testing over multi-hop networks,” Available at: https://arxiv.org/abs/1708.05198.

[6] W. Zhao and L. Lai, “Distributed testing against independence with multiple terminals,” in *Proc. 52nd Allerton Conf. Comm, Cont. and Comp.*, IL, USA, pp. 1246–1251, Oct. 2014.

[7] Y. Xiang and Y. H. Kim, “Interactive hypothesis testing against independence,” in *Proc. IEEE Int. Symp. on Info. Theory*, Istanbul, Turkey, pp. 2840–2844, Jun. 2013.

[8] M. Wigger and R. Timo, “Testing against independence with multiple decision centers,” in *Proc. of SPCOM*, Bangalore, India, Jun. 2016.

[9] S. Sreekuma and D. Gunduz, “Distributed hypothesis testing over noisy channels,” available at: https://arxiv.org/abs/1704.01535.

[10] A. El Gamal and Y. H. Kim, *Network information theory*, Cambridge Univ. Press, 2011.

[11] D. Wang, V. Chandar, S. Y. Chung and G. W. Wornell, “On reliability functions for single-message unequal error protection,” in *Proc. IEEE Int. Symp. on Info. Theory*, MIT, pp. 2934–2938, 2012.

[12] S. P. Borade, “When all information is not created equal,” Thesis, Massachusetts Institute of Technology, 2008.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, 1991.

Strategic Coordination with State Information at the Decoder

Maël Le Treust* and Tristan Tomala †
* ETIS UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS,
6, avenue du Ponceau, 95014 Cergy-Pontoise CEDEX, FRANCE
Email: mael.le-treust@ensea.fr
† HEC Paris, GREGHEC UMR 2959
1 rue de la Libération, 78351 Jouy-en-Josas CEDEX, FRANCE
Email: tomala@hec.fr

***Abstract*—We investigate the coordination of autonomous devices with strategic and non-aligned utility functions. The encoder and the decoder of the point-to-point network choose their coding strategy in order to maximize their own utility function. This paper extends our previous results on strategic coordination by considering state information at the decoder. We study the connexion between Wyner-Ziv source coding and the problem of Bayesian persuasion in the economics literature.**

I. Introduction

In this paper, we investigate a point-to-point network of autonomous devices with non-aligned utility functions, see Fig. 1. Our study is based on notion of "Empirical Coordination" which characterizes the global behavior that can be implemented by local policies. Coming originally from the literature of Game Theory [1], [2], [3], [4], [5], the problem of Coordination has attracted a lot of attention in Information Theory [6], [7], [8], [9], [10], [11]. It consists in determining the minimal exchange of information required by autonomous devices in order to implement a coordinated behavior. More precisely, a target joint distribution is achievable if there exists a coding scheme whose empirical distribution of symbols converges to that target distribution. Then, it is possible to optimize any utility function - instead of the distortion - by considering the one-shot version of the problem instead of the problem by blocks of n-symbols. The notion of Empirical Coordination generalizes the "Rate-Distortion Theory" as well as "Channel coding result" and is strongly related to the joint source-channel coding with state information at both encoder and decoder [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22].

In this paper, we investigate the coordinated behavior of two devices with non-aligned utility functions, in the spirit of [23], [24]. Fig. 1 corresponds to the problem of Wyner-Ziv source coding with state information at the decoder [25], with a noisy channel. The only difference is that the encoder and decoder are players endowed with distinct utility functions $\phi_1(u,v)$ and $\phi_2(u,v)$. If these utility functions were equal $\phi_1(u,v) = \phi_2(u,v)$, our solution would boil down to the classical result for noisy channel and Wyner-Ziv source.

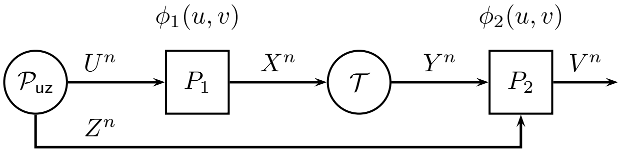


Fig. 1. The information source is i.i.d. $\mathcal{P}_{\mathsf{uz}}(u,z)$ and the channel $\mathcal{T}(y|x)$ is memoryless. The encoder P_1 and the decoder P_2 are players that maximize their own utility functions $\phi_1(u,v) \in \mathbb{R}$ and $\phi_2(u,v) \in \mathbb{R}$.

We consider a game in which the encoder and decoder are the players P_1 and P_2 that choose the encoding and decoding strategies in order to maximize their long-run utility. The equilibrium solution proposed by Stackelberg in [27] is more suited than the "Nash Equilibrium" [28], since the decoder P_2 knows in advance the encoding strategy of P_1, i.e. the encoder P_1 has "commitment power". This problem is also related to the "Strategic Transmission of Information" in the literature of Game Theory [29], [30], [31], [32], [33], [34]. In fact, our problem is closely related to the problem of "Bayesian Persuasion" [35], [36], in which a sender wants to persuade a receiver to change her action. By sending some information, the encoder is able to control the posterior beliefs of the decoder, knowing that he will choose a best-reply action. The problem of strategic communication was investigated in the literature of Information Theory [37], [38], [39], [40] for Gaussian source and channel, and the quadratic distortion functions of [29].

II. Strategic Coordination

A. Problem Statement

We consider the problem of strategic coordination depicted in Fig. 1. Notations U^n, X^n, Y^n, Z^n, V^n stand for sequences of random variables of information source $u^n = (u_1,\ldots,u_n) \in \mathcal{U}^n$, decoder's state information $z^n \in \mathcal{Z}^n$, inputs of the channel $x^n \in \mathcal{X}^n$, outputs of the channel $y^n \in \mathcal{Y}^n$ and decoder's output $v^n \in \mathcal{V}^n$, respectively. The sets $\mathcal{U}$, $\mathcal{Z}$, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{V}$ have finite cardinality. The set of probability distributions over $\mathcal{X}$ is denoted by $\Delta(\mathcal{X})$. The notation $||\mathcal{Q}-\mathcal{P}||_1 = \sum_{x\in\mathcal{X}} |\mathcal{Q}(x)-\mathcal{P}(x)|$ stands for the L_1 distance

† Tristan Tomala acknowledges financial support from the HEC foundation. The authors would like to thank Institute Henri Poincaré (IHP) in Paris France, for hosting their scientific discussions.

between the probability distributions $\mathcal{Q}$ and $\mathcal{P}$. With a slight abuse of notation, we denote by $\mathcal{Q}(x) \times \mathcal{Q}(v|x)$, the product of distributions over $\Delta(\mathcal{X} \times \mathcal{V})$. Notation $Y -\!\!\!\!\circ\!\!\!\!- X -\!\!\!\!\circ\!\!\!\!- U$ denotes the Markov chain property corresponding to $\mathcal{P}(y|x,u) = \mathcal{P}(y|x)$ for all (u, x, y). Player P_1 observes a sequence of source symbols $u^n \in \mathcal{U}^n$ and chooses at random a sequence of channel inputs $x^n \in \mathcal{X}^n$. Player P_2 observes a sequence of channel outputs $y^n \in \mathcal{Y}^n$ and state information $z^n \in \mathcal{Z}^n$ before choosing at random a sequence of actions $v^n \in \mathcal{V}^n$.

Definition II.1 (Strategies of both players) *• Player P_1 chooses a strategy σ and player P_2 chooses a strategy τ, defined as follows:*

$$\sigma : \mathcal{U}^n \longrightarrow \Delta(\mathcal{X}^n), \tag{1}$$

$$\tau : \mathcal{Y}^n \times \mathcal{Z}^n \longrightarrow \Delta(\mathcal{V}^n). \tag{2}$$

Both strategies (σ, τ) are stochastic.
• A pair of strategies (σ, τ) induces a joint probability distribution $\mathcal{P}_{\sigma,\tau} \in \Delta(\mathcal{U}^n \times \mathcal{Z}^n \times \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{V}^n)$ over the n-sequences of symbols, defined by:

$$\prod_{i=1}^{n} \mathcal{P}\Big(U_i, Z_i\Big) \times \mathcal{P}_\sigma\Big(X^n \Big| U^n\Big)$$
$$\times \prod_{i=1}^{n} \mathcal{T}\Big(Y_i \Big| X_i\Big) \times \mathcal{P}_\tau\Big(V^n \Big| Y^n, Z^n\Big). \tag{3}$$

Definition II.2 (Expected n-stage utilities) *The utilities of the n-stage game Φ_1^n and Φ_2^n are evaluated with respect to the marginal distribution $\mathcal{P}_{\sigma,\tau}$ over the sequences (U^n, V^n) and the utility functions $\phi_1(u, v) \in \mathbb{R}$, $\phi_2(u, v) \in \mathbb{R}$.*

$$\Phi_1^n(\sigma,\tau) = \mathbb{E}_{\sigma,\tau}\left[\frac{1}{n}\sum_{i=1}^{n}\phi_1(U_i, V_i)\right]$$
$$= \sum_{u^n, v^n} \mathcal{P}_{\sigma,\tau}\Big(u^n, v^n\Big) \cdot \left[\frac{1}{n}\sum_{i=1}^{n}\phi_1(u_i, v_i)\right], \tag{4}$$

$$\Phi_2^n(\sigma,\tau) = \sum_{u^n, v^n} \mathcal{P}_{\sigma,\tau}\Big(u^n, v^n\Big) \cdot \left[\frac{1}{n}\sum_{i=1}^{n}\phi_2(u_i, v_i)\right]. \tag{5}$$

Definition II.3 (Decoder's best-replies) *For any strategy σ of P_1, we define the set of n-best-reply of P_2 as follows:*

$$\mathsf{BR}_2^n(\sigma) = \Big\{\tau, \text{ s.t. } \Phi_2^n(\sigma,\tau) \geq \Phi_2^n(\sigma,\widetilde{\tau}), \ \forall \widetilde{\tau} \neq \tau\Big\}. \tag{6}$$

Definition II.4 (Characterization) *We consider an auxiliary random variable W with $|\mathcal{W}| = \min\big(|\mathcal{V}|, |\mathcal{U}|+1\big)$. We define the set $\mathbb{Q}_0$ of target probability distributions by:*

$$\mathbb{Q}_0 = \Big\{\mathcal{P}_{\mathsf{uz}}(u,z) \times \mathcal{Q}(w|u), \quad \text{s.t.},$$
$$\max_{\mathcal{P}(x)} I(X;Y) - I(U;W) + I(Z;W) \geq 0\Big\}. \tag{7}$$

We define the set $\mathbb{Q}_2\big(\mathcal{Q}(u,z,w)\big)$ of decoder's best-reply:

$$\mathbb{Q}_2\big(\mathcal{Q}(u,z,w)\big) = \mathrm{argmax}_{\mathcal{Q}(v|z,w)} \mathbb{E}_{\substack{\mathcal{Q}(u,z,w) \\ \times \mathcal{Q}(v|z,w)}}\Big[\phi_2(U,V)\Big]. \tag{8}$$

The optimal utility $\Phi_1^\star$ of P_1 is given by:

$$\Phi_1^\star = \sup_{\mathcal{Q}(u,z,w) \in \mathbb{Q}_0} \ \min_{\substack{\mathcal{Q}(v|z,w) \in \\ \mathbb{Q}_2\left(\mathcal{Q}(u,z,w)\right)}} \mathbb{E}_{\substack{\mathcal{Q}(u,z,w) \\ \times \mathcal{Q}(v|z,w)}}\Big[\phi_1(U,V)\Big]. \tag{9}$$

We prove that the n-stage game of utility $\Phi_1^n(\sigma,\tau)$ can be reformulated as a one-shot game in which the decoder chooses $\mathcal{Q}(v|z,w)$, knowing that the encoder has chosen $\mathcal{Q}(w|u)$.

Theorem II.5 (Main Result) *The limit utility of P_1 when P_2 chooses any n-best-reply $\tau \in \mathsf{BR}_2^n(\sigma)$:*

$$\forall n \in \mathbb{N}, \ \forall \sigma, \qquad \min_{\tau \in \mathsf{BR}_2^n(\sigma)} \Phi_1^n(\sigma,\tau) \leq \Phi_1^\star, \tag{10}$$

$$\forall \varepsilon > 0, \ \exists \bar{n}, \ \forall n \geq \bar{n}, \ \exists \sigma, \ \min_{\tau \in \mathsf{BR}_2^n(\sigma)} \Phi_1^n(\sigma,\tau) \geq \Phi_1^\star - \varepsilon. \tag{11}$$

The proofs of the converse (10) and achievability (11) results are stated in Sec. IV and V.

III. Example with Z-state information

The binary source U has probability $\mathcal{P}(u_1) = p_0$ with $p_0 \in [0,1]$ and the state information Z is drawn through a Z-channel with parameter $\delta \in [0,1]$ as in Fig. 2. While observing the state information Z, the decoder reactualizes his beliefs regarding the source:

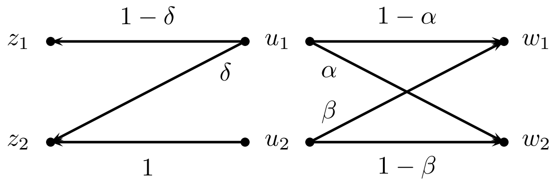


Fig. 2. Joint distribution $\mathcal{P}(u,z)$ and the signaling strategy $\mathcal{Q}(w|u)$.

$$q_1 = \mathcal{Q}(u_1|z_1) = \frac{p_0 \cdot (1-\delta)}{p_0 \cdot (1-\delta)} = 1, \tag{12}$$

$$q_2 = \mathcal{Q}(u_1|z_2) = \frac{p_0 \cdot \delta}{1 - p_0 \cdot (1-\delta)}. \tag{13}$$

We denote by (q_1, q_2) the belief ex-ante, i.e. before the transmission of W.

The binary auxiliary random variable $W \in \{w_1, w_2\}$ is drawn with distribution $\mathcal{Q}(w|u)$ and parameters $\alpha \in [0,1]$, $\beta \in [0,1]$ as in Fig. 2. After receiving the symbol W, the decoder reactualizes his posterior beliefs denoted by (p_1, p_2, p_3, p_4):

$$\mathcal{Q}(u_1|w_1,z_1) = p_1 = \mathcal{Q}(u_1|w_2,z_1) = p_3 \qquad = 1, \tag{14}$$

$$\mathcal{Q}(u_1|w_1,z_2) = \frac{p_0 \cdot (1-\alpha) \cdot \delta}{p_0 \cdot (1-\alpha) \cdot \delta + (1-p_0) \cdot \beta} = p_2, \tag{15}$$

$$\mathcal{Q}(u_1|w_2,z_2) = \frac{p_0 \cdot \alpha \cdot \delta}{p_0 \cdot \alpha \cdot \delta + (1-p_0) \cdot (1-\beta)} = p_4. \tag{16}$$

(15)-(16) reformulate into the signaling strategy $\mathcal{Q}(w|u)$:

$$\alpha = \frac{p_4 \cdot \big(p_2 \cdot (1 - p_0(1-\delta)) - p_0 \cdot \delta\big)}{p_0 \cdot \delta \cdot (p_2 - p_4)}, \tag{17}$$

$$\beta = \frac{(1-p_2) \cdot \big(p_0 \cdot \delta - p_4 \cdot (1 - p_0(1-\delta))\big)}{(p_2 - p_4) \cdot (1-p_0)}. \tag{18}$$

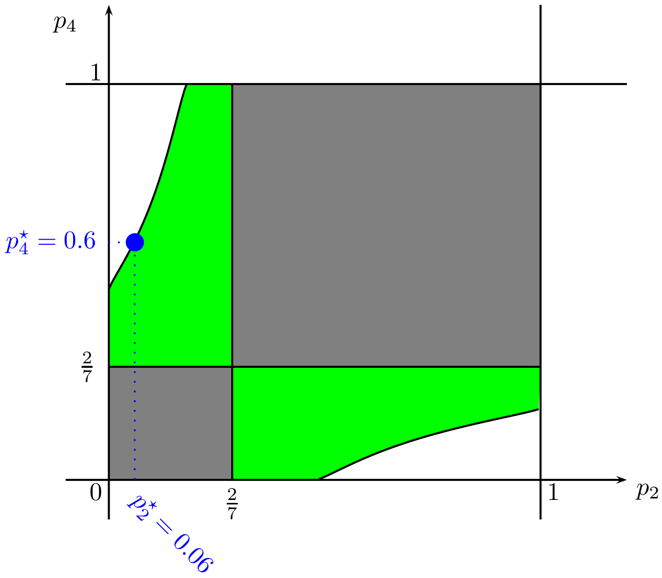


Fig. 3. Regions of posterior beliefs (p_2, p_4) satisfying information constraint $C - I(U;W) + I(Z;W) \geq 0$ for $p_0 = 0.5$, $C = 1 - H(0.25)$, $\delta = 0.4$. The threshold $\frac{2}{7}$ corresponds to the beliefs ex-ante q_2 given by (13), induced by the symbol z_2.

Lemma 1 *A pair of posterior beliefs* (p_2, p_4) *is feasible if and only if* $p_2 < q_2 < p_4$ *or* $p_4 < q_2 < p_2$.

	v_1	v_2
u_1	10	4
u_2	0	9

Fig. 4. Utility $\phi_2(u, v)$ of P_2

	v_1	v_2
u_1	1	0
u_2	1	0

Fig. 5. Utility $\phi_1(u, v)$ of P_1

The proof of Lemma 1 comes from the constraints $\alpha \in [0,1]$, $\beta \in [0,1]$ in (17) and (18). The pair of posterior beliefs (p_2, p_4) cannot belongs to the grey regions of Fig. 3. We consider that the channel capacity is fixed equal to $C = \max_{\mathcal{P}(x)} I(X;Y) = 1 - H(0.25)$. The information constraint of $\mathbb{Q}_0$ writes:

$$C - I(W;U) + I(W;Z) \geq 0 \tag{19}$$

$$\Longleftrightarrow H(U|W,Z) \geq H(U|Z) - C \tag{20}$$

$$\Longleftrightarrow \mathcal{P}(w_1, z_2)H(p_2) + \mathcal{P}(w_2, z_2)H(p_4) \geq H(U|Z) - C \tag{21}$$

$$\Longleftrightarrow \frac{p_0 \cdot \delta - p_4 \cdot (1 - p_0(1-\delta))}{p_2 - p_4} H(p_2) + \frac{p_2 \cdot (1 - p_0(1-\delta)) - p_0 \cdot \delta}{p_2 - p_4} H(p_4) \geq H(U|Z) - C. \tag{22}$$

The green regions of Fig. 3 represent the pairs of posterior beliefs (p_2, p_4) that satisfy the information constraint. We consider the utility functions of the encoder $\phi_1(u, v)$ and of the decoder $\phi_2(u, v)$, given by Fig. 4 and 5. The player P_2 holds a belief $\mathcal{P}(u_1)$ regarding the source of information U.

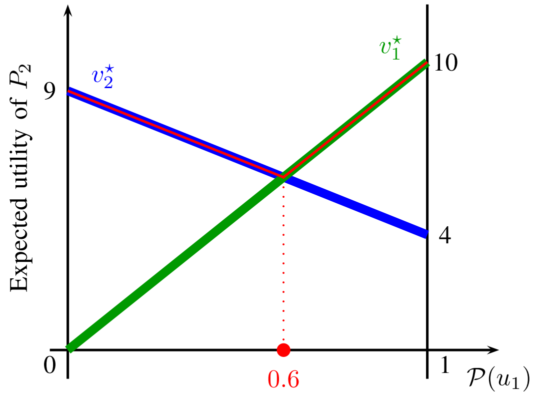


Fig. 6. The best-reply action $v^\star$ of P_2 depends on his belief $\mathcal{P}(u)$ regarding the source U: if $\mathcal{P}(u_1) \in [0, 0.6]$ he plays $v_2^\star$ and if $\mathcal{P}(u_1) \in [0.6, 1]$ he plays $v_1^\star$.

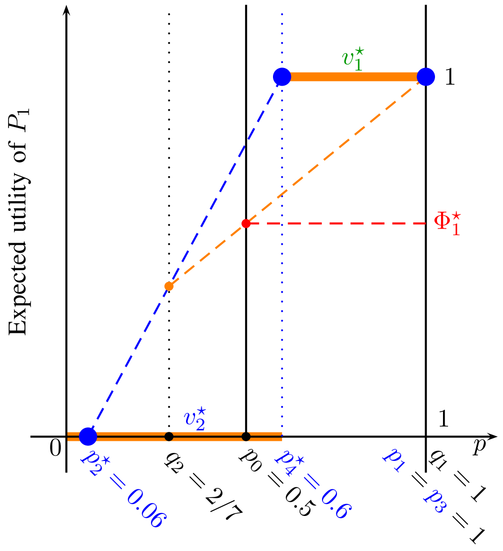


Fig. 7. The expected utility of player P_1 depending on the belief $\mathcal{P}(u_1)$ of player P_2. The optimal posterior beliefs $(p_2^\star, p_4^\star)$ satisfy the information constraint $C - I(U;W) + I(Z;W) \geq 0$, for $p_0 = 0.5$, $C = 1 - H(0.25)$, $\delta = 0.4$.

He chooses a best-reply action $v_1^\star$ or $v_2^\star$ depending on the interval $[0, 0.6]$ or $[0.6, 1]$ in which lies the belief $\mathcal{P}(u_1)$, see Fig. 6. The utility of player P_1 only depends on the action of player P_2 and is represented by the orange line in Fig. 7. The encoder would like to send some information in order to modify the posterior beliefs of P_2 such that p_4 belongs to the interval $p_4 \in [0.6, 1]$. Then the best-reply action of P_2 would be $v_1^\star$ that rewards player P_1. The optimal solution is to fix $p_4^\star = 0.6$ and to find $p_2^\star$ that satisfies the information constraint with equality:

$$C - I(W;U) + I(W;Z) = 0 \tag{23}$$

$$\begin{aligned}\Longleftrightarrow & \frac{p_0 \cdot \delta - p_4^\star \cdot (1 - p_0(1-\delta))}{p_2^\star - p_4^\star} H(p_2^\star) \\ & + \frac{p_2^\star \cdot (1 - p_0(1-\delta)) - p_0 \cdot \delta}{p_2^\star - p_4^\star} H(p_4^\star) = H(U|Z) - C.\end{aligned} \tag{24}$$

Hence the optimal solution is $p_2^\star \simeq 0.06$ and the pair $(p_2^\star, p_4^\star)$ lies at the border of the green region of Fig. 3. This solution induces the conditional entropies $H(U|W, z_2) \simeq 0.7146$ and $H(U|W,Z) = H(U|Z) - C \simeq 0.5747$. The optimal utility for player P_1 is $\Phi_1^\star \simeq 0.5925$.

IV. Converse Proof of Theorem II.5

We consider an arbitrary strategy σ of length $n \in \mathbb{N}$. We denote by T the uniform random variable $\{1, \ldots, n\}$ and we introduce the auxiliary random variable $W = (Y^n, Z^{-T}, T)$ whose joint probability distribution $\mathcal{P}_\sigma(u,z,w)$ with (U,Z) is defined by:

$$\begin{aligned}\mathcal{P}_\sigma(u,z,w) =& \mathcal{P}_\sigma\big(u_T, z_T, y^n, z^{-T}, T\big) \\ =& \mathcal{P}(T = i) \cdot \mathcal{P}_\sigma\big(u_T, z_T, y^n, z^{-T} \big| T = i\big) \\ =& \frac{1}{n} \cdot \mathcal{P}_\sigma\big(u_i, z_i, y^n, z^{-i}\big).\end{aligned} \tag{25}$$

This identification ensures that the Markov chain $W -\!\!\!\circ\!\!\!- U_T -\!\!\!\circ\!\!\!- Z_T$ is satisfied. We now prove that the distribution $\mathcal{P}(u,z,w)$ satisfies the information constraint of the set $\mathbb{Q}_0$.

$$\begin{aligned}0 \leq & I(X^n;Y^n) - I(U^n, Z^n; Y^n) && (26)\\ \leq & \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) - I(U^n;Y^n|Z^n) && (27)\\ \leq & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - \sum_{i=1}^n I(U_i; Y^n | Z^n, U^{i-1}) && (28)\\ = & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - \sum_{i=1}^n I(U_i; Y^n, Z^{-i}, U^{i-1} | Z_i) && (29)\\ \leq & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - \sum_{i=1}^n I(U_i; Y^n, Z^{-i} | Z_i) && (30)\\ = & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - n \cdot I(U_T; Y^n, Z^{-T} | Z_T, T) && (31)\\ = & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - n \cdot I(U_T; Y^n, Z^{-T}, T | Z_T) && (32)\\ = & n \cdot \max_{\mathcal{P}(x)} I(X;Y) - n \cdot I(U; W | Z) && (33)\\ = & n \cdot \bigg(\max_{\mathcal{P}(x)} I(X;Y) - I(U;W) + I(Z;W) \bigg). && (34)\end{aligned}$$

(26) comes from the Markov chain $Y^n -\!\!\!\circ\!\!\!- X^n -\!\!\!\circ\!\!\!- (U^n, Z^n)$.
(27) comes from the memoryless property of the channel and from removing the positive term $I(U^n; Z^n) \geq 0$.
(28) comes from taking the maximum $\mathcal{P}(x)$ and chain rule.
(29) comes from the i.i.d. property of the source (U,Z) that implies $I(U_i, Z_i; Z^{-i}, U^{i-1}) = I(U_i; Z^{-i}, U^{i-1} | Z_i) = 0$.
(30) comes from removing $I(U_i; U^{i-1} | Y^n, Z^{-i}, Z_i) \geq 0$.
(31) comes from the uniform random variable $T \in \{1, \ldots, n\}$.
(32) comes from the independence between T and the source (U,Z), that implies $I(U_T, Z_T; T) = I(U_T; T | Z_T) = 0$.
(33) comes from the identification $W = (Y^n, Z^{-T}, T)$.
(34) comes from the Markov chain $W -\!\!\!\circ\!\!\!- U_T -\!\!\!\circ\!\!\!- Z_T$. This proves that the distribution $\mathcal{P}_\sigma(u,z,w)$ belongs to the set $\mathbb{Q}_0$.

For any strategies (σ, τ), we reformulate the long-run utilities with the auxiliary random variable $W = (Y^n, Z^{-T}, T)$.

$$\begin{aligned}&\Phi_1^n(\sigma,\tau)\\ =& \sum_{u^n, z^n, y^n} \mathcal{P}_\sigma(u^n, z^n, y^n) \sum_{v^n} \mathcal{P}_\tau(v^n | y^n, z^n) \cdot \frac{1}{n} \sum_{i=1}^n \phi_1(u_i, v_i) && (35)\\ =& \sum_{i=1}^n \sum_{\substack{u_i, z_i,\\ z^{-i}, y^n}} \frac{1}{n} \cdot \mathcal{P}_\sigma(u_i, z^n, y^n) \sum_{v_i} \mathcal{P}_\tau(v_i | y^n, z^n) \cdot \phi_1(u_i, v_i) && (36)\\ =& \sum_{\substack{u_i, z_i, z^{-i},\\ y^n, i}} \mathcal{P}_\sigma(u_i, z_i, z^{-i}, y^n, i) \sum_{v_i} \mathcal{P}_\tau(v_i | y^n, z^{-i}, z_i, i) \cdot \phi_1(u_i, v_i) && (37)\\ =& \sum_{u,z,w} \mathcal{P}_\sigma(u,z,w) \sum_v \mathcal{P}_\tau(v | w, z) \cdot \phi_1(u,v). && (38)\end{aligned}$$

(35) - (37) are reformulations valid also for $\phi_2(u,v)$.
(38) comes from replacing the random variables (Y^n, Z^{-T}, T) by W, whose distribution is stated in (25).

By replacing $W = (Y^n, Z^{-T}, T)$, the set of n-best-reply $\mathsf{BR}_2^n(\sigma)$ is equal to the set $\mathbb{Q}_2\big(\mathcal{P}_\sigma(u,z,w)\big)$:

$$\begin{aligned}&\mathsf{BR}_2^n(\sigma)\\ =& \operatorname{argmax}_{\mathcal{P}_\tau(v^n|y^n,z^n)} \sum_{\substack{u^n, z^n,\\ x^n, y^n}} \mathcal{P}_\sigma(u^n, z^n, x^n, y^n) \\ & \times \sum_{v^n} \mathcal{P}_\tau(v^n | y^n, z^n) \cdot \frac{1}{n} \sum_{i=1}^n \phi_2(u_i, v_i) && (39)\\ =& \operatorname{argmax}_{\mathcal{P}_\tau(v|w,z)} \sum_{u,z,w} \mathcal{P}_\sigma(u,z,w) \sum_v \mathcal{P}_\tau(v|w,z) \cdot \phi_2(u,v) && (40)\\ =& \mathbb{Q}_2\big(\mathcal{P}_\sigma(u,z,w)\big). && (41)\end{aligned}$$

We conclude the proof of (10) in Theorem II.5.

$$\begin{aligned}&\min_{\tau \in \mathsf{BR}_2^n(\sigma)} \Phi_1^n(\sigma,\tau) = \min_{\tau \in \mathsf{BR}_2^n(\sigma)} \sum_{u^n, z^n, y^n} \mathcal{P}_\sigma(u^n, z^n, y^n) \\ & \times \sum_{v^n} \mathcal{P}_\tau(v^n | y^n, z^n) \cdot \frac{1}{n} \sum_{i=1}^n \phi_1(u_i, v_i) && (42)\\ =& \min_{\substack{\mathcal{P}_\tau(v|w,z) \in \\ \mathbb{Q}_2\left(\mathcal{P}_\sigma(u,z,w)\right)}} \sum_{u,z,w} \mathcal{P}_\sigma(u,z,w) \sum_v \mathcal{P}_\tau(v|w,z) \cdot \phi_1(u,v) && (43)\\ \leq & \sup_{\mathcal{Q}(u,z,w) \in \widetilde{\mathbb{Q}}_0} \min_{\substack{\mathcal{Q}(v|w,z) \in \\ \mathbb{Q}_2\left(\mathcal{Q}(u,z,w)\right)}} \mathbb{E}_{\mathcal{Q}(u,z,w) \times \mathcal{Q}(v|w,z)} \Big[\phi_1(U,V) \Big] = \Phi_1^\star. && (44)\end{aligned}$$

(42) comes from the definitions.
(43) comes from (41) that identifies the set of n-best-reply $\mathsf{BR}_2^n(\sigma)$ with the set $\mathbb{Q}_2\big(\mathcal{P}_\sigma(u,z,w)\big)$ of definition II.4.

(44) comes from (34) that shows the distribution $\mathcal{P}_\sigma(u,z,w)$ belongs to the set $\widetilde{\mathbb{Q}}_0$.

The cardinality bound $|\mathcal{W}| = \min\big(|\mathcal{V}|, |\mathcal{U}|+1\big)$ follows from Caratheodory's Lemma and Markov chain $Z -\!\!\circ\!\!- U -\!\!\circ\!\!- W$.

V. Sketch of Achievability of Theorem II.5

We denote by $\mathcal{Q}(u,z,w) \in \mathbb{Q}_0$ the distribution that is optimal for (9). We consider the concatenation of Wyner-Ziv source coding [25] with a channel code [26]. Empirical Coordination results [14] ensures that the sequences of symbols (U^n, Z^n, W^n) are jointly typical for $\mathcal{Q}(u,z,w)$ with large probability. Following the same lines as in [23], [24], we prove that the beliefs $\mathcal{P}_\sigma(u_i|y^n,z^n)$ induced by the strategy σ are close to the target belief $\mathcal{Q}(u|z,w)$:

$$\mathbb{E}_\sigma\left[\frac{1}{n}\sum_{i=1}^{n} D\bigg(\mathcal{P}_\sigma(U_i|Y^n,Z^n)\bigg|\bigg|\mathcal{Q}(U_i|W_i,Z_i)\bigg)\right] \leq \varepsilon. \quad (45)$$

This provides a lower bound on the utility of P_1:

$$\min_{\tau\in \mathsf{BR}_2^n(\sigma)} \Phi_1^n(\sigma,\tau) \geq \Phi_1^\star - \varepsilon. \quad (46)$$

The full version of the proof is in [41].

References

[1] O. Gossner, P. Hernandez, and A. Neyman, "Optimal use of communication resources," *Econometrica*, vol. 74, pp. 1603–1636, Nov. 2006.

[2] O. Gossner and N. Vieille, "How to play with a biased coin?," *Games and Economic Behavior*, vol. 41, no. 2, pp. 206–226, 2002.

[3] O. Gossner and T. Tomala, "Secret correlation in repeated games with imperfect monitoring," *Mathematics of Operation Research*, vol. 32, no. 2, pp. 413–424, 2007.

[4] O. Gossner and T. Tomala, "Empirical distributions of beliefs under imperfect observation," *Mathematics of Operation Research*, vol. 31, no. 1, pp. 13–30, 2006.

[5] O. Gossner, R. Laraki, and T. Tomala, "Informationally optimal correlation," *Mathematical Programming*, vol. 116, no. 1-2, pp. 147–172, 2009.

[6] P. Cuff, H. Permuter, and T. Cover, "Coordination capacity," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4181–4206, 2010.

[7] G. Kramer and S. Savari, "Communicating probability distributions," *Information Theory, IEEE Transactions on*, vol. 53, no. 2, pp. 518 – 525, 2007.

[8] P. Cuff and C. Schieler, "Hybrid codes needed for coordination over the point-to-point channel," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pp. 235–239, Sept 2011.

[9] M. Le Treust, A. Zaidi, and S. Lasaulce, "An achievable rate region for the broadcast wiretap channel with asymmetric side information," *IEEE Proc. of the 49th Allerton conference, Monticello, Illinois*, pp. 68 – 75.

[10] P. Cuff, "Distributed channel synthesis," *Information Theory, IEEE Transactions on*, vol. 59, pp. 7071–7096, Nov 2013.

[11] M. Le Treust, "Joint empirical coordination of source and channel," *IEEE Transactions on Information Theory*, vol. 63, pp. 5087–5114, Aug 2017.

[12] M. Le Treust, "Correlation between channel state and information source with empirical coordination constraint," in *IEEE Information Theory Workshop (ITW)*, pp. 272–276, Nov 2014.

[13] Z. Goldfeld, H. Permuter, and G. Kramer, "The ahlswede-körner coordination problem with one-sided encoder cooperation," in *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 1341–1345, June 2014.

[14] M. Le Treust, "Empirical coordination with two-sided state information and correlated source and state," in *IEEE International Symposium on Information Theory (ISIT)*, 2015.

[15] M. Le Treust, "Empirical coordination with channel feedback and strictly causal or causal encoding," in *IEEE International Symposium on Information Theory (ISIT)*, 2015.

[16] R. Blasco-Serrano, R. Thobaben, and M. Skoglund, "Polar codes for coordination in cascade networks," *in Proc. of the International Zurich Seminar on Communication, Zurich, Switzerland*, pp. 55 – 58, March 2012.

[17] B. Larrousse, S. Lasaulce, and M. Wigger, "Coordinating partially-informed agents over state-dependent networks," *IEEE Information Theory Workshop (ITW)*, 2015.

[18] M. Abroshan, A. Gohari, and S. Jaggi, "Zero error coordination," in *Information Theory Workshop - Fall (ITW), 2015 IEEE*, pp. 202–206, Oct 2015.

[19] R. Chou, M. Bloch, and J. Kliewer, "Polar coding for empirical and strong coordination via distribution approximation," in *Information Theory Proceedings (ISIT), 2015 IEEE International Symposium on*, June 2015.

[20] R. Chou, M. Bloch, and J. Kliewer, "Empirical and strong coordination via soft covering with polar codes," *submitted to IEEE Transactions on Information Theory, http://arxiv.org/abs/1608.08474*, 2016.

[21] M. Le Treust and M. Bloch, "Empirical coordination, state masking and state amplification: Core of the decoder's knowledge," *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2016.

[22] G. Cervia, L. Luzzi, M. R. Bloch, and M. Le Treust, "Polar coding for empirical coordination of signals and actions over noisy channels," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, 2016.

[23] M. Le Treust and T. Tomala, "Information design for strategic coordination of autonomous devices with non-aligned utilities," *IEEE Proc. of the 54th Allerton conference, Monticello, Illinois*, pp. 233–242, 2016.

[24] M. Le Treust and T. Tomala, "Persuasion with limited communication ressources," *draft, https://arxiv.org/abs/1711.04474*, 2017.

[25] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–11, Jan. 1976.

[26] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[27] H. von Stackelberg, *Marketform und Gleichgewicht*. Oxford University Press, 1934.

[28] J. Nash, "Non-cooperative games," *Annals of Mathematics*, vol. 54, pp. 286–295, 1951.

[29] V. P. Crawford and J. Sobel, "Strategic Information Transmission," *Econometrica*, vol. 50, no. 6, pp. 1431–1451, 1982.

[30] F. Forges, "Non-zero-sum repeated games and information transmission," *in: N. Meggido, Essays in Game Theory in Honor of Michael Maschler, Springer-Verlag*, no. 6, pp. 65–95, 1994.

[31] R. Laraki, "The splitting game and applications," *International Journal of Game Theory*, vol. 30, pp. 359–376, 2001.

[32] J. Renault, E. Solan, and N. Vieille, "Optimal dynamic information provision," *http://www.lse.ac.uk/statistics/events/2015-16-Seminar-Series/Optimal-Dynamic-Information-Provision.pdf*, February 2016.

[33] J. Ely, "Beeps," *Manuscript, Department of Economics, Northwestern University*, 2015.

[34] M. O. Jackson and H. F. Sonnenschein, "Overcoming incentive constraints by linking decisions," *Econometrica*, vol. 75, pp. 241 – 257, January 2007.

[35] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *American Economic Review*, vol. 101, pp. 2590 – 2615, 2011.

[36] M. Gentzkow and E. Kamenica, "Costly persuasion," *American Economic Review*, vol. 104, pp. 457 – 462, 2014.

[37] E. Akyol, C. Langbort, and T. Başar, "Strategic compression and transmission of information," in *Information Theory Workshop - Fall (ITW), 2015 IEEE*, pp. 219–223, Oct 2015.

[38] E. Akyol, C. Langbort, and T. Başar, "On the role of side information in strategic communication," in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1626–1630, July 2016.

[39] S. Sarıtaş, S. Yüksel, and S. Gezici, "Dynamic signaling games under nash and stackelberg equilibria," in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1631–1635, July 2016.

[40] E. Akyol, C. Langbort, and T. Başar, "Information-theoretic approach to strategic communication as a hierarchical game," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 205–218, 2017.

[41] M. Le Treust and T. Tomala, "Persuasion with partial observation at the decoder," *Technical Report*, 2017.

Distributed Information Bottleneck Method for Discrete and Gaussian Sources

Iñaki Estella Aguerri † Abdellatif Zaidi †‡

† Mathematics and Algorithmic Sciences Lab. France Research Center, Huawei Technologies, Boulogne-Billancourt, 92100, France
‡ Université Paris-Est, Champs-sur-Marne, 77454, France
{inaki.estella@huawei.com, abdellatif.zaidi@u-pem.fr}

***Abstract*—We study the problem of distributed information bottleneck, in which multiple encoders separately compress their observations in a manner such that, collectively, the compressed signals preserve as much information as possible about another signal. The model generalizes Tishby's centralized information bottleneck method to the setting of multiple distributed encoders. We establish single-letter characterizations of the information-rate region of this problem for both i) a class of discrete memoryless sources and ii) memoryless vector Gaussian sources. Furthermore, assuming a sum constraint on rate or complexity, for both models we develop Blahut-Arimoto type iterative algorithms that allow to compute optimal information-rate trade-offs, by iterating over a set of self-consistent equations.**

I. Introduction

The information bottleneck (IB) method was introduced by Tishby [1] (see [2] for an earlier equivalent formulation of the IB problem) as an information-theoretic principle for extracting the relevant information that some signal $Y \in \mathcal{Y}$ provides about another one, $X \in \mathcal{X}$, that is of interest. The approach has found remarkable applications in supervised and unsupervised learning problems such as classification, clustering and prediction, wherein one is interested in extracting the relevant features, i.e., X, of the available data Y [3], [4]. Perhaps key to the analysis, and development, of the IB method is its elegant connection with information-theoretic rate-distortion problems. Recent works show that this connection turns out to be useful also for a better understanding of deep neural networks [5]. Other connections, that are more intriguing, exist also with seemingly unrelated problems such as hypothesis testing [6] or systems with privacy constraints [7].

Motivated by applications of the IB method to settings in which the relevant features about X are to be extracted from separately encoded signals, we study the model shown in Figure 1. Here, X is the signal to be predicted and $(Y_1, \dots, Y_K)$ are correlated signals that could each be relevant to extract one or more features of X. The features could be distinct or redundant. We make the assumption that the signals $(Y_1, \dots, Y_K)$ are independent given X. This assumption holds in many practical scenarios. For example, the reader may think of $(Y_1, ..., Y_K)$ as being the results of K clinical tests that are performed independently at different clinics and are used to diagnose a disease X. A third party (decoder or detector) has to decide without access to the original data. In general, at every encoder k there is a tension among the *complexity* of the encoding, measured by the minimum description length or rate R_k at which the observation is compressed, and the information that the produced description, say U_k, provides about the signal X. The *relevance* of $(U_1, \dots, U_K)$ is measured in terms of the information that the descriptions collectively preserve about X; and is captured by Shannon's mutual information $I(U_1, \dots, U_K; X)$. Thus, the performance of the entire system can be evaluated in terms of the tradeoff between the vector $(R_1, \dots, R_K)$ of minimum description lengths and the mutual information $I(U_1, \dots, U_K; X)$.

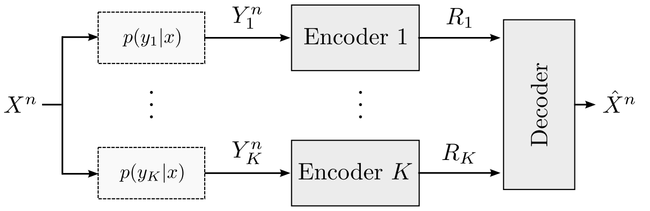


Fig. 1. A model for distributed information bottleneck (D-IB).

In this paper, we study the aforementioned tradeoff among relevant information and complexity for the model shown in Figure 1. First, we establish a single-letter characterization of the information-rate region of this model for discrete memoryless sources. In doing so, we exploit its connection with the distributed Chief Executive Officer (CEO) source coding problem under logarithmic-loss distortion measure studied in [8]. Next, we extend this result to memoryless vector Gaussian sources. Here, we prove that Gaussian test channels are optimal, thereby generalizing a similar result of [9] and [10] for the case of a single encoder IB setup.

In a second part of this paper, assuming a sum constraint on rate or complexity, we develop Blahut-Arimoto [11] type iterative algorithms that allow to compute optimal tradeoffs between information and rate, for both discrete and vector Gaussian models. We do so through a variational formulation that allows the determination of the set of self-consistent equations satisfied by the stationary solutions. In the Gaussian case, the algorithm reduces to an appropriate updating of the parameters of noisy linear projections. Here as well, our algorithms can be seen as generalizations of those developed for the single-encoder IB method, for discrete sources in [1] and for Gaussian sources in [10]; as well as a generalization of the Blahut-Arimoto algorithm proposed in [12] for the CEO source coding problem for $K = 2$ and discrete sources, to $K \geq 2$ encoders and for both discrete and Gaussian sources.

Notation: Upper case letters denote random variables, e.g., X; lower case letters denote realizations of random variables, e.g., x; and calligraphic letters denote sets, e.g., $\mathcal{X}$. The cardinality of a set is denoted by $|\mathcal{X}|$. For a random variable X with probability mass function (pmf) P_X, we use $P_X(x) = p(x)$, $x \in \mathcal{X}$ for short. Boldface upper case letters denote vectors or matrices, e.g., $\mathbf{X}$, where context makes the distinction clear. For an integer $n \in \mathbb{N}$, we denote the set $[1, n] := \{1, 2, \dots, n\}$. We denote by $D_{\mathrm{KL}}(P, Q)$ the Kullback-Leibler divergence between the pmfs P and Q. For a set of integers $\mathcal{K} \subseteq \mathbb{N}$, $X_{\mathcal{K}}$ denotes the set $X_{\mathcal{K}} = \{X_k : k \in \mathcal{K}\}$. For a zero-mean vector $\mathbf{X}$ we define the matrices $\mathbf{\Sigma_x} := \mathrm{E}[\mathbf{X}\mathbf{X}^H]$; $\mathbf{\Sigma_{x,y}} := \mathrm{E}[\mathbf{X}\mathbf{Y}^H]$, and $\mathbf{\Sigma_{x|y}} := \mathbf{\Sigma_x} - \mathbf{\Sigma_{x,y}}\mathbf{\Sigma_y}^{-1}\mathbf{\Sigma_{y,x}}$.

II. System Model

Consider the discrete memoryless D-IB model shown in Figure 1. Let $\{X_i, Y_{1,i}, \dots, Y_{K,i}\}_{i=1}^n = (X^n, Y_1^n, \dots, Y_K^n)$ be a sequence of n independent, identically distributed (i.i.d.) random variables with finite alphabets $\mathcal{X}, \mathcal{Y}_k$, $k \in \mathcal{K} := \{1, \dots, K\}$ and joint pmf $P_{X,Y_1,\dots,Y_K}$. Throughout this paper, we make the assumption that the observations at the encoders are independent conditionally on X, i.e.,

$$Y_{k,i} -\!\!\circ\!\!- X_i -\!\!\circ\!\!- Y_{\mathcal{K}/k,i} \quad \text{for } k \in \mathcal{K} \text{ and } i \in [1,n]. \tag{1}$$

Encoder $k \in \mathcal{K}$ maps the observed sequence Y_k^n to an index $J_k := \phi_k(Y_k^n)$, where $\phi_k : \mathcal{Y}_k^n \to \mathcal{M}_k$ is a given map and $\mathcal{M}_k := [1, M_k^{(n)}]$. The index J_k is sent error-free to the decoder. The decoder collects all indices $J_{\mathcal{K}} := (J_1, \dots, J_K)$ and then estimates the source X^n as $\hat{X}^n = g^{(n)}(J_{\mathcal{J}})$, where $g^{(n)} : \mathcal{M}_1 \times \dots \times \mathcal{M}_L \to \hat{\mathcal{X}}^n$ is some decoder map and $\hat{\mathcal{X}}^n$ is the reconstruction alphabet of the source.

The quality of the reconstruction is measured in terms of the n-letter *relevant information* between the unobserved source X^n and its reconstruction at the decoder $\hat{X}^n$, given by

$$\Delta^{(n)} := \frac{1}{n} I(X^n; g^{(n)}(\phi_1^{(n)}(Y_1^n), \dots, \phi_K^{(n)}(Y_K^n))). \tag{2}$$

Definition 1. *A tuple $(\Delta, R_1, \dots, R_K)$ is said to be achievable for the D-IB model if there exists a blocklength n, encoder maps $\phi_k^{(n)}$ for $k \in \mathcal{K}$, and a decoder map $g^{(n)}$, such that*

$$R_k \geq \frac{1}{n} \log M_k^{(n)},\ k \in \mathcal{K}, \ \ \text{and} \ \ \Delta \leq \frac{1}{n} I(X^n; \hat{X}^n). \tag{3}$$

where $\hat{X}^n = g^{(n)}(\phi_1^{(n)}(Y_1^n), \dots, \phi_K^{(n)}(Y_K^n))$. The information-rate region $\mathcal{R}_{\text{IB}}$ is given by the closure of all achievable rates tuples $(\Delta, R_1, \dots, R_K)$.

We are interested in characterizing the region $\mathcal{R}_{\text{IB}}$. Due to space limitations, some results are only outlined or provided without proof. We refer to [13] for a detailed version.

III. Information-Rate Region Characterization

In this section we characterize the information-rate region $\mathcal{R}_{\text{IB}}$ for a discrete memoryless D-IB model. It is well known that the IB problem is essentially a source-coding problem where the distortion measure is of logarithmic loss type [14]. Likewise, the D-IB model of Figure 1 is essentially a K-encoder CEO source coding problem under logarithmic loss (log-loss) distortion measure. The log-loss distortion between sequences is defined as

$$d_{\text{LL}}(x^n, \hat{x}^n) := -\frac{1}{n} \log \left(\frac{1}{\hat{x}^n(x^n)} \right), \tag{4}$$

where $\hat{x}^n = s(x^n | j_{\mathcal{K}})$ and s is a pmf on $\mathcal{X}^n$.

The rate-distortion region of the K-encoder CEO source coding problem under log-loss, with $K \geq 2$ which we denote hereafter as $\mathcal{RD}_{\text{CEO}}$, has been established recently in [8, Theorem 10] for the case in which the Markov chain (1) holds.

We first state the following proposition, the proof of which is easy and omitted for brevity.

Proposition 1. *A tuple $(\Delta, R_1, \dots, R_K) \in \mathcal{R}_{\text{IB}}$ if and only if $(H(X) - \Delta, R_1, \dots, R_K) \in \mathcal{RD}_{\text{CEO}}$.*

Proposition 1 implies that [8, Theorem 10] can be applied to characterize the information-rate region $\mathcal{R}_{\text{IB}}$ as given next.

Theorem 1. *In the case in which the Markov chain (1) holds, the rate-information region $\mathcal{R}_{\text{IB}}$ of the D-IB model is given by the set of all tuples $(\Delta, R_1, \dots, R_K)$ which satisfy for $\mathcal{S} \subseteq \mathcal{K}$*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(Y_k; U_k | X, Q)] + I(X; U_{\mathcal{S}^c} | Q), \tag{5}$$

for some joint pmf $p(q)p(x) \prod_{k=1}^K p(y_k|x) \prod_{k=1}^K p(u_k|y_k, q)$.

A. Memoryless Vector Gaussian D-IB

Consider now the following memoryless vector Gaussian D-IB problem. In this model, the source vector $\mathbf{X} \in \mathbb{C}^N$ is Gaussian and has zero mean and covariance matrix $\mathbf{\Sigma_x}$, i.e., $\mathbf{X} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma_x})$. Encoder k, $k \in \mathcal{K}$, observes a noisy observation $\mathbf{Y}_k \in \mathbb{C}^{M_k}$, that is given by

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X} + \mathbf{N}_k, \tag{6}$$

where $\mathbf{H}_k \in \mathbb{C}^{M_k \times N}$ is the channel connecting the source to encoder k, and $\mathbf{N}_k \in \mathbb{C}^{M_k}$, $k \in \mathcal{K}$, is the noise vector at encoder k, assumed to be Gaussian, with zero-mean and covariance matrix $\mathbf{\Sigma}_{\mathbf{n}_k}$, and independent from all other noises and the source vector $\mathbf{X}$.

The studied Gaussian model satisfies the Markov chain (1); and thus, the result of Theorem 1, which can be extended to continuous sources using standard techniques, characterizes the information-rate region of this model, denoted by $\mathcal{R}_{\text{IB}}^{\text{G}}$. Next theorem characterizes $\mathcal{R}_{\text{IB}}^{\text{G}}$, shows that the optimal test channels $P_{U_k|Y_k}$, $k \in \mathcal{K}$, are Gaussian and that there is not need for time-sharing, i.e., $Q = \emptyset$.

Theorem 2. *If $(\mathbf{X}, \mathbf{Y}_1, \dots, \mathbf{Y}_K)$ are jointly Gaussian as in* (6)*, the information-rate region $\mathcal{R}_{\text{IB}}^{\text{G}}$ is given by the set of all tuples $(\Delta, R_1, \dots, R_L)$ satisfying that for all $\mathcal{S} \subseteq \mathcal{K}$*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k + \log |\mathbf{I} - \mathbf{B}_k|] + \log \left| \sum_{k \in \mathcal{S}^c} \bar{\mathbf{H}}_k^H \mathbf{B}_k \bar{\mathbf{H}}_k + \mathbf{I} \right|,$$

for some $\mathbf{0} \preceq \mathbf{B}_k \preceq \mathbf{I}$ and where $\bar{\mathbf{H}}_k = \mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2} \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}}^{1/2}$. In addition, the information-rate tuples in $\mathcal{R}_{\text{IB}}^{\text{G}}$ are achievable with $Q = \emptyset$ and $p^*(\mathbf{u}_k | \mathbf{y}_k, q) = \mathcal{CN}(\mathbf{y}_k, \Sigma_{\mathbf{n}_k}^{1/2}(\mathbf{B}_k - \mathbf{I})\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2})$.

Proof. An outline of the proof is given in Appendix A. □

IV. Computation of the Information Rate Region under Sum-Rate Constraint

In this section, we describe an iterative Blahut-Arimoto (BA)-type algorithm to compute the pmfs $P_{U_k|Y_k}$, $k \in \mathcal{K}$, that maximize information Δ under sum-rate constraint, i.e., $R_{\text{sum}} := \sum_{k=1}^K R_k$, for tuples $(\Delta, R_1, \dots, R_K)$ in $\mathcal{R}_{\text{IB}}$. From Theorem 1 we have:

$$\mathcal{R}_{\text{sum}} := \text{convex-hull}\{(\Delta, R_{\text{sum}}) : \Delta \leq \Delta_{\text{sum}}(R_{\text{sum}})\}, \tag{7}$$

where we define the information-rate function

$$\Delta_{\text{sum}}(R) := \max_{\mathbf{P}} \min \left\{ I(X; U_{\mathcal{K}}), R - \sum_{k=1}^K I(Y_k; U_k | X) \right\},$$

and where the optimization is over the set of K conditional pmfs $P_{U_k|Y_k}$, $k \in \mathcal{K}$, which, for short, we define as

$$\mathbf{P} := \{P_{U_1|Y_1}, \dots, P_{U_K|Y_K}\}. \tag{8}$$

Next proposition provides a characterization of the pairs $(\Delta, R_{\text{sum}}) \in \mathcal{R}_{\text{sum}}$ in terms of a parameter $s \geq 0$.

Proposition 2. *Each tuple $(\Delta, R_{\mathrm{sum}})$ on the information rate curve $\Delta = \Delta_{\mathrm{sum}}(R_{\mathrm{sum}})$, can be obtained for some $s \geq 0$, as (Δ_s, R_s), parametrically defined by*

$$(1+s)\Delta_s = (1+sK)H(X) + sR_s - \min_{\mathbf{P}} F_s(\mathbf{P}), \tag{9}$$

$$R_s = I(Y_{\mathcal{K}}; U_{\mathcal{K}}^*) + \sum_{k=1}^{K} [I(Y_k; U_k^*) - I(X; U_k^*)], \tag{10}$$

where $\mathbf{P}^*$ are the pmfs yielding the minimum in (9) *and*

$$F_s(\mathbf{P}) := H(X|U_{\mathcal{K}}) + s\sum_{k=1}^{K} [I(Y_k; U_k) + H(X|U_k)]. \tag{11}$$

Proof. The proof of Proposition 2 follows along the lines of [12, Theorem 2] and is omitted for brevity. Note that the rate expression in Theorem 1 is different to that in [12]. □

From Proposition 2, the information-rate function can be computed by solving (9) and evaluating (10) for all $s \geq 0$. Inspired by the standard Blahut-Arimoto (BA) method [11], and following similar steps as for the BA-type algorithm proposed in [12] for the CEO problem with $K = 2$ encoders, we show that problem (9) can be solved with an alternate optimization procedure, with respect to $\mathbf{P}$ and some appropriate auxiliary pmfs Q_{U_k}, $Q_{X|U_k}$, $k \in \mathcal{K}$ and $Q_{X|U_1,\ldots,U_K}$, denoted as

$$\mathbf{Q} := \{Q_{U_1}, \ldots, Q_{U_K}, Q_{X|U_1}, \ldots, Q_{X|U_K}, Q_{X|U_1,\ldots,U_K}\}.$$

To this end, we define the function $\bar{F}_s(\cdot)$ and write (9) as a minimization over the pmfs $\mathbf{P}$ and pmfs $\mathbf{Q}$, where

$$\begin{aligned} \bar{F}_s(\mathbf{P}, \mathbf{Q}) :=& - s\sum_{k=1}^{K} H(U_k|Y_k) - s\sum_{k=1}^{K} \mathrm{E}_{U_k}[\log q(U_k)] \\ &- s\sum_{k=1}^{K} \mathrm{E}_{X,U_k}[\log q(X|U_k)] - \mathrm{E}_{X,U_{\mathcal{K}}}[\log q(X|U_{\mathcal{K}})]. \end{aligned} \tag{12}$$

Lemma 1. *We have*

$$F^* := \min_{\mathbf{P}} F_s(\mathbf{P}) = \min_{\mathbf{P}} \min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q}). \tag{13}$$

Algorithm 1 describes the steps to successively minimize $\bar{F}_s(\mathbf{P}, \mathbf{Q})$ by optimizing a convex problem over $\mathbf{P}$ and over $\mathbf{Q}$ at each iteration. The proof of Lemma 1 and the steps of the proposed algorithm are justified with the following lemmas, whose proofs are along the lines of Lemma 1, Lemma 2, Lemma 3 in [12], and are omitted due to space limitations.

Lemma 2. *$\bar{F}_s(\mathbf{P}, \mathbf{Q})$ is convex in $\mathbf{P}$ and convex in $\mathbf{Q}$.*

Lemma 3. *For fixed pmfs $\mathbf{P}$, $\bar{F}_s(\mathbf{P}, \mathbf{Q}) \geq F_s(\mathbf{P})$ for all pmfs $\mathbf{Q}$, and there exists a unique $\mathbf{Q}$ that achieves the minimum $\min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q}) = F_s(\mathbf{P})$, given by*

$$Q_{U_k}^* = P_{U_k}, \; Q_{X|U_k}^* = P_{X|U_k}, \quad k \in \mathcal{K}, \tag{14}$$

$$Q_{X|U_1,\ldots,U_K}^* = P_{X|U_1,\ldots,U_K}, \tag{15}$$

where P_{U_k}, $P_{X|U_k}$ and $P_{X|U_1,\ldots,U_K}$ are computed from $\mathbf{P}$.

Lemma 4. *For fixed $\mathbf{Q}$, there exists a $\mathbf{P}$ that achieves the minimum $\min_{\mathbf{P}} \bar{F}_s(\mathbf{P}, \mathbf{Q})$, where $P_{U_k|Y_k}$ is given by*

$$p^*(u_k|y_k) = q(u_k)\frac{\exp\left(-\psi_s(u_k, y_k)\right)}{\sum_{u_k \in \mathcal{U}_k} q(u_k)\exp(-\psi_s(u_k, y_k))}, \tag{16}$$

for $u_k \in \mathcal{U}_k$ and $y_k \in \mathcal{Y}_k$, $k \in \mathcal{K}$, and where we define

$$\begin{aligned} \psi_s(u_k, y_k) :=& D_{\mathrm{KL}}(P_{X|y_k} || Q_{X|u_k}) \\ &+ \frac{1}{s}\mathrm{E}_{U_{\mathcal{K}\setminus k}|y_k}[D_{\mathrm{KL}}(P_{X|U_{\mathcal{K}\setminus k}, y_k} || Q_{X|U_{\mathcal{K}\setminus k}, u_k}))]. \end{aligned} \tag{17}$$

Algorithm 1 BA-type algorithm for the Discrete D-IB

```
input: pmf $P_{X,Y_1,\ldots,Y_k}$, parameter $s \geq 0$.
output: optimal $P_{U_k|Y_k}^*$, pair $(\Delta_s, R_s)$.
initialization Set $t = 0$ and set $\mathbf{P}^{(0)}$ with $p(u_k|y_k) = \frac{1}{|\mathcal{U}_k|}$
               for $u_k \in \mathcal{U}_k$, $y_k \in \mathcal{Y}_k$, $k = 1, \ldots, K$.
repeat
    Compute $\mathbf{Q}^{(t+1)}$ as (14) and (15) from $\mathbf{P}^{(t)}$.
    Compute $\mathbf{P}^{(t+1)}$ as (16) from $\mathbf{Q}^{(t+1)}$ and $\mathbf{P}^{(t)}$.
    $t \leftarrow t + 1$.
until convergence.
```

Algorithm 1 essentially falls in the Successive Upper-Bound Minimization (SUM) framework [15] in which $\bar{F}_s(\mathbf{P}, \mathbf{Q})$ acts as a globally tight upper bound on $F_s(\mathbf{P})$. Algorithm 1 provides a sequence $\mathbf{P}^{(t)}$ for each iteration t, which converges to a stationary point of the optimization problem (13).

Proposition 3. *Every limit point of the sequence $\mathbf{P}^{(t)}$ generated by Algorithm 1 converges to a stationary point of* (13).

Proof. Let $\mathbf{Q}^*(\mathbf{P}) := \arg\min_{\mathbf{Q}} \bar{F}_s(\mathbf{P}, \mathbf{Q})$. From Lemma 3, $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}')) \geq \bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P})) = F_s(\mathbf{P})$ for $\mathbf{P}' \neq \mathbf{P}$. It follows that $F_s(\mathbf{P})$ and $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}'))$ satisfy [15, Proposition 1] and thus $\bar{F}_s(\mathbf{P}, \mathbf{Q}^*(\mathbf{P}'))$ satisfies A1-A4 in [15]. Convergence to a stationary point of (13) follows from [15, Theorem 1]. □

Remark 1. *The resulting set of self consistent equations* (14), (15) *and* (17) *satisfied by any stationary point of the D-IB problem, remind that of the original IB problem [9]. Note the additional divergence term in* (17) *for encoder k averaged over the descriptions at the other $\mathcal{K} \setminus k$ encoders.*

V. Computation of the Information Rate Region for the Vector Gaussian D-IB

Computing the maximum information under sum-rate constraint from Theorem 2 is a convex optimization problem on $\mathbf{B}_k$, which can be efficiently solved with generic tools. Alternatively, next we extend Algorithm 1 for Gaussian sources.

For finite alphabet sources the updates of $\mathbf{Q}^{(t+1)}$ and $\mathbf{P}^{(t+1)}$ in Algorithm 1 are simple, but become unfeasible for continuous alphabet sources. We leverage on the optimality of Gaussian descriptions, shown in Theorem 2, to restrict the optimization of $\mathbf{P}$ to Gaussian distributions, which are easily represented by a finite set of parameters, namely its mean and covariance. We show that if $\mathbf{P}^{(t)}$ are Gaussian pmfs, then $\mathbf{P}^{(t+1)}$ are also Gaussian pmfs, which can be computed with an efficient update algorithm of its representing parameters. In particular, if at time t, the k-th pmf $P_{\mathbf{U}_k|\mathbf{Y}_k}^{(t)}$ is given by

$$\mathbf{U}_k^t = \mathbf{A}_k^t \mathbf{Y}_k + \mathbf{Z}_k^t, \tag{18}$$

where $\mathbf{Z}_k^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^t})$; we show that for $\mathbf{P}^{(t+1)}$ updated as in (16), $P_{\mathbf{U}_k|\mathbf{Y}_k}^{(t+1)}$ corresponds to $\mathbf{U}_k^{t+1} = \mathbf{A}_k^{t+1}\mathbf{Y}_k + \mathbf{Z}_k^{t+1}$, where $\mathbf{Z}_k^{t+1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}})$ and $\mathbf{A}_k^{t+1}, \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}$ are updated as

$$\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} = \left(\left(1 + \frac{1}{s}\right)\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1} - \frac{1}{s}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\setminus k}^t}^{-1}\right)^{-1}, \tag{19}$$

$$\begin{aligned} \mathbf{A}_k^{t+1} =& \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1}\left(\left(1 + \frac{1}{s}\right)\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}^{-1}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1})\right. \\ &\left. - \frac{1}{s}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}_{\mathcal{K}\setminus k}^t}^{-1}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}_{\mathcal{K}\setminus k}^t}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1})\right). \end{aligned} \tag{20}$$

The detailed update procedure is given in Algorithm 2.

Remark 2. *Algorithm 2 generalizes the iterative algorithm for single encoder Gaussian D-IB in [10] to the Gaussian D-IB with K encoders and sum-rate constraint. Similarly to the solution in [10], the optimal description at each encoder is given by a noisy linear projection of the observation, whose dimensionality is determined by the parameter s and the second order moments between the observed data and the source of interest, as well as a term depending on the observed data with respect to the descriptions at the other encoders.*

A. Derivation of Algorithm 2

In this section, we derive the update rules in Algorithm 2 and show that the Gaussian distribution is invariant to the update rules in Algorithm 1, in line with Theorem 2.

First, we recall that if $(\mathbf{X}_1, \mathbf{X}_2)$ are jointly Gaussian, then

$$P_{\mathbf{X}_2|\mathbf{X}_1=\mathbf{x}_1} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1}, \mathbf{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}), \tag{21}$$

where $\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1} := \mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1}\mathbf{x}_1$, with $\mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1} := \mathbf{\Sigma}_{\mathbf{x}_2,\mathbf{x}_1}\mathbf{\Sigma}_{\mathbf{x}_1}^{-1}$.

Then, for $\mathbf{Q}^{(t+1)}$ computed as in (14) and (15) from $\mathbf{P}^{(t)}$, which is a set of Gaussian distributions, we have

$$Q^{(t+1)}_{\mathbf{X}|\mathbf{u}_k} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_k^t}, \mathbf{\Sigma}_{\mathbf{x}|\mathbf{u}_k^t}), Q^{(t+1)}_{\mathbf{X}|\mathbf{u}_{\mathcal{K}}} = \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_{\mathcal{K}}^t}, \mathbf{\Sigma}_{\mathbf{x}|\mathbf{u}_{\mathcal{K}}^t}).$$

Next, we look at the update $\mathbf{P}^{(t+1)}$ as in (16) from given $\mathbf{Q}^{(t+1)}$. First, we have that $p(\mathbf{u}_k^t)$ is the marginal of $\mathbf{U}_k^t$, given by $\mathbf{U}_k^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{u}_k^t})$ where $\mathbf{\Sigma}_{\mathbf{u}_k^t} = \mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k}\mathbf{A}_k^{t,H} + \mathbf{\Sigma}_{\mathbf{z}_k^t}$.

Then, to compute $\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)$, first, we note that

$$\begin{aligned}&E_{U_{\mathcal{K}\setminus k}|y_k}[D_{\mathrm{KL}}(P_{X|U_{\mathcal{K}\setminus k},y_k}||Q_{X|U_{\mathcal{K}\setminus k},u_k})] \quad (22)\\ =&D_{\mathrm{KL}}(P_{X,U_{\mathcal{K}\setminus k}|y_k}||Q_{X,U_{\mathcal{K}\setminus k}|u_k}) - D_{\mathrm{KL}}(P_{U_{\mathcal{K}\setminus k}|y_k}||Q_{U_{\mathcal{K}\setminus k}|u_k}),\end{aligned}$$

and that for two generic multivariate Gaussian distributions $P_1 \sim \mathcal{CN}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $P_2 \sim \mathcal{CN}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ in $\mathbb{C}^N$,

$$\begin{aligned}D_{\mathrm{KL}}(P_1, P_2) =&(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^H\mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\\ &+ \log|\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1}| - N + \mathrm{tr}\{\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\}. \quad (23)\end{aligned}$$

Applying (22) and (23) in (17) and noting that all involved distributions are Gaussian, it follows that $\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)$ is a quadratic form. Then, since $p(\mathbf{u}_k^t)$ is Gaussian, the product $\log(p(\mathbf{u}_k^t)\exp(-\psi_s(\mathbf{u}_k^t, \mathbf{y}_k)))$ is also a quadratic form, and identifying constant, first and second order terms, we can write

$$\begin{aligned}\log p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k) =&Z(\mathbf{y}_k) + (\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k})^H\mathbf{\Sigma}^{-1}_{\mathbf{z}_k^{t+1}}\\ &\cdot(\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}), \quad (24)\end{aligned}$$

where $Z(\mathbf{y}_k)$ is a normalization term independent of $\mathbf{u}_k$, and

$$\begin{aligned}\mathbf{\Sigma}^{-1}_{\mathbf{z}_k^{t+1}} =&\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t} + \mathbf{K}^H_{\mathbf{x}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x}|\mathbf{u}_k^t}\mathbf{K}_{\mathbf{x}|\mathbf{u}_k^t}\\ &+ \frac{1}{s}\mathbf{K}^H_{\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{K}_{\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\\ &- \frac{1}{s}\mathbf{K}^H_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{K}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}, \quad (25)\\ \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} =&\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}\Big(\mathbf{K}^H_{\mathbf{x}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k}\\ &+ \frac{1}{s}\mathbf{K}_{\mathbf{x},\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x},\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{x},\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{y}_k}\\ &-\frac{1}{s}\mathbf{K}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{y}_k}\Big). \quad (26)\end{aligned}$$

This shows that $p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k)$ is a Gaussian distribution and that $\mathbf{U}_k^{t+1}|\{\mathbf{Y}_k = \mathbf{y}_k\}$ is distributed as $\mathcal{CN}(\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}, \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}})$.

Algorithm 2 BA-type algorithm for the Gaussin Vector D-IB

```
input: covariance $\mathbf{\Sigma}_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k}$, parameter $s \geq 0$.
output: optimal pairs $(\mathbf{A}_k^*, \mathbf{\Sigma}_{\mathbf{z}_k^*})$, $k = 1, \ldots, K$.
initialization Set randomly $\mathbf{A}_k^0$ and $\mathbf{\Sigma}_{\mathbf{z}_k^0} \succeq 0$, $k \in \mathcal{K}$.
repeat
    Compute $\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}^t_{\mathcal{K}\setminus k}}$ and update for $k \in \mathcal{K}$
        $\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}} = \mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\mathbf{A}_k^{t,H} + \mathbf{\Sigma}_{\mathbf{z}_k^t}$    (27)
        $\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}^t_{\mathcal{K}\setminus k}} = \mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}^t_{\mathcal{K}\setminus k}}\mathbf{A}_k^{t,H} + \mathbf{\Sigma}_{\mathbf{z}_k^t}$,    (28)
    Compute $\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}$ as in (19) for $k \in \mathcal{K}$.
    Compute $\mathbf{A}_k^{t+1}$ as (20), $k \in \mathcal{K}$.
    $t \leftarrow t + 1$.
until convergence.
```

Next, we simplify (25) and (26) to obtain the update rules (19) and (20). From the matrix inversion lemma, similarly to [10], for $(\mathbf{X}_1, \mathbf{X}_2)$ jointly Gaussian we have

$$\mathbf{\Sigma}^{-1}_{\mathbf{x}_2|\mathbf{x}_1} = \mathbf{\Sigma}^{-1}_{\mathbf{x}_2} + \mathbf{K}^H_{\mathbf{x}_1|\mathbf{x}_2}\mathbf{\Sigma}^{-1}_{\mathbf{x}_1|\mathbf{x}_2}\mathbf{K}_{\mathbf{x}_1|\mathbf{x}_2}. \tag{29}$$

Applying (29), in (25) we have

$$\begin{aligned}\mathbf{\Sigma}^{-1}_{\mathbf{z}_k^{t+1}} &= \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}} + \frac{1}{s}\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}} - \frac{1}{s}\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{u}^t_{\mathcal{K}\setminus k}}, \quad (30)\\ &= \left(1 + \frac{1}{s}\right)\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}} - \frac{1}{s}\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{u}^t_{\mathcal{K}\setminus k}}, \quad (31)\end{aligned}$$

where (31) is due to the Markov chain $\mathbf{U}_k \mathrel{-\!\!\circ\!\!-} \mathbf{X} \mathrel{-\!\!\circ\!\!-} \mathbf{U}_{\mathcal{K}\setminus k}$.

Then, also from the matrix inversion lemma, we have for jointly Gaussian $(\mathbf{X}_1, \mathbf{X}_2)$,

$$\mathbf{\Sigma}^{-1}_{\mathbf{x}_2|\mathbf{x}_1}\mathbf{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\mathbf{\Sigma}^{-1}_{\mathbf{x}_1} = \mathbf{\Sigma}^{-1}_{\mathbf{x}_2}\mathbf{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\mathbf{\Sigma}^{-1}_{\mathbf{x}_1|\mathbf{x}_2}. \tag{32}$$

Applying (32) in (26), for the first term, we have

$$\begin{aligned}\mathbf{K}^H_{\mathbf{x}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k} =&\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}}\mathbf{\Sigma}_{\mathbf{x},\mathbf{u}_k^t}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}_k} \quad (33)\\ =&\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}}\mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{x}}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{\Sigma}_{\mathbf{x},\mathbf{y}_k}\mathbf{\Sigma}^{-1}_{\mathbf{y}_k}\mathbf{y}_k\\ =&\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\mathbf{\Sigma}^{-1}_{\mathbf{y}_k})\mathbf{y}_k, \quad (34)\end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{x},\mathbf{u}_k^t} = \mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{x}}$; and (34) is due to the definition of $\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}$. Similarly, for the second term, we have

$$\begin{aligned}&\mathbf{K}_{\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{x},\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{y}_k}\\ &=\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}}\mathbf{\Sigma}^{-1}_{\mathbf{y}_k})\mathbf{y}_k, \quad (35)\\ &=\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{x}}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}}\mathbf{\Sigma}^{-1}_{\mathbf{y}_k})\mathbf{y}_k, \quad (36)\end{aligned}$$

where we use $\mathbf{\Sigma}_{\mathbf{u}_k^t,\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}} = \mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{x}\mathbf{u}^t_{\mathcal{K}\setminus k}}$; and (36) is due to the Markov chain $\mathbf{U}_k \mathrel{-\!\!\circ\!\!-} \mathbf{X} \mathrel{-\!\!\circ\!\!-} \mathbf{U}_{\mathcal{K}\setminus k}$. For the third term,

$$\begin{aligned}&\mathbf{K}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\mathbf{\Sigma}^{-1}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{u}_k^t}\boldsymbol{\mu}_{\mathbf{u}^t_{\mathcal{K}\setminus k}|\mathbf{y}_k}\\ &=\mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{u}^t_{\mathcal{K}\setminus k}}\mathbf{A}_k^t(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}^t_{\mathcal{K}\setminus k}}\mathbf{\Sigma}^{-1}_{\mathbf{y}_k})\mathbf{y}_k. \quad (37)\end{aligned}$$

Equation (20) follows by noting that $\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} = \mathbf{A}_k^{t+1}\mathbf{y}_k$, and that from (26) $\mathbf{A}_k^{t+1}$ can be identified as given in (20).

Finally, we note that due to (18), $\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{u}^t_{\mathcal{K}\setminus k}}$ are given as in (27) and (28), where $\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{n}_k}$ and $\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{u}^t_{\mathcal{K}\setminus k}}$ can be computed from its definition.

VI. Numerical Results

In this section, we consider the numerical evaluation of Algorithm 2, and compare the resulting relevant information to two upper bounds on the performance for the D-IB: i) the information-rate pairs achievable under centralized IB encoding, i.e., if $(Y_1, \dots, Y_K)$ are encoded jointly at a rate equal to the total rate $R_{\text{sum}} = R_1 + \dots + R_K$, characterized in [10]; ii) the information-rate pairs achievable under centralized IB encoding when $R_{\text{sum}} \to \infty$, i.e., $\Delta = I(X; Y_1, \dots, Y_K)$.

Figure 2 shows the resulting (Δ, R_{sum}) tuples for a Gaussian vector model with $K = 2$ encoders, source dimension $N = 4$, and observations dimension $M_1 = M_2 = 2$ for different values of s calculated as in Proposition 2 using Algorithm 2, and its upper convex envelope. As it can be seen, the distributed IB encoding of sources performs close to the Tishby's centralized IB method, particularly for low R_{sum} values.

Appendix A

Let $(\mathbf{X}, \mathbf{U})$ be two complex random vectors. The conditional Fischer information is defined as $\mathbf{J}(\mathbf{X}|\mathbf{U}) := \mathrm{E}[\nabla \log p(\mathbf{X}|\mathbf{U}) \nabla \log p(\mathbf{X}|\mathbf{U})^H]$, and the MMSE is given by $\mathrm{mmse}(\mathbf{X}|\mathbf{U}) := \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{U}])(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{U}])^H]$. Then [16]

$$\log |(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U})| \leq h(\mathbf{X}|\mathbf{U}) \leq \log |(\pi e)\mathrm{mmse}(\mathbf{X}|\mathbf{U})|. \tag{38}$$

We outer bound the information-rate region in Theorem 1 for $(\mathbf{X}, \mathbf{Y}_{\mathcal{K}})$ as in (6). For $q \in \mathcal{Q}$ and fixed $\prod_{k=1}^{K} p(\mathbf{u}_k|\mathbf{y}_k, q)$, choose $\mathbf{B}_{k,q}$, $k \in \mathcal{K}$ satisfying $\mathbf{0} \preceq \mathbf{B}_{k,q} \preceq \mathbf{\Sigma}_{\mathbf{n}_k}^{-1}$ such that

$$\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q) = \mathbf{\Sigma}_{\mathbf{n}_k} - \mathbf{\Sigma}_{\mathbf{n}_k}\mathbf{B}_{k,q}\mathbf{\Sigma}_{\mathbf{n}_k}. \tag{39}$$

Such $\mathbf{B}_{k,q}$ always exists since $\mathbf{0} \preceq \mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q) \preceq \mathbf{\Sigma}_{\mathbf{n}_k}^{-1}$, for all $q \in \mathcal{Q}$, and $k \in \mathcal{K}$. We have from (5),

$$\begin{aligned} I(\mathbf{Y}_k; \mathbf{U}_k|\mathbf{X}, q) &\geq \log |\mathbf{\Sigma}_{\mathbf{n}_k}| - \log |\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,q}, q)| \\ &= -\log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}\mathbf{B}_{k,q}\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}|, \end{aligned} \tag{40}$$

where the inequality is due to (38), and (40) is due to (39). Let $\bar{\mathbf{B}}_k := \sum_{q \in \mathcal{Q}} p(q)\mathbf{B}_{k,q}$. Then, we have from (40)

$$\begin{aligned} I(\mathbf{Y}_k; \mathbf{U}_k|\mathbf{X}, Q) &\geq - \sum_{q \in \mathcal{Q}} p(q) \log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}\mathbf{B}_{k,q}\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}| \\ &\geq -\log |\mathbf{I} - \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}\bar{\mathbf{B}}_k\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}|, \end{aligned} \tag{41}$$

where (41) follows from the concavity of the log-det function and Jensen's inequality. On the other hand, we have

$$I(\mathbf{X}; \mathbf{U}_{S^c,q}|q) \leq \log |\mathbf{\Sigma}_{\mathbf{x}}| - \log |\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U}_{S^c,q}, q)| \tag{42}$$

$$= \log \left| \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_{\mathbf{x}}^{1/2}\mathbf{H}_k^H\mathbf{B}_{k,q}\mathbf{H}_k\mathbf{\Sigma}_{\mathbf{x}}^{1/2} + \mathbf{I} \right|, \tag{43}$$

where (42) is due to (38); and (43) is due to to the following equality connecting the MMSE matrix (39) and the Fisher information as in [16]–[19] (We refer to [13] for details):

$$\mathbf{J}(\mathbf{X}|\mathbf{U}_{S^c,q}, q) = \sum_{k \in \mathcal{S}^c} \mathbf{H}_k^H\mathbf{B}_{k,q}\mathbf{H}_k + \mathbf{\Sigma}_{\mathbf{x}}^{-1}. \tag{44}$$

Similarly to (41), from (43) and Jensen's Inequality we have

$$I(\mathbf{X}; \mathbf{U}_{S^c}|Q) \leq \log \left| \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_{\mathbf{x}}^{1/2}\mathbf{H}_k^H\bar{\mathbf{B}}_k\mathbf{H}_k\mathbf{\Sigma}_{\mathbf{x}}^{1/2} + \mathbf{I} \right|. \tag{45}$$

Substituting (41) and (45) in (5) and letting $\mathbf{B}_k := \mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2}\bar{\mathbf{B}}_k\mathbf{\Sigma}_{\mathbf{n}_k}^{-1/2}$ gives the outer bound. The proof is completed by noting that the outer bound is achieved with $Q = \emptyset$ and $p^*(\mathbf{u}_k|\mathbf{y}_k, q) = \mathcal{CN}(\mathbf{y}_k, \mathbf{\Sigma}_{\mathbf{n}_k}^{1/2}(\mathbf{B}_k - \mathbf{I})\mathbf{\Sigma}_{\mathbf{n}_k}^{1/2})$.

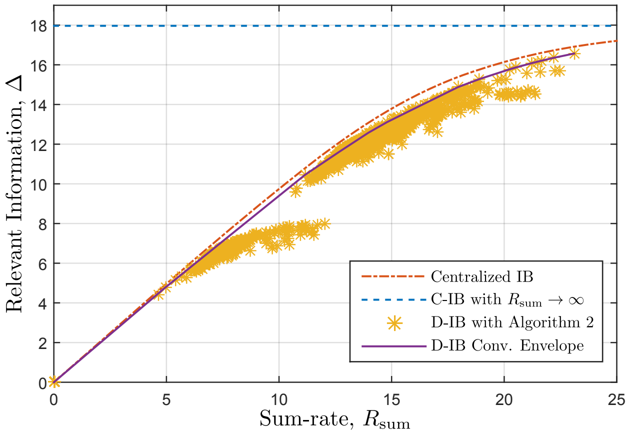


Fig. 2. Information vs. sum-rate for vector Gaussian D-IB with $K = 2$ encoders, source dimension $N = 4$, and observation dimension $M_1 = M_2 = 2$.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conf. on Comm., Control, and Computing*, 1999, pp. 368–377.

[2] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.

[3] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. of 23rd Ann. Int'l ACM SIGIR Conf. on Res. and Dev. in Info. Retrieval*, 2000, pp. 208–215.

[4] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, "Information-based clustering," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 297–18 302, 2005.

[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information." [Online]. Available: http://arxiv.org/abs/1703.00810

[6] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Info. Theory*, vol. 54, no. 10, pp. 4666–4681, Oct. 2008.

[7] K. Kittichokechai and G. Caire, "Privacy-constrained remote source coding," in *IEEE Int. Symp. Inf. Th. (ISIT)*, Jul. 2016, pp. 1078–1082.

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

[9] A. Globerson and N. Tishby, "On the optimality of the Gaussian information bottleneck curve," *Hebrew University Tech. Report*, 2004.

[10] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables." *Journal of Machine Learning Research*, vol. 6, pp. 165–188, Feb. 2005.

[11] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[12] Y. Ugur, I. E. Aguerri, and A. Zaidi, "A generalization of Blahut-Arimoto algorithm to computing rate-distortion regions of multiterminal source coding under logarithmic loss," in *Proc. of IEEE Info. Theory Workshop, ITW*, Kaohsiung, Taiwan, Nov. 2017.

[13] I. Estella and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources." [Online]. Available: http://www-syscom.univ-mlv.fr/ zaidi/publications/proofs-dIB-izs2017.pdf

[14] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. IEEE Int. Symp. Information Theory*, Jun. 2007, pp. 566–570.

[15] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. on Opt.*, vol. 23, pp. 1126–1153, Jun. 2013.

[16] E. Ekrem and S. Ulukus, "An outer bound for the vector Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, Nov. 2014.

[17] Y. Zhou, Y. Xu, W. Yu, and J. Chen, "On the optimal fronthaul compression and decoding strategies for uplink cloud radio access networks," *IEEE Tr. Inf. Th.*, vol. 62, no. 12, pp. 7402–7418, Dec. 2016.

[18] I. Estella, A. Zaidi, G. Caire, and S. Shamai, "On the capacity of cloud radio access networks," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2017.

[19] I. E. Aguerri, A. Zaidi, G. Caire, and S. Shamai, "On the capacity of cloud radio access networks with oblivious relaying." [Online]. Available: http://arxiv.org/abs/1710.09275

Near Maximum Likelihood Decoding with Deep Learning

Eliya Nachmani, Yaron Bachar, Elad Marciano, David Burshtein and Yair Be'ery
School of Electrical Engineering, Tel-Aviv University, Israel
Emails: enk100@gmail.com, yaronbac@gmail.com, eladmarc@gmail.com, burstyn@eng.tau.ac.il, ybeery@eng.tau.ac.il

***Abstract*—A novel and efficient neural decoder algorithm is proposed. The proposed decoder is based on the neural Belief Propagation algorithm and the Automorphism Group. By combining neural belief propagation with permutations from the Automorphism Group we achieve near maximum likelihood performance for High Density Parity Check codes. Moreover, the proposed decoder significantly improves the decoding complexity, compared to our earlier work on the topic. We also investigate the training process and show how it can be accelerated. Simulations of the hessian and the condition number show why the learning process is accelerated. We demonstrate the decoding algorithm for various linear block codes of length up to 63 bits.**

I. INTRODUCTION

In the last few years Deep Learning methods were applied to communication systems, for example in [1]–[8]. Furthermore, Deep Neural decoders is a new approach for decoding linear block codes. In [9]–[12] it has been shown that deep neural decoders can improve the existing belief propagation methods for decoding high density parity check codes (HDPCs). Other methods for using deep learning to decode error correcting codes were proposed in [13]–[15]. In this work we combine the deep recurrent neural decoder of [10] with permutations from the Automorphism Group as defined in [16]. The combined architecture is defined by $I_{permutations}$ blocks, each of which contains I_{BP} iterations of neural belief propagation followed by permutation. We show that this architecture achieves near maximum-likelihood performance for various BCH codes of up to 63 bits long with significantly lower complexity then the corresponding mRRD decoder [17]. We also investigate the training process of the deep neural decoder and show how the learning can be accelerated by adding penalties to the loss function. We argue that this penalties transform the manifold of the loss function into an isotropic manifold which is easy to optimize. Simulations of the Hessian matrix of the loss function support this claim.

II. THE NEURAL BELIEF PROPAGATION ALGORITHM

We start with a brief description of the deep neural network proposed in [9], [10]. The deep neural decoder is a message passing algorithm parameterized as a deep neural network. The input to the neural network is the set of LLR values, $v = 1, 2, \dots, N$,

$$l_v = \log \frac{\Pr(C_v = 1|y_v)}{\Pr(C_v = 0|y_v)}$$

where N is the block length of the code, y_v is the channel output corresponding to the vth codebit. The neural decoder consists of pairs of odd and even layers. For odd i layer,

$$\begin{aligned} x_{i,e=(v,c)} &= \\ &= \tanh\left(\frac{1}{2}\left(l_v + \sum_{e'=(c',v),\, c'\neq c} w_{e,e'} x_{i-1,e'}\right)\right) \end{aligned} \tag{1}$$

for even i layer,

$$x_{i,e=(c,v)} = 2\tanh^{-1}\left(\prod_{e'=(v',c),\, v'\neq v} x_{i-1,e'}\right) \tag{2}$$

and for output layer,

$$o_v = \sigma\left(l_v + \sum_{e'=(c',v)} \tilde{w}_{v,e'} x_{2L,e'}\right) \tag{3}$$

where $\sigma(x) \equiv \left(1 + e^{-x}\right)^{-1}$ is a sigmoid function. Please note that equations (1),(2) define recurrent neural network, as the learnable weights $w_{e,e'}, \tilde{w}_{v,e'}$ are tied.

III. THE PROPOSED DEEP NEURAL NETWORK DECODER

A. Architecture

The proposed neural network is composed of $I_{permutations}$ blocks. Each block contains I_{BP} layers of neural belief propagation, which are described below. Between each two successive blocks, we apply a permutation from the automorphism group. Lastly, we apply the corresponding inverse permutation, to obtain the decoded codeword. The proposed architecture is illustrated in Figure 1.

We re-parameterized the deep neural network decoder from section II. In the j-th block, I_{BP} iterations of neural belief-propagation are performed as follows:

For each variable node in the i-th layer,

$$x^j_{i,e=(v,c)} = \tanh\left(\frac{1}{2}(o^j_{i-1,v} - x^j_{i-1,e=(c,v)})\right) \tag{4}$$

For each check node in the i-th layer,

$$x^j_{i,e=(c,v)} = 2\tanh^{-1}\left(\prod_{e'=(v',c),\, v'\neq v} x^j_{i-1,e'}\right) \tag{5}$$

For mid-output node in the i-th layer,

$$o_{i,v}^{j} = o_{0,v}^{j} + \sum_{e'=(c',v)} w_{e'} x_{i,e'}^{j} \tag{6}$$

The output of the j-th block:

$$c_v^j = \pi_j(o_{i=I_{BP},v}^{j}) \tag{7}$$

The initialization:

$$o_{0,v}^{j} = \begin{cases} l_v & ,j = 0 \\ c_v^{j-1} & ,j > 0 \end{cases} \tag{8}$$

$$x_{0,e}^{j} = 0 \tag{9}$$

After each BP block, an appropriate inverse permutation is applied:

$$d_v^j = (\pi_1^{-1} \cdot \pi_2^{-1} \cdot ... \cdot \pi_j^{-1}) c_v^j \tag{10}$$

Note that the neural weights $w_{e'}$ are tied along the neural network graph. Also note, that the new parametrization of the neural belief propagation decoder was easier to optimize, and converged faster to a better performance.

As in [9]–[12] the proposed architecture preserves the symmetry conditions, therefore we can train the neural network with noisy versions of a single codeword.

Also note, that in order to be consistent with the BP algorithm, one needs to multiply $x_{i-1,e=(c,v)}^{j}$ in (4) by w_e. However, this multiplication did not have any significant influence on the results obtained.

B. Loss function

The loss of the neural network is composed of three constituents:

- A multi-loss cross-entropy between $\tilde{d}_v^j \equiv \sigma(d_v^j)$ and the correct codeword y_v. This is a loss term concerned with the output of the BP blocks:

$$L_1^j = -\frac{1}{N}\sum_{v=1}^{N} y_v \log(\tilde{d}_v^j) + (1-y_v)\log(1-\tilde{d}_v^j) \tag{11}$$

- A sub multi-loss cross-entropy between $\tilde{o}_{i,v}^{j} \equiv \sigma(o_{i,v}^{j})$ and the correct codeword y_v. This term involves inner-BP marginalizations:

$$L_2^{j,i} = -\frac{1}{N}\sum_{v=1}^{N} y_v \log(o_{i,v}^{j}) + (1-y_v)\log(1-o_{i,v}^{j}) \tag{12}$$

- l_2-norm of the weights w_v, $w_{v,e'}$:

$$L_3 = \sum_{v} \|w_v\|^2 + \sum_{v,e} \|w_{v,e}\|^2 \tag{13}$$

The total loss is:

$$L = \sum_{j} (L_1^j + \lambda \cdot L_3) + \sum_{j,i} L_2^{j,i} \tag{14}$$

IV. EXPERIMENTS

A. Neural Network Training and Dataset

We implemented the proposed neural network in TensorFlow framework. The neural network was optimized with RMSPROP [18]. As in [9]–[12], the dataset consisted of the zero codeword and an AWGN channel. We used the cycle reduced parity check matrix from [19]. Due to large number of layers in our network, and the fact that the network is a recurrent neural network, gradient clipping was applied to avoid gradient exploding throughout the learning process. Clipping threshold of $c_{grad} = 0.1$ was used. The l_2-Loss term was added with a factor of λ. Note that we use the three terms L_1, L_2, L_3 of the loss for training. The weights were constrained to have non-negative values. In all of our experiments no overfitting was observed.

The architecture was tested on BCH codes. Their automorphism group is described in detail in [20]. The permutations were chosen randomly using the product-replacement algorithm [19], which has the N_{pr} and K_{pr} parameters. N_{pr} is the size of the group of permutations the algorithm builds, and K_{pr} is the initial number of iterations, used to build this permutations-reservoir. In Table I we provide details about the parameters configurations of the network.

B. BCH(63,45)

Batch size was set to 160, with 20 examples per SNR. The SNR varied from $1dB$ to $8dB$ in the training process, and from $1dB$ to $5dB$ in the validation process. The neural network comprises $I_{permutations} = 50$ permutations, and each block contains $I_{BP} = 2$ BP iterations. A total of 100 BP iterations correspond to a deep neural network with 200 layers.

C. BCH(63,36)

Batch size was set to 120, with 30 examples per SNR. The SNR varied from $1dB$ to $6dB$ in the training process, and from $3dB$ to $4.5dB$ in the validation process. The neural network comprised $I_{permutations} = 300$ permutation, and each block contains $I_{BP} = 2$ belief propagation iteration. This configuration represents 600 Belief Propagation iterations which correspond to deep neural network with 1200 layers.

	Parameter	BCH(63,45)	BCH(63,36)
BP	I_{BP}	2	2
	llr clip	15	15
RRD	$I_{Permutatoins}$	50	300
	N_{pr}	20	1000
	K_{pr}	60	4000
Neural Network	learning rate	1e-3	1e-3
	batch size	160	120
	batch size / snr	20	30
	SNR range	1-8dB	1-6dB
	λ	100	10^{12}
	gradient clipping	0.1	0.1
	network depth	200	1200

TABLE I: Parameter Configuration of the Model

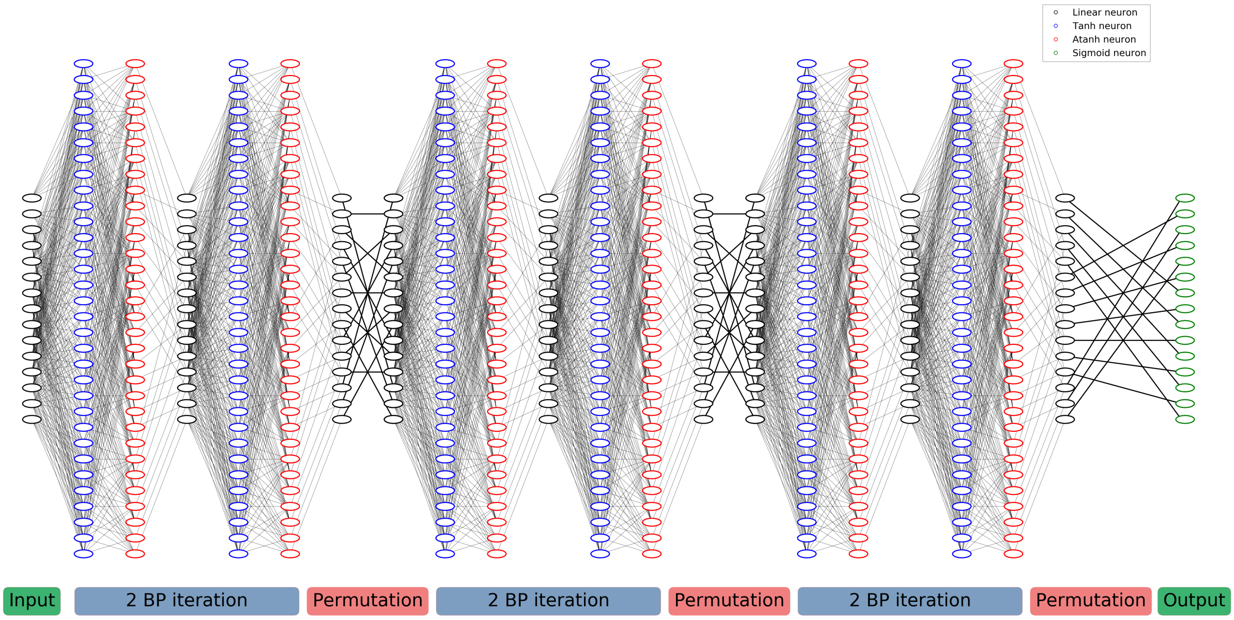


Fig. 1: Deep Neural Network Architecture For BCH(15,11) with 3 permutations and 2 belief propagation iterations for each permutation. The permutations have bold lines. The self message $o^j_{i,v}$ was removed from the diagram for a cleaner view.

D. Results

In the following figures, "Perm-RNN-i-j-k" denotes our proposed decoder, with i parallel branches, j permutations and k BP iterations between two consecutive permutations; "mRRD-i" denotes the classical mRRD decoder with i branches; and "mRRD-RNN-i-j-k" denotes the mRRD-RNN decoder with i branches, j blocks of BP, each with k iterations.

In Figure 2 we provide the bit-error-rate for $BCH(63, 45)$ code for our proposed decoder. The maximum-likelihood estimate was obtained by the OSD algorithm [21]. We observe near maximum likelihood performance with our proposed decoder, with a gap of up to 0.2dB to ML. The runtime of the proposed neural decoder is lower than OSD's when SNR is bigger than 3.8dB, as shown in figure 4. In Figures 5 and 6 we provide the bit-error-rate and the running time for $BCH(63, 36)$ code for our proposed decoder. The maximum-likelihood estimate was obtained by the 2nd order OSD algorithm [21], and the mRRD performance was obtained using 10-parallel mRRD decoder [17]. We have a gap of 0.25-0.5dB to achieve maximum likelihood performance with our proposed decoder.

Note, that the overall decoding time of our decoder is substantially smaller than the mRRD's decoding time for the (63,36) code, with a factor of up to 3.5. In addition, only one neural decoder was needed to match the performance 10-parallel mRRD decoder. Also note, that OSD's main disadvantage of parallel implementation is not encountered in the neural decoder.

In Figure 3 we provide the learning curve for $BCH(63, 45)$ code. The learning rate was constant during the training process, yet the loss significantly drops at some stage of the training. For training without l_2-norm, the drop occurs in epoch 265, and most of the improvement occurs at the same time. Training with l_2-norm accelerated the learning process: the loss dropped at epoch 8, as if the training process was accelerated by factor 33. We will investigate and discuss the dropping phenomenon and the l_2-acceleration at section V.

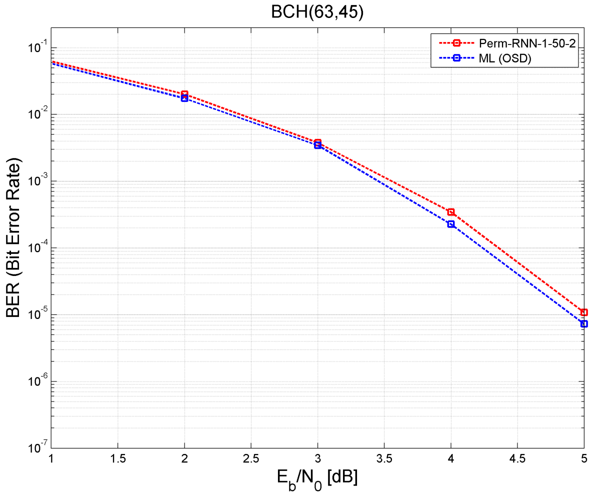


Fig. 2: BER results for BCH(63,45) code

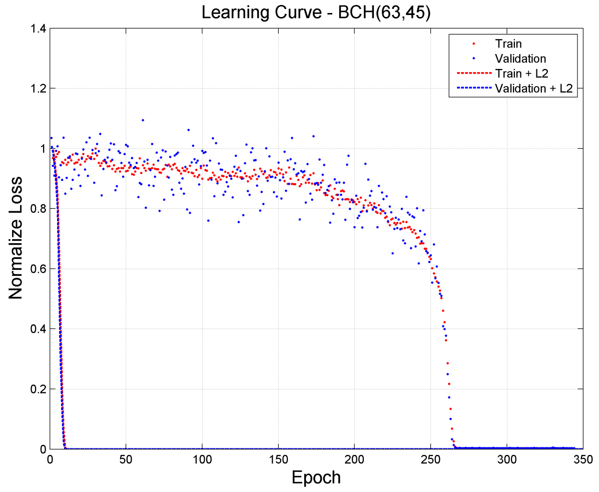


Fig. 3: Learning Curve for BCH(63,45) code

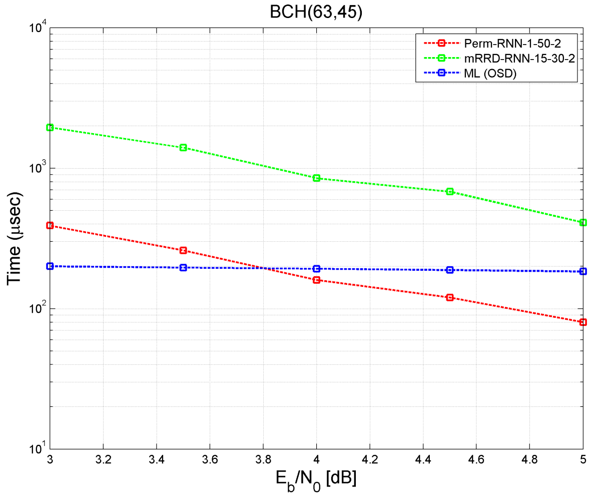


Fig. 4: Running time comparison for BCH(63,45) code

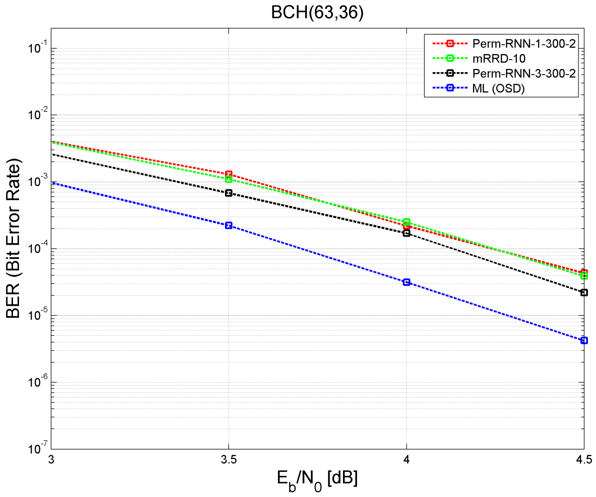


Fig. 5: BER results for BCH(63,36) code

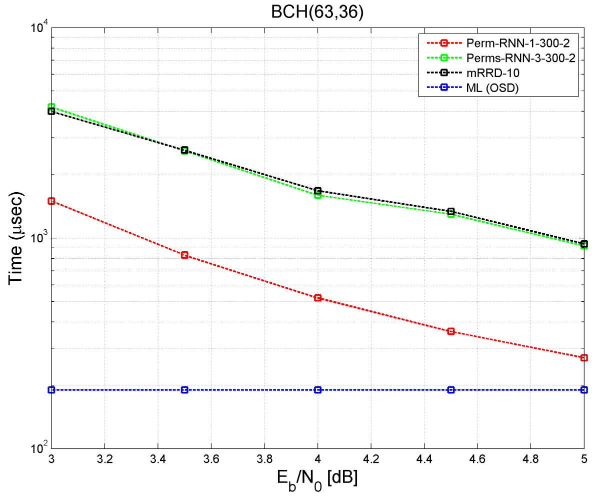


Fig. 6: Running time comparison for BCH(63,36) code

V. Training Acceleration

As introduced in the previous section, during the training the loss drops significantly. This phenomenon usually occurs while training very deep neural networks. Note that our proposed network for $BCH(63,45)$ contained 200 layers. As shown in [22], this phenomenon can be explained by the existence of saddle points in the loss-surface of the network.

A. Hessian simulation

To further investigate the phenomenon of the significant loss-drop and the l_2 acceleration, we computed the Hessian matrix of the deep neural decoder. Since the Hessian calculation demands high resources, we investigated the training process of a similar and smaller code, $BCH(31,16)$. As shown in Figures 7 and 8, the training process of the $BCH(31,16)$ behaves in the same manner as the $BCH(63,45)$ code.

The Hessian matrix was evaluated during the training process. We calculated the condition number and the distribution of the eigenvalues of the Hessian matrix. The setting for the $BCH(31,16)$ code was: $I_{permutations} = 10$ permutations, $I_{BP} = 2$ BP-iterations, $c_{grad} = 0.1$, $\lambda = 100$.

Figures 7 and 9 demonstrate the significant loss-drop properties. Whereas in epochs 1-20 the loss and the BER do not improve significantly and the positive eigenvalues ratio is low, epoch 20-40 serves as a turning point: the loss and the BER decrease rapidly and the positive eigenvalues ratio increases at the same time.

Figures 8 and 9 further stress this matter: the positive eigenvalues ratio is high right from the beginning, and accordingly the loss presents no initial-plateau to begin with. Put in other words, the Hessian rapidly becomes similar to a scaled identity matrix. The equivalent loss-surface is isotropic, which results in an accelerated learning process.

It is of no surprise that adding an l_2 term to the loss brings the Hessian closer to an identity matrix. Yet, the notable training acceleration and the performance improvement are a result of a gentle setting of parameters and the specific optimization problem discussed.

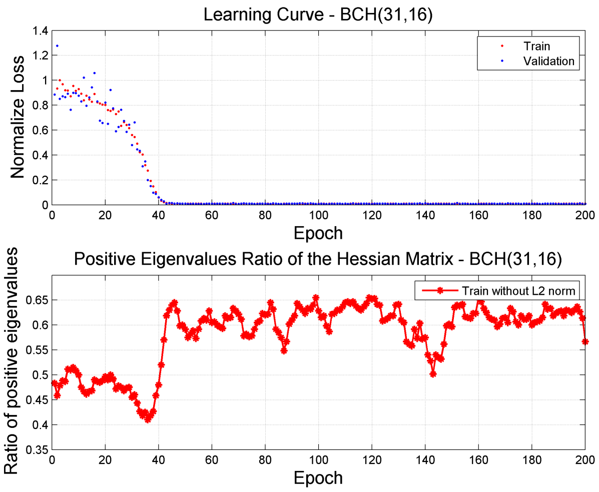


Fig. 7: Learning Curve and Positive Eigenvalues Ratio of the Hessian Matrix for BCH(31,16) For Training without l_2-norm

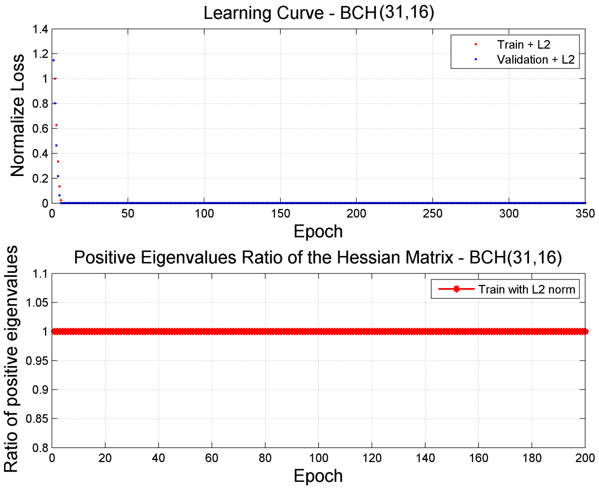


Fig. 8: Learning Curve and Positive Eigenvalues Ratio of the Hessian Matrix for BCH(31,16) For Training with l_2-norm

References

[1] N. Farsad and A. Goldsmith, “Detection algorithms for communication systems using deep learning,” *arXiv preprint arXiv:1705.08044*, 2017.

[2] T. J. O’Shea and J. Hoydis, “An introduction to machine learning communications systems,” *arXiv preprint arXiv:1702.00832*, 2017.

[3] N. Samuel, T. Diskin, and A. Wiesel, “Deep mimo detection,” *arXiv preprint arXiv:1706.01151*, 2017.

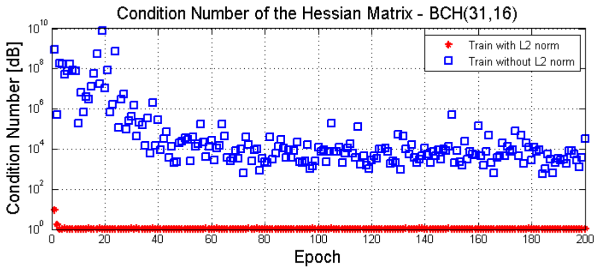


Fig. 9: Condition Number of the Hessian Matrix during training for BCH(31,16) code

[4] F. Liang, C. Shen, and F. Wu, “An iterative bp-cnn architecture for channel decoding,” *arXiv preprint arXiv:1707.05697*, 2017.

[5] S. Dorner, S. Cammerer, J. Hoydis, and S. ten Brink, “Deep learning-based communication over the air,” *arXiv preprint arXiv:1707.03384*, 2017.

[6] P. Shengliang, J. Hanyu, W. Huaxia, and Y. Yu-Dong, “Deep learning and its applications in communications systems - modulation classification,” *Submitted to IEEE Communications Magazine*, 2017.

[7] Y. Hao, Y. L. Geoffrey, and F. J. Biing-Hwang, “Power of deep learning for channel estimation and signal detection in ofdm systems,” *arXiv preprint arXiv:1708.08514*, 2017.

[8] H. Sihao and L. Haowen, “Fully optical spacecraft communications: Implementing an omnidirectional pv-cell receiver and 8mb/s led visible light downlink with deep learning error correction,” *arXiv preprint arXiv:1709.03222*, 2017.

[9] E. Nachmani, Y. Be’ery, and D. Burshtein, “Learning to decode linear codes using deep learning,” in *54’th Annual Allerton Conf. On Communication, Control and Computing*, September 2016, arXiv preprint arXiv:1607.04793.

[10] E. Nachmani, E. Marciano, D. Burshtein, and Y. Be’ery, “Rnn decoding of linear block codes,” *arXiv preprint arXiv:1702.07560*, 2017.

[11] L. Lugosch and W. J. Gross, “Neural offset min-sum decoding,” in *2017 IEEE International Symposium on Information Theory*, June 2017, arXiv preprint arXiv:1701.05931.

[12] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Beery, “Deep learning methods for improved decoding of linear codes,” *IEEE Journal of Selected Topics In Signal Processing*, Feb. 2018.

[13] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, “On deep learning-based channel decoding,” *accepted for CISS 2017, arXiv preprint arXiv:1701.07738*, 2017.

[14] S. Cammerer, T. Gruber, J. Hoydis, and S. ten Brink, “Scaling deep learning-based decoding of polar codes via partitioning,” *arXiv preprint arXiv:1702.06901*, 2017.

[15] S. Krastanov and L. Jiang, “Deep neural network probabilistic decoder for stabilizer codes,” *arXiv preprint arXiv:1705.09334*, 2017.

[16] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977.

[17] I. Dimnik and Y. Be’ery, “Improved random redundant iterative hdpc decoding,” *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1982–1985, 2009.

[18] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

[19] T. R. Halford and K. M. Chugg, “Random redundant soft-in soft-out decoding of linear block codes,” in *Information Theory, 2006 IEEE International Symposium on*. IEEE, 2006, pp. 2230–2234.

[20] C.-C. Lu and L. R. Welch, “On automorphism groups of binary primitive bch codes,” in *Proc. IEEE Symposium on Information Theory*, June 1994, p. 1951.

[21] M. P. Fossorier and S. Lin, “Soft-decision decoding of linear block codes based on ordered statistics,” *IEEE Transactions on Information Theory*, vol. 41, no. 5, pp. 1379–1396, 1995.

[22] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, “Identifying and attacking the saddle point problem in high-dimensional non-convex optimization,” in *Advances in neural information processing systems*, 2014, pp. 2933–2941.

Detection Over Unknown Channels via Machine Learning

Nariman Farsad and Andrea Goldsmith
Electrical Engineering, Stanford University

I. Introduction

One of the important modules in reliable recovery of data sent over a communication channel is the detection algorithm, where the transmitted signal is estimated from a noisy and corrupted version observed at the receiver. The design and analysis of this module has traditionally relied on mathematical models that describe the transmission process, signal propagation, receiver noise, and many other components of the system that affect the end-to-end signal transmission and reception. However, there are cases where tractable mathematical descriptions of the channel are elusive, either because the EM signal propagation is very complicated or when it is poorly understood. Even when the underlying channel models are known, since the channel conditions may change with time, many model-based detection algorithms rely on the estimation of the instantaneous channel state information (CSI) (i.e., channel model parameters) for detection. This estimation process typically entails overhead that decreases the data transmission rate. Moreover, the accuracy of the estimation effects the performance of the detection algorithm.

We demonstrate that, using known neural network (NN) architectures such as a recurrent neural network (RNN) [1], it is possible to train a detector without any knowledge of the underlying system model. In this scheme, the receiver goes through a training phase where a NN detector is trained using known transmission signals. We also propose a real-time sequence detector, which we call the *sliding bidirectional RNN (SBRNN) detector*, that detects the symbols corresponding to a data stream, as they arrive at the destination. This technique could be extended to any type of real-time estimation of data streams. We demonstrate that training the SBRNN on a diverse dataset that contains transmission sequences in different channel conditions yields a detector that is resilient to changing channel conditions and outperforms the Viterbi detector (VD) with CSI estimation error.

II. Sliding BRNN Detector Performance

Let L be the maximum length of the BRNN. For this maximum length, during training, blocks of $\ell \leq L$ consecutive transmissions are used for training. Note that sequences of different length could be used during training as long as all sequence lengths are smaller than or equal to L. Inspired by some of the techniques used in speech recognition, we propose a dynamic programing scheme we call the *sliding BRNN (SBRNN) detector*. The first $\ell \leq L$ symbols are detected using the BRNN. Then as each new symbol arrives at the destination, the position of the BRNN slides ahead by one symbol. Let the set $\mathcal{J}_k = \{j \mid j \leq k \ \wedge \ j + L > k\}$ be the set of all valid starting positions for a BRNN detector of length L, such that the detector overlaps with the k^{th} symbol. For example, if $L = 3$ and $k = 4$, then $j = 1$ is not in the set $\mathcal{J}_k$ since the BRNN detector overlaps with symbol positions 1, 2, and 3, and not the symbol position 4. Let $\hat{\mathbf{p}}_k^{(j)}$ be the estimated PMF for the k^{th} symbol, when the start of the sliding BRNN is on $j \in \mathcal{J}_k$. The final PMF corresponding to the k^{th} symbol is given by the weighted sum of the estimated PMFs for each of the relevant windows: $\hat{\mathbf{p}}_k = \frac{1}{|\mathcal{J}_k|} \sum_{j \in \mathcal{J}_k} \hat{\mathbf{p}}_k^{(j)}$. One of the main benefits of this approach is that, after the first L symbols are received and detected, as the signal corresponding to a new symbol arrives at the destination, the detector immediately estimates that symbol. The detector also updates its estimate for the previous $L - 1$ symbols dynamically. Therefore, this algorithm is similar to a dynamic programming algorithm.

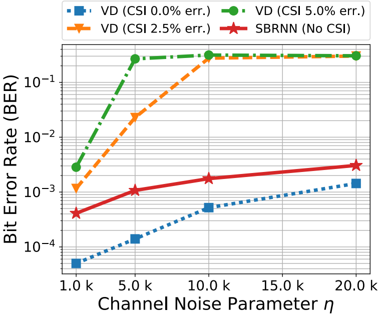


Fig. 1. The BER for different values of noise rates η.

To evaluate the performance of the SBRNN we consider the Poisson channel, which is used to model optical and molecular communication systems. Figure 1 compares the performance of the SBRNN to the VD with perfect CSI (i.e., the maximum-likelihood detector), as well as to the VDs with 2.5 and 5 percent error in CSI estimation. We see that the SBRNN outperforms VD with estimation error and comes close to the performance of the VD with perfect CSI estimation. More details can be found in [2].

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] N. Farsad and A. Goldsmith, “Neural network detection of data sequences in communication systems,” *IEEE Transactions on Signal Processing*, 2018, submitted. [Online]. Available: http://narimanfarsad.com/papers/NN%20Detectors%20Journal.pdf

Learning to Optimize: Training Deep Neural Networks for Interference Management

Haoran Sun*, Xiangyi Chen*, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D. Sidiropoulos

I. Introduction

We consider an interference management problem for an interference channel consisting of K single-antenna transceiver pairs. Let $h_{kk} \in \mathbb{C}$ denote the direct channel between transmitter k and receiver k, and $h_{kj} \in \mathbb{C}$ denote the interference channel from transmitter j to receiver k. Furthermore, we assume that the transmitted symbol of each transmitter k is an independent Gaussian random variable with zero mean and variance p_k (which is also referred to as the transmission power of transmitter k). Then the signal to interference-plus-noise ratio (SINR) for each receiver k is given by

$$\text{sinr}_k \triangleq \frac{|h_{kk}|^2 p_k}{\sum_{j \neq k} |h_{kj}|^2 p_j + \sigma_k^2},$$

where σ_k^2 denotes the noise power at receiver k.

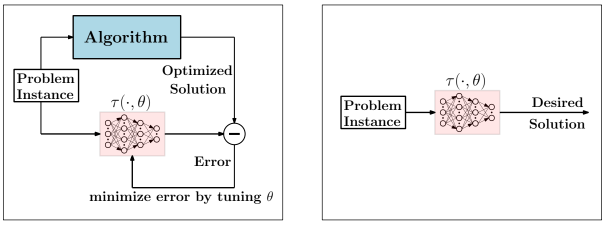
To optimally allocate power for each transmitter so that the weighted system throughput is maximized, the problem can be formulated as the following *nonconvex* problem

$$\begin{aligned} \max_{p:=\{p_1,\ldots,p_K\}} \quad & \sum_{k=1}^{K} \alpha_k \log\left(1 + \frac{|h_{kk}|^2 p_k}{\sum_{j\neq k} |h_{kj}|^2 p_j + \sigma_k^2}\right) \\ \text{s.t.} \quad & 0 \leq p_k \leq P_{\max},\ \forall\, k = 1, 2, \ldots, K, \end{aligned} \tag{1}$$

where $P_{\max}$ denotes the power budget of each transmitter and α_k denotes the nonnegative weight. Problem (1) is known to be NP-hard [1]. To obtain a good solution for problem (1), many transceiver design algorithms developed in the literature, such as WMMSE [2], SCALE [3], and the pricing algorithm [4]. A particular version of the WMMSE [5, Figure 1] also applied to solve the power control problem (1). However, optimization algorithms often entail considerable complexity, which creates a serious gap between theoretical design/analysis and real-time processing.

II. The Learning to Optimize Approach

To resolve the computational issues arisen on the above interference management problem with stringent real-time requirements, we design a new 'learning to optimize' based framework as shown in Figure 1. The main idea is to treat a given algorithm as a "black box", and try to *learn* its input/output relation by using a *deep neural network* (DNN) [6]. If the nonlinear mapping can be learned accurately by a DNN of moderate size, then the interference management tasks can be performed in almost *real time* – since passing the input through a DNN only requires a small number of simple operations.



(a) Training Stage

(b) Testing Stage

Fig. 1: The Proposed Method. The key idea is to treat the input and output of an algorithm as an unknown nonlinear mapping and use a DNN to approximate it. In the figure $\tau(\cdot,\theta)$ represent a DNN parameterized by θ.

Unlike all existing works on approximating optimization algorithms such as those using unfolding [7]–[10] , our approach is justified by rigorous theoretical analysis. We show that there are conditions under which an algorithm is learnable by a DNN [5], and indicate that it is possible to learn a well-defined optimization algorithm very well by using finite-sized deep neural networks. To concisely state the result, let us make the following definitions. Given an input channel vector $h := \{h_{ij}\} \in \mathbb{R}^{K^2}$, let us use $v(h)_i^t$ to denote the variable v_i at t^{th} iteration generated by WMMSE [5] (which basically represents $\sqrt{p_i}$ at tth iteration). Also let $H_{\min}, H_{\max} > 0$ denote the minimum and maximum channel strength and let $V_{\min} > 0$ be a given positive number. Let $NET(x, z)$ represent a neural network with (x, z) as input.

Theorem 1 *Suppose that WMMSE is randomly initialized with $(v_k^0)^2 \leq P_{\max}$, $\sum_{i=1}^{K} v(h)_i^0 \geq V_{\min}$, and it is executed for T iterations. Define the following set of 'admissible' channel realizations*

$$\mathcal{H} := \left\{ h \mid H_{\min} \leq |h_{jk}| \leq H_{\max}, \forall j, k,\ \sum_{i=1}^{K} v(h)_i^t \geq V_{\min}, \forall t \right\}.$$

Given $\epsilon > 0$, there exists a neural network with $h \in \mathbb{R}^{K^2}$ and $v^0 \in \mathbb{R}_+^K$ as input and $NET(h, v^0) \in \mathbb{R}_+^K$ as output, with the following number of layers

$$O\left(T^2 \log\left(\max\left(K, P_{\max}, H_{\max}, \frac{1}{\sigma}, \frac{1}{H_{\min}}, \frac{1}{P_{\min}}\right)\right) + T \log\left(\frac{1}{\epsilon}\right)\right)$$

and the following number of ReLUs and binary units

$$O\left(T^2 K^2 \log\left(\max\left(K, P_{\max}, H_{\max}, \frac{1}{\sigma}, \frac{1}{H_{\min}}, \frac{1}{P_{\min}}\right)\right) + TK^2 \log\left(\frac{1}{\epsilon}\right)\right),$$

TABLE I: Sum-Rate and Computational Performance for IMAC

# of base stations	average sum-rate (bit/sec.)			total CPU time (sec.)			
and users (N,K)	DNN	WMMSE	DNN/WMMSE	DNN	WMMSE (MATLAB)	WMMSE(C)	DNN/WMMSE (C)
$(3, 12)$	17.722	18.028	98.30%	0.021	22.33	0.27	7.78%
$(3, 18)$	20.080	20.606	97.45%	0.022	42.77	0.48	4.58%
$(3, 24)$	21.928	22.648	96.82%	0.025	67.59	0.89	2.81%
$(7, 28)$	33.513	35.453	94.53%	0.038	140.44	2.41	1.58%
$(20, 80)$	79.357	87.820	90.36%	0.141	890.19	23.0	0.61%

such that the relation below holds true

$$\max_{h\in\mathcal{H}}\max_{i} |(v(h)_i^T)^2 - NET(h, v^0)_i| \leq \epsilon \tag{2}$$

Remark 1 *The bounds in Theorem 1 provide an intuitive understanding of how the size of the network should be dependent on various system parameters. A key observation is that having a neural network with multiple layers is essential in achieving our rate bounds. Another observation is that the effect of the approximation error on the size of the network is rather minor [the dependency is in the order of $\mathcal{O}(\log(1/\epsilon))$]. However, we do want to point out that the numbers predicted by Theorem 1 represent some upper bounds on the size of the network. In practice, much smaller networks are often used to achieve the best tradeoff between computational speed and solution accuracy.*

III. Numerical Results

To demonstrate the achievable performance of the proposed approach, a multi-cell interfering multiple Access Channel (IMAC) model is considered with a total of N cells and K users. In each cell, one BS is placed at the center of the cell and the users are randomly and uniformly distributed in the area; The channel between each user and each BS is randomly generated according to a Rayleigh fading distribution; see [5] for more detail. We perform the training and testing following the procedures outlined in Figure 1 and summarize the testing results in TABLE I. It can be seen that the proposed DNN approach can be trained to well-approximate the behavior of the state-of-the-art algorithm WMMSE [2], and achieve relatively high sum-rate performance. It is also shown that DNNs can achieve orders of magnitude speedup in computational time compared to state-of-the-art power allocation algorithms based on optimization.

Note that in the table, WMMSE (C)/(MATLAB) represents the WMMSE algorithm implemented using either C or MATLAB. The proposed DNN approach is implemented in Python 3.6.0 with TensorFlow 1.0.0 on one computer node with two 8-core Intel Haswell processors, two Nvidia K20 Graphical Processing Units (GPUs), and 128 GB of memory. The GPUs are used in the training stage to reduce the training time, but are *not* used in the testing stage.

References

[1] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, 2008.

[2] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[3] J. Papandriopoulos and J. S. Evans, "SCALE: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3711–3724, 2009.

[4] D. Schmidt, C. Shi, R. Berry, M. Honig, and W. Utschick, "Distributed resource allocation schemes," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 53 –63, 2009.

[5] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management," in *Signal Processing Advances in Wireless Communications (SPAWC), 2017 IEEE 18th International Workshop on*. IEEE, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 399–406.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imgaging Science*, vol. 2, no. 1, pp. 183 – 202, 2009.

[9] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint arXiv:1409.2574*, 2014.

[10] P. Sprechmann, R. Litman, T. B. Yakar, A. M. Bronstein, and G. Sapiro, "Supervised sparse analysis and synthesis operators," in *Advances in Neural Information Processing Systems*, 2013, pp. 908–916.

Communication Algorithms via Deep Learning

Hyeji Kim*, Yihan Jiang†, Ranvir Rana*, Sreeram Kannan†, Sewoong Oh*, Pramod Viswanath*
University of Illinois at Urbana Champaign*, University of Washington†

Reliable digital communication, both wireline (ethernet, cable and DSL modems) and wireless (cellular, satellite, deep space), is a primary workhorse of the modern information age. A critical aspect of reliable communication involves the design of codes that allow transmissions to be robustly (and computationally efficiently) decoded under noisy conditions. This is the discipline of coding theory; over the past century and especially the past 70 years (since the birth of information theory [1]) much progress has been made in the design of near optimal codes. Landmark codes include convolutional codes, turbo codes, low density parity check (LDPC) codes and, recently, polar codes. The impact on humanity is enormous – every cellular phone designed uses one of these codes, which feature in global cellular standards ranging from the 2nd generation to the 5th generation respectively, and are text book material [2].

The canonical setting is one of point-to-point reliable communication over the additive white Gaussian noise (AWGN) channel and performance of a code in this setting is its gold standard. The AWGN channel fits much of wireline and wireless communications although the front end of the receiver may have to be specifically designed before being processed by the decoder (example: intersymbol equalization in cable modems, beamforming and sphere decoding in multiple antenna wireless systems); again this is text book material [3]. There are two long term goals in coding theory: (a) design of new, computationally efficient, codes that improve the state of the art (probability of correct reception) over the AWGN setting. Since the current codes already operate close to the information theoretic "Shannon limit", the emphasis is on *robustness* and *adaptability* to deviations from the AWGN settings (a list of channel models motivated by practical settings, (such as urban, pedestrian, vehicular) in the recent 5th generation cellular standard is available in Annex B of 3GPP TS 36.101. (b) design of new codes for multi-terminal (i.e., beyond point-to-point) settings – examples include the feedback channel, the relay channel and the interference channel.

Progress over these long term goals has generally been driven by individual human ingenuity and, befittingly, is sporadic. For instance, the time duration between convolutional codes (2nd generation cellular standards) to polar codes (5th generation cellular standards) is over 4 decades. Deep learning is fast emerging as capable of learning sophisticated algorithms from observed data (input, action, output) alone and has been remarkably successful in a large variety of human endeavors (ranging from language [4] to vision [5] to playing Go [6]). Motivated by these successes, we posit that deep learning methods can play a crucial role in solving both the aforementioned goals of coding theory and show that we can make significant progress on both these goals in this work.

While the learning framework is clear and there is virtually unlimited training data available, there are two main challenges: (a) The space of codes is very vast and the sizes astronomical; for instance a rate 1/2 code over 100 information bits involves designing 2^{100} codewords in a 200 dimensional space. Computationally efficient encoding and decoding procedures are a must, apart from high reliability over the AWGN channel. (b) Generalization is highly desirable across block lengths and data rate that each work very well over a wide range of channel signal to noise ratios (SNR). In other words, one is looking to design a family of codes (parametrized by data rate and number of information bits) and their performance is evaluated over a range of channel SNRs.

In part due to these challenges, recent deep learing works on coding theory focus on decoding known codes using data-driven neural decoders for short block lengths [7, 8, 9]. The main challenge is to restrict oneself to a class of codes that neural networks can naturally encode and decode. In this work, we restrict ourselves to a class of *sequential* encoding and decoding schemes, of which convolutional and turbo codes are part of. These sequential coding schemes naturally meld with the family of recurrent neural network (RNN) architectures, which have recently seen large success

*H. Kim, R. Rana and P. Viswanath are with Coordinated Science Lab and Department of Electrical Engineering at University of Illinois at Urbana Champaign. S. Oh is with Coordinated Science Lab and Department of Industrial and Enterprise Systems Engineering at University of Illinois at Urbana Champaign. Email: {hyejikim,rbrana2,swoh,pramodv}@illinois.edu (H. Kim, R. Rana, S.Oh, and P.Viswanath)

†Y. Jiang and S. Kannan are with the Department of Electrical Engineering at University of Washington. Email: yihanrogerjiang@gmail.com (Y. J), ksreeram@uw.edu (S. K.)

The full manuscript of this work is under review at the 6th International Conference on Learning Representations.

in a wide variety of time-series tasks. The ancillary advantage of sequential schemes is that arbitrarily long information bits can be encoded and also at a large variety of coding rates. Working within sequential codes parametrized by RNN architectures, we make the following contributions.

(1) Focusing on *convolutional codes* we aim to decode them on the AWGN channel using RNN architectures. Efficient optimal decoding of convolutional codes has represented historically fundamental progress in the broad arena of algorithms; optimal bit error decoding is achieved by the 'Viterbi decoder' [10] which is simply dynamic programming or Dijkstra's algorithm on a specific graph (the 'trellis') induced by the convolutional code. Optimal block error decoding is the BCJR decoder [11] which is part of a family of forward-backward algorithms. While early work had shown that vanilla-RNNs are capable in *principle* of emulating both Viterbi and BCJR decoders [12, 13] we show empirically, through a careful construction of RNN architectures and training methodology, that neural network decoding is possible at very near optimal performances (both bit error rate (BER) and block error rate (BLER)). The key point is that we train a RNN decoder at a *specific* SNR and over *short information bit* lengths (100 bits) and show *strong generalization* capabilities by testing over a wide range of SNR and block lengths (up to 10,000 bits). The specific training SNR is closely related to the Shannon limit of the AWGN channel at the rate of the code and provides strong information theoretic collateral to our empirical results.

(2) *Turbo codes* are naturally built on top of convolutional codes, both in terms of encoding and decoding. A natural generalization of our RNN convolutional decoders allow us to decode turbo codes at BER comparable to, and at certain regimes, even *better* than state of the art turbo decoders on the AWGN channel. That data driven, SGD-learnt, RNN architectures can decode comparably is fairly remarkable since turbo codes already operate near the Shannon limit of reliable communication over the AWGN channel.

(3) We show the afore-described neural network decoders for both convolutional and turbo codes are *robust* to variations to the AWGN channel model. We consider a problem of contemporary interest: communication over a "bursty" AWGN channel (where a small fraction of noise has much higher variance than usual) which models inter-cell interference in OFDM cellular systems (used in 4G and 5G cellular standards) or co-channel radar interference. We demonstrate empirically the neural network architectures can adapt to such variations and beat state of the art heuristics comfortably (despite evidence elsewhere that neural network are sensitive to models they are trained on [14]). Via an innovative local perturbation analysis (akin to [15]), we demonstrate the neural network to have learnt sophisticated preprocessing heuristics in engineering of real world systems [16].

(4) We demonstrate new RNN-driven encoders (with matching decoders) that operate significantly better than state of the art on the AWGN channel with (noisy) output feedback. While feedback does not improve the Shannon capacity of the AWGN channel [17], it is known to provide better reliability at finite block lengths [18], although very sensitive to even tiny amounts of noise in the output feedback; more generally any linear code incorporating the noisy output feedback cannot achieve a non-zero reliable rate of communication [19] – this is very troubling since all practical codes are linear and linear codes are known to achieve capacity (without feedback) [20]. Our RNN parameterized encoders are inherently *nonlinear* and map information bits *directly* to real-valued transmissions. Their performance vastly improves the state of the art on the long standing open problem in information theory on communicating over the AWGN channel with noisy output feedback.

References

[1] C. E. Shannon, "A mathematical theory of communication, part i, part ii," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[2] T. Richardson and R. Urbanke, *Modern coding theory*. Cambridge university press, 2008.

[3] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[7] T. J. O'Shea and J. Hoydis, "An introduction to machine learning communications systems," *CoRR*, vol. abs/1702.00832, 2017. [Online]. Available: http://arxiv.org/abs/1702.00832

[8] S. Dörner, S. Cammerer, J. Hoydis, and S. t. Brink, "Deep learning-based communication over the air," *arXiv preprint arXiv:1707.03384*, 2017.

[9] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *Information Sciences and Systems (CISS), 2017 51st Annual Conference on*. IEEE, 2017, pp. 1–6.

[10] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[11] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," *IEEE Transactions on information theory*, vol. 20, no. 2, pp. 284–287, 1974.

[12] X.-A. Wang and S. B. Wicker, "An artificial neural net viterbi decoder," *IEEE Transactions on Communications*, vol. 44, no. 2, pp. 165–171, Feb 1996.

[13] M. H. Sazlı and C. Icsık, "Neural network implementation of the bcjr algorithm," *Digital Signal Processing*, vol. 17, no. 1, pp. 353 – 359, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200406000029

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 1135–1144. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939778

[16] J. Li, X. Wu, and R. Laroia, *OFDMA mobile broadband communications: A systems approach*. Cambridge University Press, 2013.

[17] C. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.

[18] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–i: No bandwidth constraint," *IEEE Transactions on Information Theory*, vol. 12, no. 2, pp. 172–182, 1966.

[19] Y.-H. Kim, A. Lapidoth, and T. Weissman, "The gaussian channel with noisy feedback," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE, 2007, pp. 1416–1420.

[20] P. Elias, "Coding for noisy channels," in *IRE Convention record*, vol. 4, 1955, pp. 37–46.

End-to-end Learning for Physical Layer Communications

Sebastian Cammerer*, Sebastian Dörner*, Jakob Hoydis† and Stephan ten Brink*
* Institute of Telecommunications, Pfaffenwaldring 47, University of Stuttgart, 70659 Stuttgart, Germany
{cammerer,doerner,tenbrink}@inue.uni-stuttgart.de
†Nokia Bell Labs, Route de Villejust, 91620 Nozay, France, jakob.hoydis@nokia-bell-labs.com

I. Introduction

Since Shannon's groundbreaking work on the fundamental limits of communications [1], engineers have been seeking to solve the task of "reproducing at one point either exactly or approximately a message selected at another point" [1] or, in other words, reliably transmitting a message from a source to a destination over a channel by the use of a transmitter and a receiver as illustrated in Fig. 1. *"Classical"* block-based signal processing has shown to be close to optimal while each sub-block can be optimized individually for a specific task such as equalization, modulation or channel coding.

At first glance, machine learning techniques do not appear to be a good match to communications on the physical layer, with 50 years of tremendous progress based on classic signal processing, communication and information theory, approaching close-to-optimal Shannon limit performance on many channels. However, several open problems remain, e.g., pertaining adaptivity and complexity of joint processing, where first results using machine learning-based approaches are promising (see [2], [3] and references therein).

Recently, the idea of deep learning (DL)-based communication was proposed in the literature [3], [4] based on the autoencoder concept ([5, Ch. 14]). In contrast to component-wise optimizations, the autoencoder approach now enables end-to-end training over any type of channel without the need for detailed prior mathematical abstraction of the channel model, breaking up restrictions commonplace in conventional block-based signal processing by moving away from handcrafted, carefully optimized sub-blocks towards adaptive and flexible (artificial) neural networks, leading to many attractive research questions. The benefits of machine learning approaches may include more flexible hardware, highly adaptive systems and less overall complexity. We thus pose the seemingly naive, yet, in fact, rather complicated and attractive research question: *Can we learn to communicate?*

We demonstrate the practical potential and viability of such a system by extending the idea of end-to-end learning of communications systems through deep neural network-based autoencoders to orthogonal frequency division multiplex (OFDM) with cyclic prefix (CP). This allows learning of transmitter and receiver implementations—without any prior knowledge—that are optimized for an arbitrary differentiable end-to-end performance metric, e.g., block error rate (BLER). Our implementation shares the same benefits as a conventional OFDM system, namely single-tap equalization and robustness against sampling synchronization errors, which turned out to be one of the major challenges in prior single-carrier implementations [6]. We show that the proposed scheme can be realized with state-of-the-art deep learning software libraries, since transmitter and receiver solely consist of differentiable layers required for gradient-based training.

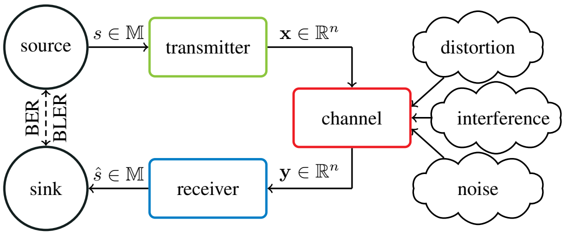


Fig. 1: Illustration of a simple communications system.

II. Autoencoder-based Communication

As described in [3], a communications system can be interpreted as an autoencoder [5]. This is schematically shown in Fig. 2. An autoencoder describes a deep neural network consisting of various hidden layers that is trained to reconstruct the input (a so-called one-hot encoded vector representing one of the m possible messages) at the output. As the information must pass each layer, the network needs to find a robust representation of the input message at every layer. In particular, the transmitter output (a real vector of dimension n) must be robust with respect to various channel impairments. Note that the channel is also represented by network layers (without trainable weights) that carry out stochastic transformations of the input data. It is crucial to have a good model that accurately reflects the real channel. The autoencoder is trained end-to-end using stochastic gradient descent (SGD). After training, the transmitter and receiver are fully described by their respective layer dimensions and weights and can operate in standalone mode to generate/process radio signals, e.g., on a software-defined radio (SDR) platform as shown in [6].

During training, the encoder part of the autoencoder has learned robust symbol sequence representations of all messages. Fig. 3 shows constellation diagrams of the IQ-symbols

We would like to thank Maximilian Arnold for many helpful discussions.

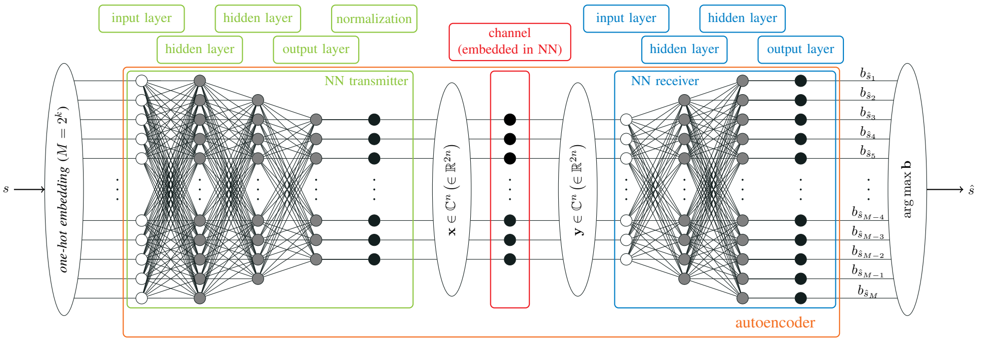


Fig. 2: Illustration of an end-to-end communications system as an autoencoder.

of all of the $m = 256$ possible messages of the single-carrier system, i.e., per subcarrier of the multi-carrier system. Each diagram shows all symbols at the same symbol position within a message, as each message consists of $\frac{1}{2}\log_2(m) = 4$ complex-valued IQ-symbols (we assume $n = 8$ and consider the first half of the transmitter output as the real and the second half as the imaginary part). Interestingly, we can observe that the autoencoder has learned some form of superimposed piloting since the center of the constellations is shifted away from the origin. For further details we refer to [6].

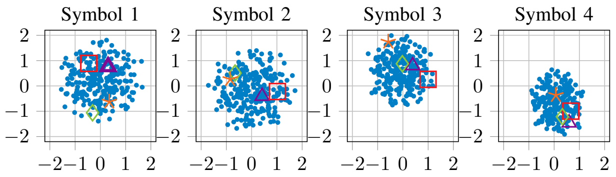


Fig. 3: Scatter plot of the learned constellations for all $M = 256$ messages using average power normalization $\|\mathbf{x}\|^2 \leq n$. The symbols of four individual messages are highlighted by different color markers.

III. OFDM Extensions

We extend our work of [6] from single-carrier to multi-carrier, i.e., OFDM with CP as shown in Fig. 4. Note that a single autoencoder message $\mathbf{x}$ is represented by $\frac{n}{2}$ complex-valued IQ-symbols. Instead of directly transmitting the encoder's output $\mathbf{x}$, an inverse discrete Fourier-transform (DFT) of width w_{FFT} is applied on a set of w_{FFT} independent autoencoder messages, i.e., w_{FFT} equivalent independent sub-channels are created, where independent autoencoder messages are assigned to each subcarrier.[1] As each autoencoder still requires $\frac{n}{2}$ channel uses, we generate $\frac{n}{2}$ complex-valued OFDM symbols $\mathbf{x}_{\text{OFDM}}$, each of length w_{FFT}. For additional robustness against sampling synchronization errors and to avoid inter-symbol interference (ISI) , we further add a CP of length ℓ_{CP}, i.e., w_{FFT} independent autoencoder symbols form one single OFDM symbol $\mathbf{x}_{\text{OFDM,CP}}$ of total length $w_{\text{FFT}} + \ell_{\text{CP}}$. Thus, a sequence of $\frac{n}{2}(w_{\text{FFT}} + \ell_{\text{CP}})$ complex-valued symbols is subsequently transmitted over the (mutlipath) channel.

At the receiver side, the CP can be used for frame synchronization through autocorrelation with peak detection; synchronization turned out to be a challenging step in singe-carrier autoencoder-based communication [6]. Finally, a DFT recovers the inputs for the w_{FFT} independent autoencoder receivers.

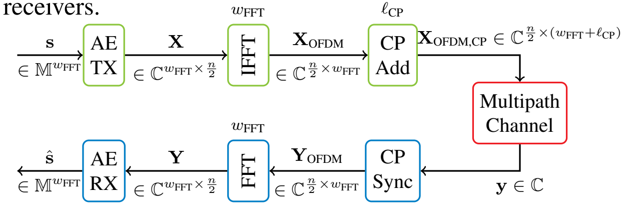

$\mathbf{s}$
$\in \mathbb{M}^{w_{\text{FFT}}}$
AE TX
$\mathbf{X}$
$\in \mathbb{C}^{w_{\text{FFT}} \times \frac{n}{2}}$
w_{FFT}
IFFT
$\mathbf{X}_{\text{OFDM}}$
$\in \mathbb{C}^{\frac{n}{2} \times w_{\text{FFT}}}$
ℓ_{CP}
CP Add
$\mathbf{X}_{\text{OFDM,CP}} \in \mathbb{C}^{\frac{n}{2} \times (w_{\text{FFT}} + \ell_{\text{CP}})}$
Multipath Channel
$\hat{\mathbf{s}}$
$\in \mathbb{M}^{w_{\text{FFT}}}$
AE RX
$\mathbf{Y}$
$\in \mathbb{C}^{w_{\text{FFT}} \times \frac{n}{2}}$
w_{FFT}
FFT
$\mathbf{Y}_{\text{OFDM}}$
$\in \mathbb{C}^{\frac{n}{2} \times w_{\text{FFT}}}$
CP Sync
$\mathbf{y} \in \mathbb{C}$

Fig. 4: OFDM extension to the autoencoder system.

At first glance it may appear counterintuitive that the autoencoder system benefits from such an explicit structure as it could also *learn* to compensate for these effects with a single (large) neural network. However, we observe for a *single* neural network that training complexity tremendously increases and practically limits the system performance (see [7]). Thus, the benefits of the proposed system are:

1) robustness against sampling synchronization errors
2) single-tap equalization[2]
3) moderate training complexity due to independent and short length sub-carrier messages (i.e., small n)

This enables reliable communication over multipath channels and makes the communication scheme suitable for commodity hardware with imprecise oscillators.

References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] M. Ibnkahla, “Applications of neural networks to digital communications: A survey,” *Signal Process.*, vol. 80, no. 7, pp. 1185–1215, 2000.

[3] T. J. O'Shea and J. Hoydis, “An introduction to machine learning communications systems,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, Dec 2017.

[4] T. J. O'Shea, K. Karra, and T. C. Clancy, “Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention,” *IEEE Int. Symp. Signal Process. Inform. Tech. (ISSPIT)*, pp. 223–228, 2016.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[6] S. Dörner, S. Cammerer, J. Hoydis, and S. ten Brink, “Deep learning-based communication over the air,” *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[7] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, “On deep learning-based channel decoding,” in *Proc. of CISS*, 2017.

[1] Remark: as no additional piloting is assumed, we cannot simply distribute the $\frac{n}{2}$ symbols of a message within the same OFDM symbol. Otherwise the unknown phase rotation per subcarrier would destroy the message.

[2] The autoencoder inherently has to *learn* how to synchronize.

Massive multiuser MIMO downlink with low-resolution converters

Sven Jacobsson[1,2], Giuseppe Durisi[1], Mikael Coldrey[2], and Christoph Studer[3]

[1] Chalmers University of Technology, Gothenburg, Sweden
[2] Ericsson Research, Gothenburg, Sweden
[3] Cornell University, Ithaca, NY, USA

***Abstract*—In this review paper, we analyze the downlink of a massive multiuser multiple-input multiple-output system in which the base station is equipped with low-resolution digital-to-analog converters (DACs). Using Bussgang's theorem, we characterize the sum-rate achievable with a Gaussian codebook and scaled nearest-neighbor decoding at the user equipments (UE). For the case of 1-bit DACs, we show how to evaluate the sum-rate using Van Vleck's arcsine law. For the case of multi-bit DACs, for which the sum-rate cannot be expressed in closed-form, we present two approximations. The first one, which is obtained by ignoring the overload (or clipping) distortion caused by the DACs, turns out to be accurate provided that one can adapt the dynamic range of the quantizer to the received-signal strength so as to avoid clipping. The second approximation, which is obtained by modeling the distortion noise as a white process, both in time and space, is accurate whenever the resolution of the DACs is sufficiently high and when the oversampling ratio is small. We conclude the paper by discussing extensions to orthogonal frequency-division multiplexing systems; we also touch upon the problem of out-of-band emissions in low-precision-DAC architectures.**

I. Introduction

Nontrivial fronthaul connectivity challenges must be solved if one wants to enable massive multiple-input multiple-output (MIMO) operation over the relatively large bandwidth available in the higher portion of the frequency spectrum assigned to 5G systems. Consider, for example, a base station (BS) equipped with 100 antennas, each one connected to two high-precision (e.g., 10-bit resolution) digital-to-analog converters (DACs) and analog-to-digital converters (ADCs) operating at 1 GS/s. In such a system, 2 Tbit/s of data would need to be transferred to and from the radio unit (typically co-located with the antenna array) to the baseband-processing unit (typically located at the base of the tower hosting the BS). This exceeds by far the rate supported by the common public radio interface (CPRI) used over today's fiber-optical fronthaul links [1].

One promising approach to reduce this fronthaul bottleneck is to lower the resolution of the data converters. Several aspects of massive MIMO systems equipped with low-precision converters have been recently investigated in the literature, including achievable rates [2]–[7], channel-estimation and data-detection algorithms [8], [9], precoding design [10]–[14], energy efficiency [15], and out-of-band spectral emissions [16]. In this review paper, we provide an overview of some of the most recent results. Our focus will be exclusively on the *downlink* of a multi-user (MU) massive MIMO system in which a BS serves multiple user equipments (UEs) concurrently in the same frequency band.

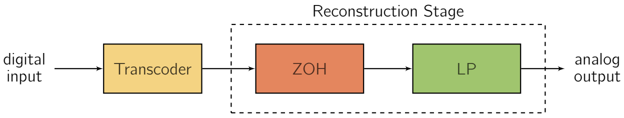


Fig. 1. Block diagram of the basic components of a DAC [17, Fig. 1.1].

II. System Model and Digital-to-Analog Converters

We consider a massive-MIMO BS equipped with B antennas and serving U UEs. Each BS antenna is fed by two DACs, which generate the in-phase and the quadrature components of the transmitted signal. A DAC performs two basic operations: (i) it transforms the digital input sequence into its analog representation (*transcoder* stage) and (ii) it maps the transcoder output to a continuous-time waveform (*reconstruction* stage, typically consisting of a zero-order hold followed by a low-pass filter [17]; see Fig. 1 for an illustration).

Under the simplifying assumption that the digital input to the DAC has infinite precision, we can view the transcoding step of the two DACs as a quantizer, i.e., a nonlinear function $\mathcal{Q}(\cdot)$ that maps a sample in $\mathbb{C}$ to a finite-cardinality set $\mathcal{X} = \{q_0, \dots, q_{2^Q-1}\} \times \{q_0, \dots, q_{2^Q-1}\}$. Here, Q is the number of DAC bits. Throughout the paper, we shall consider only symmetric, uniform quantizers and denote their step size by Δ and the number of levels by $L = 2^Q$. Furthermore, we shall assume that the output of the DACs is scaled by a factor α so as to satisfy an average transmit-power constraint.

III. Achievable Rates via Bussgang's Decomposition

To begin with, we focus, for simplicity, on the case of transmission over flat-fading channels. We also assume that the

The work of SJ and GD was supported in part by the Swedish Foundation for Strategic Research under grant ID14-0022, and by the Swedish Governmental Agency for Innovation Systems (VINNOVA) within the competence center ChaseOn. The work of CS was supported in part by Xilinx, Inc. and by the US National Science Foundation (NSF) under grants ECCS-1408006, CCF-1535897, CAREER CCF-1652065, and CNS-1717559.

DACs operate at symbol time and that their reconstruction stage involve an ideal rectangular low-pass filter (see [16] for details). Generalizations to more realistic setups are discussed in Section V. Under these assumptions, the input-output relation of the downlink channel can be modeled as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{1}$$

Here, the vector $\mathbf{y} \in \mathbb{C}^U$ contains the signal received at the U UEs; $\mathbf{H} \in \mathbb{C}^{U \times B}$ is the fading channel, which is assumed to be perfectly known at the BS. The vector $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_U)$ models the additive noise. Finally $\mathbf{x} \in \mathcal{X}^B$ is the output of the transcoder stage of the DACs.

We assume that

$$\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s}) \tag{2}$$

where $\mathbf{s} \in \mathbb{C}^U$ contains the data symbols intended for the U UEs and $\mathbf{P} \in \mathbb{C}^{B \times U}$ is the precoding matrix, which is a function of the fading channel $\mathbf{H}$. The precoding structure in (2) is referred to in [12] as *linear-quantized precoding*, to distinguish from more general nonlinear precoder structures, which offer superior performance at the cost of additional computational complexity.

By substituting (2) into (1), we see that the presence of the quantizer $\mathcal{Q}(\cdot)$ makes the channel output $\mathbf{y}$ depend on the symbol vector $\mathbf{s}$ in a nonlinear way. We next use Bussgang's theorem [18], a special case of Price's theorem [19], to linearize the input-output relation and to enable a theoretical analysis [20]. Then, we will use the *generalized mutual information* (GMI) [21] to estimate the rate achievable at each UE by scaled nearest-neighbor decoding [22] and a Gaussian codebook ensemble.

Theorem 1 below follows from a simple adaptation of Bussgang's theorem to the quantizer output $\mathcal{Q}(\mathbf{P}\mathbf{s})$.

Theorem 1: Assume that $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_U)$. Then for a fixed precoding matrix $\mathbf{P}$, we have that

$$\mathbb{E}\left[\mathcal{Q}(\mathbf{P}\mathbf{s})(\mathbf{P}\mathbf{s})^H\right] = \mathbf{G}\mathbf{P}\mathbf{P}^H \tag{3}$$

where $\mathbf{G}$ is the following real-valued diagonal matrix:[1]

$$\mathbf{G} = \frac{\alpha\Delta}{\sqrt{\pi}} \operatorname{diag}\left(\mathbf{P}\mathbf{P}^H\right)^{-1/2} \times \sum_{i=1}^{L-1} \exp\left(-\Delta^2\left(i - \frac{L}{2}\right)^2 \operatorname{diag}\left(\mathbf{P}\mathbf{P}^H\right)^{-1}\right). \tag{4}$$

It follows from (3) that, under the assumption that $\mathbf{s}$ is Gaussian, we can rewrite (2) in the linearized form

$$\mathbf{x} = \mathbf{G}\mathbf{P}\mathbf{s} + \mathbf{d} \tag{5}$$

where $\mathbf{d}$ is the zero-mean quantization-noise vector, which is uncorrelated with $\mathbf{s}$. Note that $\mathbf{G}\mathbf{P}\mathbf{s}$ is the linear minimum-mean-square estimate of $\mathbf{x}$ given $\mathbf{P}\mathbf{s}$, and $\mathbf{d}$ is the corresponding estimation error.

Substituting (5) into (1), we obtain a linear input-output relation, with non-Gaussian additive noise $\mathbf{H}\mathbf{d} + \mathbf{n}$. The ergodic rate[2] achievable over this channel using a Gaussian codebook and scaled nearest-neighbor decoding at the receiver can be established from a GMI analysis similar to the one reported in [22], [24]. Specifically, we have the following result.

Theorem 2: Assume that UE u has knowledge of the channel gain[3] $\mathbf{h}_u^T\mathbf{G}\mathbf{p}_u$, where $\mathbf{h}_u^T$ is the uth row of the channel matrix $\mathbf{H}$ and $\mathbf{p}_u$ is the uth column of the precoding matrix $\mathbf{P}$. Then the GMI R_u achievable with a Gaussian codebook and scaled nearest-neighbor decoding at the uth UE is

$$R_u = \mathbb{E}[\log(1 + \gamma_u)] \tag{6}$$

where the signal-to-interference-noise-and-distortion ratio (SINDR) γ_u is given by

$$\gamma_u = \frac{\left|\mathbf{h}_u^T\mathbf{G}\mathbf{p}_u\right|^2}{\sum_{v \neq u}\left|\mathbf{h}_u^T\mathbf{G}\mathbf{p}_v\right|^2 + \mathbf{h}_u^T\,\mathbb{E}[\mathbf{d}\mathbf{d}^H]\,\mathbf{h}_u^* + N_0}. \tag{7}$$

IV. Statistics of the Quantization Noise

Evaluating (7) requires knowledge of the correlation matrix $\mathbb{E}\left[\mathbf{d}\mathbf{d}^H\right]$ of the zero-mean quantization noise $\mathbf{d}$. It follows from (5) that

$$\mathbb{E}\left[\mathbf{d}\mathbf{d}^H\right] = \mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right] - \mathbf{G}\mathbf{P}\mathbf{P}^H\mathbf{G}. \tag{8}$$

For the case $L = 2$ (1-bit DACs), the covariance matrix $\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right]$ of the quantizer output admits a well-known closed-form expression, commonly referred to as the *arcsine law* and reported first by Van Vleck [25]:

$$\begin{aligned}\mathbb{E}&\left[\mathbf{x}\mathbf{x}^H\right] \\ &= \frac{2P}{\pi B}\left(\sin^{-1}\left(\operatorname{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}}\Re\{\mathbf{P}\mathbf{P}^H\}\operatorname{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}}\right)\right. \\ &\left.+\, j\sin^{-1}\left(\operatorname{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}}\Im\{\mathbf{P}\mathbf{P}^H\}\operatorname{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}}\right)\right).\end{aligned} \tag{9}$$

Here, P denotes the power constraint. However, for any finite L larger than 2, no closed-form expression is available for $\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right]$ and this matrix needs to be evaluated numerically (see [26], [7]).

Alternatively, one can seek closed-form approximations to $\mathbb{E}\left[\mathbf{d}\mathbf{d}^H\right]$. Two such approximations are discussed in [26]. The first one, referred to as *diagonal approximation*, involves neglecting spatial correlation, i.e., assuming that $\mathbb{E}\left[\mathbf{d}\mathbf{d}^H\right]$ is a diagonal matrix. Then, one exploits that the entries on the main diagonal of $\mathbb{E}\left[\mathbf{d}\mathbf{d}^H\right]$ can be computed in closed form even when $L > 2$. This approximation is accurate only for DACs with medium-to-high resolution (i.e., when $L \geq 4$).

The second one, referred to as *rounding approximation*, involves replacing each DAC by a one-dimensional midrise lattice quantizer (which implies $L = \infty$) with step size Δ, for which the covariance matrix of the quantization error is known in closed form [27]. This approximation is accurate also for low-precision DACs, provided that the step size Δ is chosen so that the distortion due to clipping/saturation is negligible compared to the granular quantization distortion. This requires adapting Δ to the signal strength.

[1] In (4), the operator $\operatorname{diag}(\cdot)$ returns a diagonal matrix whose main diagonal coincides with that of the matrix it is applied to; furthermore, the exponential function is applied elementwise to the diagonal entries of $\operatorname{diag}\left(\mathbf{P}\mathbf{P}^H\right)^{-1/2}$.

[2] We assume that coding can be performed over sufficiently many independent realization of the channel matrix $\mathbf{H}$. See [23] for an analysis of the impact of imperfect channel-state information on the system performance.

[3] This is the scaling factor in the scaled nearest-neighbor decoding rule.

V. Extensions

Extensions of the analysis described above to the frequency-selective case and to the use of orthognal-frequency-division multiplexing (OFDM) and oversampling DACs are discussed in [26], [28], [16]. In the oversampling case, the diagonal approximation involves neglecting also temporal correlation, and it turns out to be accurate only when the oversampling ratio is small (e.g., less than four for $L = 4$). The rounding approximation does not suffer from this limitation.

The use of low-precision DACs in the massive MIMO downlink causes unwanted out-of-band (OOB) emissions, which may be incompatible with the spectral requirements imposed by regulatory bodies. An extension of the Bussgang's decomposition (3) to OFDM systems with nonideal analog filters is used in [16] to study such OOB emissions. There, it is shown that by an appropriate design of the DACs' low-pass filter and by employing simple digital pre-equalization techniques, one can significantly reduce OOB emissions, at the cost of a small decrease in the SINDR (7) and of a small increase in the peak-to-average power ratio of the transmitted signal.

References

[1] Ericsson AB, Huawei Technologies, NEC Corporation, Alcatel Lucent, and Nokia Siemens Networks, *Common public radio interface (CPRI); Interface Specification*, CPRI specification v7.0, Sep. 2015.

[2] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," Apr. 2014. [Online]. Available: http://arxiv.org/abs/1404.7736

[3] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: channel estimation and high-order modulations," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, to appear.

[4] J. Zhang, L. Dai, S. Sun, and Z. Wang, "On the spectral efficiency of massive MIMO systems with low-resolution ADCs," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 842 – 845, Dec. 2015.

[5] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 983 – 997, Apr. 2016.

[6] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.

[7] B. Li, N. Liang, and W. Zhang, "On transmission model for massive mimo under low-resolution output quantization," in *Proc. IEEE Veh. Technol. Conf. Spring (VTC-Spring)*, Sydney, Australia, Jun. 2017.

[8] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[9] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[10] H. Jedda, J. A. Nossek, and A. Mezghani, "Minimum BER precoding in 1-bit massive MIMO systems," in *IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, Rio de Janeiro, Brazil, Jul. 2016.

[11] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sep. 2017.

[12] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[13] O. Castañeda, S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "1-bit massive MU-MIMO precoding in VLSI," *IEEE Trans. Emerg. Sel. Topics Circuits Syst*, vol. 7, no. 4, pp. 508–522, Dec. 2017.

[14] A. Nedelcu, F. Steiner, M. Staudacher, G. Kramer, W. Zirwas, R. S. Ganesan, P. Baracca, and S. Wesemann, "Quantized precoding for multi-antenna downlink channels with MAGIQ," Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.08735

[15] M. Sarajlić, L. Liu, and O. Edfors, "When are low resolution ADCs energy efficient in massive MIMO?" *IEEE ACCESS*, 2017.

[16] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "On out-of-band emissions of quantized precoding in massive MU-MIMO-OFDM," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove CA, U.S.A., Dec. 2017.

[17] F. Maloberti, *Data converters*. Dordrecht, The Netherlands: Springer, Mar. 2007.

[18] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Elec., Cambridge, MA, Tech. Rep. 216, Mar. 1952.

[19] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IEEE Trans. Inf. Theory*, vol. 4, no. 2, pp. 69–72, Jun. 1958.

[20] H. E. Rowe, "Memoryless nonlinearities with Gaussian inputs: Elementary results," *Bell Labs Tech. J.*, vol. 61, no. 7, pp. 1519–1525, Sep. 1982.

[21] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in fading environment," *Int. J. Electron. Commun. (AEÜ)*, vol. 47, no. 4, pp. 228–239, 1993.

[22] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[23] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove CA, Nov. 2016.

[24] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[25] J. H. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proc. IEEE*, vol. 54, no. 1, pp. 2–19, Jan. 1966.

[26] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," Sep. 2017. [Online]. Available: https://arxiv.org/abs/1709.04846

[27] A. B. Sripad and D. L. Snyder, "a necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 5, pp. 442–448, Oct. 1977.

[28] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Massive MU-MIMO-OFDM downlink with one-bit DACs and linear precoding," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Singapore, Dec. 2017.

On the Information Dimension Rate of Multivariate Gaussian Processes

Bernhard C. Geiger*, Tobias Koch†‡
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
†Signal Theory and Communications Department, Universidad Carlos III de Madrid, 28911, Leganés, Spain
‡Gregorio Marañón Health Research Institute, 28007, Madrid, Spain.
Emails: geiger@ieee.org, koch@tsc.uc3m.es

Abstract—**The authors have recently defined the Rényi information dimension rate $d(\{X_t\})$ of a stationary stochastic process $\{X_t,\, t \in \mathbb{Z}\}$ as the entropy rate of the uniformly-quantized process divided by minus the logarithm of the quantizer step size $1/m$ in the limit as $m \to \infty$ (B. Geiger and T. Koch, "On the information dimension rate of stochastic processes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, June 2017). For Gaussian processes with a given spectral distribution function F_X, they showed that the information dimension rate is given by the Lebesgue measure of the set of harmonics where the derivative of F_X is positive. This paper extends this result to multivariate Gaussian processes with a given matrix-valued spectral distribution function $F_{\mathbf{X}}$. It is demonstrated that the information dimension rate equals the average rank of the derivative of $F_{\mathbf{X}}$. As side results, it is shown that the scale and translation invariance of information dimension carries over from random variables to stochastic processes.**

I. Introduction

In 1959, Rényi [1] proposed the information dimension and the d-dimensional entropy to measure the information content of general random variables (RVs). In recent years, it was shown that the information dimension is of relevance in various areas of information theory, including rate-distortion theory, almost lossless analog compression, or the analysis of interference channels. For example, Kawabata and Dembo [2] showed that the information dimension of a RV is equal to its rate-distortion dimension, defined as twice the rate-distortion function $R(D)$ divided by $-\log(D)$ in the limit as $D \downarrow 0$. Koch [3] demonstrated that the rate-distortion function of a source with infinite information dimension is infinite, and that for any source with finite information dimension and finite differential entropy the Shannon lower bound on the rate-distortion function is asymptotically tight. Wu and Verdú [4] analyzed both linear encoding and Lipschitz decoding of discrete-time, independent and identically distributed (i.i.d.), stochastic processes and showed that the information dimension plays a fundamental role in achievability and converse results. Wu *et al.* [5] showed that the degrees of freedom of the K-user Gaussian interference channel can be characterized through the sum of information dimensions. Stotz and Bölcskei [6] later generalized this result to vector interference channels.

In [7], [8], we proposed the information dimension rate as a generalization of information dimension from RVs to univariate (real-valued) stochastic processes. Specifically, consider the stationary process $\{X_t, t \in \mathbb{Z}\}$, and let $\{[X_t]_m,\, t \in \mathbb{Z}\}$ be the process obtained by uniformly quantizing $\{X_t\}$ with step size $1/m$. We defined the information dimension rate $d(\{X_t\})$ of $\{X_t\}$ as the entropy rate of $\{[X_t]_m\}$ divided by $\log m$ in the limit as $m \to \infty$ [8, Def. 2]. We then showed that, for any stochastic process, $d(\{X_t\})$ coincides with the rate-distortion dimension of $\{X_t\}$ [8, Th. 5]. We further showed that for stationary Gaussian processes with spectral distribution function F_X, the information dimension rate $d(\{X_t\})$ equals the Lebesgue measure of the set of harmonics on $[-1/2, 1/2]$ where the derivative of F_X is positive [8, Th. 7]. This implies an intuitively appealing connection between the information dimension rate of a stochastic process and its bandwidth.

In this work, we generalize our definition of $d(\{X_t\})$ to multivariate processes. Consider the L-variate (real-valued) stationary process $\{\mathbf{X}_t\}$, and let $\{[\mathbf{X}_t]_m\}$ be the process obtained by quantizing every component process of $\{\mathbf{X}_t\}$ uniformly with step size $1/m$. As in the univariate case, the information dimension rate $d(\{\mathbf{X}_t\})$ of $\{\mathbf{X}_t\}$ is defined as the entropy rate of $\{[\mathbf{X}_t]_m\}$ divided by $\log m$ in the limit as $m \to \infty$. Our main result is an evaluation of $d(\{\mathbf{X}_t\})$ for L-variate Gaussian processes with spectral distribution matrix $F_{\mathbf{X}}$. We demonstrate that for such processes $d(\{\mathbf{X}_t\})$ equals the Lebesgue integral of the rank of the derivative of $F_{\mathbf{X}}$. As a corollary, we show that the information dimension rate of univariate complex-valued Gaussian processes is maximized if the process is proper, in which case it is equal to twice the Lebesue measure of the set of harmonics where the derivative of its spectral distribution function F_X is positive.

As side results, we show that $d(\{\mathbf{X}_t\})$ is scale and translation invariant. These properties are known for the information dimension of RVs (cf. [9, Lemma 3]), but they do not directly carry over to our definition of $d(\{\mathbf{X}_t\})$, which is why we state them explicitly in this paper.

Due to space limitations, some of the proofs are only sketched or omitted altogether. The full proofs appear in [10].

The work of Bernhard C. Geiger has been funded by the Erwin Schrödinger Fellowship J 3765 of the Austrian Science Fund. The work of Tobias Koch has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 714161), from the Spanish Ministerio de Economía y Competitividad under Grants TEC2013-41718-R, RYC-2014-16332 and TEC2016-78434-C3-3-R (AEI/FEDER, EU), and from the Comunidad de Madrid under Grant S2103/ICE-2845.

II. Notation and Preliminaries

We denote by $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}$ the set of real numbers, the set of complex numbers, and the set of integers, respectively. We use a calligraphic font, such as $\mathcal{F}$, to denote other sets, and we denote complements as $\mathcal{F}^{\mathsf{c}}$.

We denote RVs by upper case letters, e.g., X. For a finite or countably infinite collection of RVs we abbreviate $X_\ell^k \triangleq (X_\ell, \dots, X_{k-1}, X_k)$, $X_\ell^\infty \triangleq (X_\ell, X_{\ell+1}, \dots)$, and $X_{-\infty}^k \triangleq (\dots, X_{k-1}, X_k)$. Univariate discete-time stochastic processes are denoted as $\{X_t,\, t \in \mathbb{Z}\}$ or, in short, as $\{X_t\}$. For L-variate stochastic processes we use the same notation but with X_t replaced by $\mathbf{X}_t \triangleq (X_{1,t}, \dots, X_{L,t})$. We call $\{X_{i,t},\, t \in \mathbb{Z}\}$ a *component process*.

We define the quantization of X with precision m as

$$[X]_m \triangleq \frac{\lfloor mX \rfloor}{m} \tag{1}$$

where $\lfloor a \rfloor$ is the largest integer less than or equal to a. Likewise, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. We denote by $[X_\ell^k]_m = ([X_\ell]_m, \dots, [X_k]_m)$ the component-wise quantization of X_ℓ^k (and similarly for other collections of RVs or random vectors). Likewise, for complex RVs Z with real part R and imaginary part I, the quantization $[Z]_m$ is equal to $[R]_m + \imath [I]_m$ where $\imath \triangleq \sqrt{-1}$.

Let $H(\cdot)$, $h(\cdot)$, and $D(\cdot \| \cdot)$ denote entropy, differential entropy, and relative entropy, respectively, and let $I(\cdot;\cdot)$ denote the mutual information [11]. We take logarithms to base $e \approx 2.718$, so mutual informations and entropies have dimension *nats*. The entropy rate of a discrete-valued, stationary, L-variate stochastic process $\{\mathbf{X}_t\}$ is [11, Th. 4.2.1]

$$H'(\{\mathbf{X}_t\}) \triangleq \lim_{k\to\infty} \frac{H(\mathbf{X}_1^k)}{k}. \tag{2}$$

Rényi defined the information dimension of a collection of RVs X_ℓ^k as [1]

$$d(X_\ell^k) \triangleq \lim_{m\to\infty} \frac{H([X_\ell^k]_m)}{\log m} \tag{3}$$

provided the limit exists. If the limit does not exist, one can define the upper and lower information dimension $\overline{d}(X_\ell^k)$ and $\underline{d}(X_\ell^k)$ by replacing the limit with the limit superior and limit inferior, respectively. If a result holds for both the limit superior and the limit inferior but it is unclear whether the limit exists, then we shall write $\overline{\underline{d}}(X_\ell^k)$. We shall follow this notation throughout this document: an overline $\overline{(\cdot)}$ indicates that the quantity in the brackets has been computed using the limit superior over m, an underline $\underline{(\cdot)}$ indicates that it has been computed using the limit inferior, both an overline and an underline $\overline{\underline{(\cdot)}}$ indicates that a result holds irrespective of whether the limit superior or limit inferior over m is taken.

If $H([X_\ell^k]_1) < \infty$, then [1, Eq. 7], [4, Prop. 1]

$$0 \le \underline{d}(X_\ell^k) \le \overline{d}(X_\ell^k) \le k - \ell + 1. \tag{4}$$

If $H([X_\ell^k]_1) = \infty$, then $\overline{\underline{d}}(X_\ell^k) = \infty$. As shown in [9, Lemma 3], information dimension is invariant under scaling and translation, i.e., $\overline{\underline{d}}(a \cdot X_\ell^k) = \overline{\underline{d}}(X_\ell^k)$ and $\overline{\underline{d}}(X_\ell^k + c) = \overline{\underline{d}}(X_\ell^k)$ for every $a \neq 0$ and $c \in \mathbb{R}^{k-\ell+1}$.

III. Information Dimension of Univariate Processes

In [7], [8], we generalized (3) by defining the information dimension rate of a univariate stationary process $\{X_t\}$ as

$$d(\{X_t\}) \triangleq \lim_{m\to\infty} \frac{H'(\{[X_t]_m\})}{\log m} = \lim_{m\to\infty} \lim_{k\to\infty} \frac{H([X_1^k]_m)}{k \log m} \tag{5}$$

provided the limit exists. (The limit over k exists by stationarity.)

If $H([X_1]_1) < \infty$, then [8, Lemma 4]

$$0 \le \underline{d}(\{X_t\}) \le \overline{d}(\{X_t\}) \le 1. \tag{6}$$

If $H([X_1]_1) = \infty$, then $\overline{\underline{d}}(\{X_t\}) = \infty$. Moreover, the information dimension rate of the process cannot exceed the information dimension of the marginal RV, i.e.,

$$\overline{\underline{d}}(\{X_t\}) \le \overline{\underline{d}}(X_1). \tag{7}$$

Kawabata and Dembo [2, Lemma 3.2] showed that the information dimension of a RV equals its rate-distortion dimension. By emulating the proof of [2, Lemma 3.2], we generalized this result to stationary processes by demonstrating that the information dimension rate is equal to the rate-distortion dimension. Specifically, let $R(X_1^k, D)$ denote the rate-distortion function of the k-dimensional source X_1^k, i.e.,

$$R(X_1^k, D) \triangleq \inf_{\mathsf{E}[\|\hat{X}_1^k - X_1^k\|^2] \le D} I(X_1^k; \hat{X}_1^k) \tag{8}$$

where the infimum is over all conditional distributions of $\hat{X}_1^k$ given X_1^k such that $\mathsf{E}[\|\hat{X}_1^k - X_1^k\|^2] \le D$ (where $\|\cdot\|$ denotes the Euclidean norm). The rate-distortion dimension of the stationary process $\{X_t\}$ is defined as

$$\mathsf{dim}_R(\{X_t\}) \triangleq 2 \lim_{D \downarrow 0} \lim_{k\to\infty} \frac{R(X_1^k, kD)}{-k \log D} \tag{9}$$

provided the limit as $D \downarrow 0$ exists. By stationarity, the limit over k always exists [12, Th. 9.8.1]. We showed that [8, Th. 5]

$$\overline{\underline{\mathsf{dim}}}_R(\{X_t\}) = \overline{\underline{d}}(\{X_t\}). \tag{10}$$

This result directly generalizes to non-stationary process (possibly with the limit over k replaced by the limit superior or limit inferior).

IV. Information Dimension of Multivariate Processes

In this section, we generalize the definition of the information dimension rate (5) to multivariate (real-valued) processes and study its properties.

Definition 1 (Information Dimension Rate): The information dimension rate of the stationary, L-variate process $\{\mathbf{X}_t\}$ is

$$\begin{aligned} d(\{\mathbf{X}_t\}) &\triangleq \lim_{m\to\infty} \frac{H'(\{[\mathbf{X}_t]_m\})}{\log m} \\ &= \lim_{m\to\infty} \lim_{k\to\infty} \frac{H([X_{1,1}^k]_m, \dots, [X_{L,1}^k]_m)}{k \log m} \end{aligned} \tag{11}$$

provided the limit over m exists.

We next summarize some basic properties of the information dimension rate.

Lemma 1 (Finiteness and Bounds): Let $\{\mathbf{X}_t\}$ be a stationary, L-variate process. If $H([\mathbf{X}_1]_1) < \infty$, then

$$0 \leq \overline{\underline{d}}(\{\mathbf{X}_t\}) \leq \overline{\underline{d}}(\mathbf{X}_1) \leq L. \tag{12}$$

If $H([\mathbf{X}_1]_1) = \infty$, then $\overline{\underline{d}}(\{\mathbf{X}_t\}) = \infty$.

Proof: Suppose first that $H([\mathbf{X}_1]_1) < \infty$. Then, the rightmost inequality in (12) follows from (4). The leftmost inequality follows from the nonnegativity of entropy. Finally, the center inequality follows since conditioning reduces entropy, hence $H'(\{[\mathbf{X}_t]_m\}) \leq H([\mathbf{X}_1]_m)$.

Now suppose that $H([\mathbf{X}_1]_1) = \infty$. By stationarity and since $[\mathbf{X}_1]_1$ is a function of $[\mathbf{X}_1^k]_m$ for every m and every k, we have

$$H([\mathbf{X}_1]_1) \leq H([\mathbf{X}_1^k]_m). \tag{13}$$

This implies that $H'(\{[\mathbf{X}_t]_m\}) = \infty$ and the claim $\overline{\underline{d}}(\{\mathbf{X}_t\}) = \infty$ follows from Definition 1. ■

It was shown in [9, Lemma 3] that information dimension is invariant under scaling and translation. The same properties hold for the information dimension rate.

Lemma 2 (Scale Invariance): Let $\{\mathbf{X}_t\}$ be a stationary, L-variate process and let $a_i > 0$, $i = 1, \ldots, L$. Further let $Y_{i,t} \triangleq a_i X_{i,t}$, $i = 1, \ldots, L$, $t \in \mathbb{Z}$. Then, $\overline{\underline{d}}(\{\mathbf{Y}_t\}) = \overline{\underline{d}}(\{\mathbf{X}_t\})$.

Proof: The proof is based on [4, Lemma 16] and appears in [10]. For brevity, let us focus on the case $L = 2$. The case $L > 2$ follows analogously. For $L = 2$, we have

$$\begin{aligned}
&H([a_1 X_{1,1}^k]_m, [a_2 X_{2,1}^k]_m) \\
&\quad \leq H([X_{1,1}^k]_m, [X_{2,1}^k]_m) + H([a_1 X_{1,1}^k]_m | [X_{1,1}^k]_m) \\
&\qquad + H([a_2 X_{2,1}^k]_m | [X_{2,1}^k]_m) \\
&\quad \leq H([X_{1,1}^k]_m, [X_{2,1}^k]_m) \\
&\qquad + k \log(\lceil a_1 \rceil + 1) + k \log(\lceil a_2 \rceil + 1)
\end{aligned} \tag{14}$$

where the second step follows because, given $[X_{i,1}^k]_m$, $[a_i X_{i,1}^k]_m$ can have at most $\lceil a_i \rceil + 1$ possible values. By following the same steps with a_i replaced by $1/a_i$, we obtain the reverse inequality

$$\begin{aligned}
H([a_1 X_{1,1}^k]_m, [a_2 X_{2,1}^k]_m) &\geq H([X_{1,1}^k]_m, [X_{2,1}^k]_m) \\
&\quad - k \log(\lceil 1/a_1 \rceil + 1) - k \log(\lceil 1/a_2 \rceil + 1).
\end{aligned} \tag{15}$$

The lemma then follows by dividing (14) and (15) by $k \log m$ and by letting k and m tend to infinity. ■

Lemma 3 (Translation Invariance): Let $\{\mathbf{X}_t\}$ be a stationary, L-variate process and let $\{\mathbf{c}_t\}$, $t \in \mathbb{Z}$ be a sequence of L-dimensional vectors. Then, $\overline{\underline{d}}(\{\mathbf{X}_t + \mathbf{c}_t\}) = \overline{\underline{d}}(\{\mathbf{X}_t\})$.

Proof: The lemma follows from [9, Lemma 30], which states that

$$|H(U_1^{kL}) - H(V_1^{kL})| \leq \sum_{i=1}^{kL} \log(1 + A_i + B_i) \tag{16}$$

for any collection of integer-valued RVs U_1^{kL} and V_1^{kL} satisfying almost surely $-B_i \leq U_i - V_i \leq A_i$, $i = 1, \ldots, kL$. Applying this result with $U_{\ell L+j} = \lfloor m X_{\ell,j} + m c_{\ell,j} \rfloor$ and $V_{\ell L+j} = \lfloor m X_{\ell,j} \rfloor + \lfloor m c_{\ell,j} \rfloor$ gives the desired result. Indeed, we have that $-1 \leq U_{\ell L+j} - V_{\ell L+j} \leq 2$, so (16) yields

$$\left| H([\mathbf{X}_1^k]_m) - H([\mathbf{X}_1^k + \mathbf{c}_1^k]_m) \right| \leq kL \log(4). \tag{17}$$

We thus obtain $|d(\{\mathbf{X}_t\}) - d(\{\mathbf{X}_t + \mathbf{c}_t\})| = 0$ by dividing (17) by $k \log m$ and by letting k and m tend to infinity. ■

We finally observe that the information dimension rate of a stationary stochastic process equals its rate-distortion dimension. This generalizes [8, Th. 5] to multivariate processes.

Theorem 1: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate process. Then,

$$\overline{\underline{d}}(\{\mathbf{X}_t\}) = \overline{\underline{\mathsf{dim}}}_R\{\mathbf{X}_t\} \tag{18}$$

where $\mathsf{dim}_R\{\mathbf{X}_t\}$ is defined as in (9) but with $\{X_t\}$ replaced by $\{\mathbf{X}_t\}$.

Proof: The proof is analog to that of [2, Lemma 3.2] and [8, Th. 5] and is therefore omitted. ■

V. Information Dimension of Gaussian Processes

Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued Gaussian process with mean vector $\boldsymbol{\mu}$ and (matrix-valued) spectral distribution function (SDF) $\theta \mapsto F_{\mathbf{X}}(\theta)$. Thus, $F_{\mathbf{X}}$ is bounded, non-decreasing, and right-continuous on $[-1/2, 1/2]$, and it satisfies [13, (7.3), p. 141]

$$K_{\mathbf{X}}(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta} \mathrm{d}F_{\mathbf{X}}(\theta), \quad \tau \in \mathbb{Z} \tag{19}$$

where $K_{\mathbf{X}}(\tau) \triangleq \mathsf{E}\left[(\mathbf{X}_{t+\tau} - \boldsymbol{\mu})(\mathbf{X}_t - \boldsymbol{\mu})^{\mathsf{T}}\right]$ denotes the autocovariance function and $(\cdot)^{\mathsf{T}}$ denotes the transpose. It can be shown that $\theta \mapsto F_{\mathbf{X}}(\theta)$ has a derivative almost everywhere, which has positive semi-definite, Hermitian values [13, (7.4), p. 141]. We shall denote the derivative of $F_{\mathbf{X}}$ by $F'_{\mathbf{X}}$.

For univariate stationary Gaussian processes with SDF F_X, we have shown that the information dimension rate is equal to the Lebesgue measure of the set of harmonics on $[-1/2, 1/2]$ where the derivative of F_X is is positive [8, Th. 7], i.e.,

$$d(\{X_t\}) = \lambda(\{\theta\colon F'_X(\theta) > 0\}) \tag{20}$$

where $\lambda(\cdot)$ denotes the Lebesgue measure on $[-1/2, 1/2]$. This result can be directly generalized to the multivariate case where the component processes are independent. Indeed, suppose that $\{\mathbf{X}_t\}$ is a collection of L independent Gaussian processes $\{X_{i,t},\, t \in \mathbb{Z}\}$ with SDFs F_{X_i}. This corresponds to the case where the (matrix-valued) SDF is a diagonal matrix with the SDFs of the individual processes on the main diagonal. For independent processes, the joint entropy rate can be written as the sum of the entropy rates of the component processes. It follows that

$$d(\{\mathbf{X}_t\}) = \sum_{i=1}^{L} d(\{X_{i,t}\}) = \sum_{i=1}^{L} \lambda(\{\theta\colon F'_{X_i}(\theta) > 0\}). \tag{21}$$

The expression on the right-hand side (RHS) of (21) can alternatively be written as

$$\int_{-1/2}^{1/2} \sum_{i=1}^{L} \mathbf{1}\{F'_{X_i}(\theta) > 0\} \mathrm{d}\theta = \int_{-1/2}^{1/2} \mathrm{rank}(F'_{\mathbf{X}}(\theta)) \mathrm{d}\theta \tag{22}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Observe that it is immaterial at which frequencies the component processes contain signal power. For example, the information dimension rate of two independent Gaussian processes with bandwidth $1/4$ equals 1 regardless of where the derivatives of their SDFs have their support. The following theorem shows that this result continuous to hold for general L-variate Gaussian processes.

Theorem 2: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate Gaussian process with mean vector $\boldsymbol{\mu}$ and SDF $F_{\mathbf{X}}$. Then,

$$d(\{\mathbf{X}_t\}) = \int_{-1/2}^{1/2} \operatorname{rank}(F'_{\mathbf{X}}(\theta))\mathrm{d}\theta. \tag{23}$$

Proof: Due to space limitations, we only provide a proof outline. The full proof can be found in [10].

We first note that we can assume, without loss of optimality, that $\{\mathbf{X}_t\}$ has zero mean and that every component process of $\{\mathbf{X}_t\}$ has unit variance. Indeed, by Lemma 3, the information dimension rate of $\{\mathbf{X}_t\}$ is translation invariant, so we can subtract the mean without affecting the information dimension rate. Likewise, by Lemma 2, the information dimension rate of $\{\mathbf{X}_t\}$ is scale invariant, so any component process with positive variance can be normalized to a unit-variance process without affecting the information dimension rate. Furthermore, zero-variance component processes can be omitted without affecting neither the left-hand side (LHS) nor the RHS of (23).

We next write the entropy of $[\mathbf{X}_1^k]_m$ as

$$H([\mathbf{X}_1^k]_m) = h(\mathbf{W}_1^k) + kL\log m \tag{24}$$

where $\mathbf{W}_t \triangleq [\mathbf{X}_t]_m + \mathbf{U}_t$, $t \in \mathbb{Z}$ and $\{\mathbf{U}_t\}$ is a sequence of i.i.d. random vectors that are uniformly distributed on the L-dimensional hypercube $[0, 1/m)^L$. Denoting by $(\mathbf{W}_1^k)_G$ a Gaussian vector with the same mean and covariance matrix as $\mathbf{W}_1^k$, and denoting by $f_{\mathbf{W}_1^k}$ and $g_{\mathbf{W}_1^k}$ the probability density functions of $\mathbf{W}_1^k$ and $(\mathbf{W}_1^k)_G$, respectively, this can be expressed as

$$H([\mathbf{X}_1^k]_m) = h\big((\mathbf{W}_1^k)_G\big) + D(f_{\mathbf{W}_1^k}\|g_{\mathbf{W}_1^k}) + kL\log m. \tag{25}$$

The entropy rate of a stationary, multivariate, Gaussian process is given by [13, Th. 7.10]

$$\lim_{k\to\infty} \frac{h((\mathbf{W}_1^k)_G)}{k} = \frac{1}{2}\int_{-1/2}^{1/2} \log\big(2\pi e \det F'_{\mathbf{W}}(\theta)\big)\mathrm{d}\theta. \tag{26}$$

Furthermore, the relative entropy $D(f_{\mathbf{W}_1^k}\|g_{\mathbf{W}_1^k})$ is bounded by [10, Lemma 6]

$$\frac{D(f_{\mathbf{W}_1^k}\|g_{\mathbf{W}_1^k})}{k} \leq L\left(\frac{\log\big(2\pi(1+\frac{1}{12})\big)}{2} + \frac{75}{2} + \frac{24}{\pi}\right). \tag{27}$$

Thus, dividing (25) by $k\log m$, and letting first k and then m tend to infinity yields

$$d(\{\mathbf{X}_t\}) = L + \lim_{m\to\infty}\int_{-1/2}^{1/2} \frac{\log\det F'_{\mathbf{W}}(\theta)}{2\log m}\mathrm{d}\theta. \tag{28}$$

It remains to show that the RHS of (28) is equal to the RHS of (23). To this end, we use that for zero-mean processes $\{\mathbf{X}_t\}$ with unit-variance component processes the SDF of $\{[\mathbf{X}_t]_m\}$ can be expressed as [10, Lemma 4]

$$F_{[\mathbf{X}]_m}(\theta) = (2a-1)F_{\mathbf{X}}(\theta) + F_{\mathbf{N}}(\theta) \tag{29}$$

where $a \triangleq \mathsf{E}\left[X_{1,1}[X_{1,1}]_m\right]$ and the diagonal elements of $F_{\mathbf{N}}(\theta)$ satisfy

$$\int_{-1/2}^{1/2} \mathrm{d}F_{N_i}(\theta) \leq \frac{1}{m^2}. \tag{30}$$

We can thus express the derivative of the SDF of $\{\mathbf{W}_t\}$ as

$$F'_{\mathbf{W}}(\theta) = (2a-1)F'_{\mathbf{X}}(\theta) + F'_{\mathbf{N}}(\theta) + \frac{1}{12m^2}I_L \tag{31}$$

where I_L denotes the $L\times L$ identity matrix. By performing an analysis similar to that in [8, App. C-A], one can show that

$$\lim_{m\to\infty}\int_{-1/2}^{1/2} \frac{\log\det F'_{\mathbf{W}}(\theta)}{2\log m}\mathrm{d}\theta = -\sum_{i=1}^{L}\lambda(\{\theta\colon \mu_i(\theta)=0\}) \tag{32}$$

where $\mu_i(\theta)$ denotes the i-th eigenvalue of $F'_{\mathbf{X}}(\theta)$. (For the details, see [10, App. A].). Combining (32) with (28) gives

$$\begin{aligned} d(\{\mathbf{X}_t\}) &= \sum_{i=1}^{L}\big[1-\lambda(\{\theta\colon \mu_i(\theta)=0\})\big] \\ &= \sum_{i=1}^{L}\lambda(\{\theta\colon \mu_i(\theta)>0\}) \end{aligned} \tag{33}$$

which as in (21) and (22) can be shown to be equal to the RHS of (23). ■

VI. Information Dimension of Complex Gaussian Processes

Theorem 2 allows us to study the information dimension of stationary, univariate, complex-valued Gaussian processes by treating them as bivariate, real-valued processes. Let $\{Z_t\}$ be a stationary, univariate, complex-valued, Gaussian process with mean μ and SDF F_Z, i.e.,

$$K_Z(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta}\mathrm{d}F_Z(\theta), \quad \tau\in\mathbb{Z} \tag{34}$$

where $K_Z(\tau) \triangleq \mathsf{E}\left[(Z_{t+\tau}-\mu)(Z_t-\mu)^*\right]$ is the autocovariance function, and $(\cdot)^*$ denotes complex conjugation.

Alternatively, $\{Z_t\}$ can be expressed in terms of its real and imaginary part. Indeed, let $Z_t = R_t + \imath I_t$, $t\in\mathbb{Z}$. The stationary, bivariate, real-valued process $\{(R_t, I_t),\, t\in\mathbb{Z}\}$ is jointly Gaussian and has SDF

$$F_{(R,I)}(\theta) = \begin{pmatrix} F_R(\theta) & F_{RI}(\theta) \\ F_{IR}(\theta) & F_I(\theta) \end{pmatrix}, \quad -\frac{1}{2}\leq\theta\leq\frac{1}{2} \tag{35}$$

where F_R and F_I are the SDFs of $\{R_t\}$ and $\{I_t\}$, respectively, and F_{RI} and F_{IR} are the cross SDFs between $\{R_t\}$ and $\{I_t\}$. The derivatives of F_Z and $F_{(R,I)}$ are connected as follows:

$$\begin{aligned} F'_Z(\theta) &= F'_R(\theta) + F'_I(\theta) + \imath\big(F'_{IR}(\theta) - F'_{RI}(\theta)\big) \\ &= F'_R(\theta) + F'_I(\theta) + 2\mathfrak{Im}\big(F'_{RI}(\theta)\big) \end{aligned} \tag{36}$$

where the last equality follows because $F'_{(R,I)}$ is Hermitian. Here we use $\mathfrak{Im}(\cdot)$ to denote the imaginary part. It can be further shown that $\theta \mapsto F'_R(\theta)$ and $\theta \mapsto F'_I(\theta)$ are real-valued and symmetric, and that $\theta \mapsto \mathfrak{Im}\big(F'_{RI}(\theta)\big)$ is anti-symmetric.

A stationary, complex-valued process $\{Z_t\}$ is said to be *proper* if its mean μ and its pseudo-autocovariance function

$$\overline{K_Z}(\tau) \triangleq \mathsf{E}\left[(Z_{t+\tau}-\mu)(Z_t-\mu)\right], \quad \tau \in \mathbb{Z}$$

are both zero [14, Def. 17.5.4]. Since, by Lemma 3, the information dimension rate is independent of μ, we shall slightly abuse notation and say that a stationary, complex-valued process is proper if its pseudo-autocovariance function is identically zero, irrespective of its mean. Properness implies that, for all θ, $F_R(\theta) = F_I(\theta)$ and $F_{RI}(\theta) = -F_{IR}(\theta)$. Since $\theta \mapsto F'_{(R,I)}(\theta)$ is Hermitian, this implies that for a proper process the function $\theta \mapsto F'_{RI}(\theta)$ is purely imaginary.

The following corollary to Theorem 2 shows that proper Gaussian processes maximize information dimension. This parallels the result that proper Gaussian vectors maximize differential entropy [15, Th. 2].

Corollary 1: Let $\{Z_t\}$ be a stationary, complex-valued Gaussian process with mean μ and SDF F_Z. Then

$$d(\{Z_t\}) \leq 2 \cdot \lambda(\{\theta\colon F'_Z(\theta) > 0\}) \tag{37}$$

with equality if $\{Z_t\}$ is proper.

Proof: We know from Theorem 2 that

$$d(\{Z_t\}) = \int_{-1/2}^{1/2} \operatorname{rank}(F'_{(R,I)}(\theta))\mathrm{d}\theta. \tag{38}$$

For a given θ, the eigenvalues of $F'_{(R,I)}(\theta)$ are given by

$$\frac{F'_R(\theta)+F'_I(\theta)}{2} \pm \sqrt{\frac{(F'_R(\theta)-F'_I(\theta))^2}{4} + |F'_{RI}(\theta)|^2}. \tag{39}$$

Since $F'_{(R,I)}(\theta)$ is positive semi-definite, these eigenvalues are nonnegative and

$$F'_R(\theta)F'_I(\theta) \geq |F'_{RI}(\theta)|^2. \tag{40}$$

The larger of these eigenvalues, say $\mu_1(\theta)$, is zero on

$$\mathcal{F}_1 \triangleq \{\theta\colon F'_R(\theta) = F'_I(\theta) = 0\}. \tag{41}$$

The smaller eigenvalue, $\mu_2(\theta)$, is zero on

$$\mathcal{F}_2 \triangleq \big\{\theta\colon F'_R(\theta)F'_I(\theta) = |F'_{RI}(\theta)|^2\big\}. \tag{42}$$

Clearly, $\mathcal{F}_1 \subseteq \mathcal{F}_2$. By (38), we have that

$$\begin{aligned} d(\{Z_t\}) &= \lambda(\{\theta\colon \mu_1(\theta) > 0\}) + \lambda(\{\theta\colon \mu_2(\theta) > 0\}) \\ &= 1 - \lambda(\mathcal{F}_1) + 1 - \lambda(\mathcal{F}_1) - \lambda(\mathcal{F}_1^{\mathsf{c}} \cap \mathcal{F}_2). \end{aligned} \tag{43}$$

We next note that, by (36) and (40), the derivative $F'_Z(\theta)$ is zero if either $F'_R(\theta) = F'_I(\theta) = 0$ or if $F'_R(\theta) + F'_I(\theta) > 0$ and $F'_R(\theta) + F'_I(\theta) = -2\mathfrak{Im}(F'_{RI}(\theta))$. Since $\theta \mapsto F'_R(\theta)$ and $\theta \mapsto F'_I(\theta)$ are symmetric and $\theta \mapsto \mathfrak{Im}(F'_{RI}(\theta))$ is anti-symmetric, it follows that for any $\theta \in \mathcal{F}_1^c$ satisfying $F'_R(\theta) + F'_I(\theta) = -2\mathfrak{Im}(F'_{RI}(\theta))$ we have that $F'_R(-\theta) + F'_I(-\theta) = 2\mathfrak{Im}(F'_{RI}(-\theta))$. Thus, defining

$$\mathcal{F}_3 \triangleq \big\{\theta\colon F'_R(\theta) + F'_I(\theta) = 2|\mathfrak{Im}(F'_{RI}(\theta))|\big\} \tag{44}$$

we can express the Lebesgue measure of the set of harmonics where $F'_Z(\theta) = 0$ as

$$\lambda(\{\theta\colon F'_Z(\theta) = 0\}) = \lambda(\mathcal{F}_1) + \frac{1}{2}\lambda(\mathcal{F}_1^{\mathsf{c}} \cap \mathcal{F}_3). \tag{45}$$

Combining (43) and (45), we obtain

$$\begin{aligned} d(\{Z_t\}) &= 2\lambda(\{\theta\colon F'_Z(\theta) > 0\}) \\ &\qquad + \lambda(\mathcal{F}_1^{\mathsf{c}} \cap \mathcal{F}_3) - \lambda(\mathcal{F}_1^{\mathsf{c}} \cap \mathcal{F}_2). \end{aligned} \tag{46}$$

Since the arithmetic mean is greater than or equal to the geometric mean, and with (40), we have that

$$\begin{aligned} (F'_R(\theta) + F'_I(\theta))^2 &\geq 4F'_R(\theta)F'_I(\theta) \\ &\geq 4|F'_{RI}(\theta)|^2 \geq 4\mathfrak{Im}(F'_{RI}(\theta))^2. \end{aligned} \tag{47}$$

Hence, $\mathcal{F}_3 \subseteq \mathcal{F}_2$ and the second line in (46) is less than or equal to zero. This proves (37).

If $\{Z_t\}$ is proper, then we have $F'_R(\theta) = F'_I(\theta)$ and $|F'_{RI}(\theta)| = |\mathfrak{Im}(F'_{RI}(\theta))|$. In this case, $F'_R(\theta)F'_I(\theta) = |F'_{RI}(\theta)|^2$ implies $F'_R(\theta) + F'_I(\theta) = 2|\mathfrak{Im}(F'_{RI}(\theta))|$, so $\mathcal{F}_2 \subseteq \mathcal{F}_3$. It follows that $\mathcal{F}_2 = \mathcal{F}_3$ and the second line in (46) is zero. Hence, (37) holds with equality. ■

Remark 1: There are also non-proper processes for which (37) holds with equality. For example, this is the case for any stationary Gaussian process for which real and imaginary parts are independent and F'_R and F'_I have matching support but are different otherwise.

References

[1] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 193–215, Mar. 1959.

[2] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1564–1572, Sep. 1994.

[3] T. Koch, "The Shannon lower bound is asymptotically tight," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6155–6161, Nov. 2016.

[4] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[5] Y. Wu, S. Shamai (Shitz), and S. Verdú, "Information dimension and the degrees of freedom of the interference channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 256–279, Jan. 2015.

[6] D. Stotz and H. Bölcskei, "Degrees of freedom in vector interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4172–4197, Jul. 2016.

[7] B. C. Geiger and T. Koch, "On the information dimension rate of stochastic processes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 888–892.

[8] ——, "On the information dimension of stochastic processes," `arXiv:1702.00645v2 [cs.IT]`, Feb. 2017.

[9] Y. Wu, "Shannon theory for compressed sensing," Ph.D. dissertation, Princeton University, 2011.

[10] B. C. Geiger and T. Koch, "On the information dimension of multivariate Gaussian processes," `arXiv:1712.07863v1 [cs.IT]`, Dec. 2017.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. Wiley Interscience, 1991.

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

[13] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes," *Acta Mathematica*, vol. 98, no. 1, pp. 111–150, 1957.

[14] A. Lapidoth, *A Foundation in Digital Communication*. Cambridge: Cambridge University Press, 2009.

[15] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

Quantizations Preserving Kullback-Leibler Divergence

Wasim Huleihel
MIT
wasimh@mit.edu

Matthew Brennan
MIT
brennanm@mit.edu

Guy Bresler
MIT
gbresler@mit.edu

Abstract—**We consider the problem of recovering a hidden community of size K from a graph where edges between members of the community have label drawn independently and identically distributed (i.i.d.) according to P and all other edges have labels drawn i.i.d. according to Q. The information limits for this problem were characterized by Hajek-Wu-Xu in 2016 in terms of the Kullback-Leibler (KL) divergence between P and Q. We complement their work by showing that for a broad class of distributions P and Q one may reduce to the case $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$. Specifically, given $X \sim P$ and $Y \sim Q$, we show that there exists some map (or, qunatizer) $\Phi : \mathbb{R} \to \{0,1\}$, preserving the KL-divergence, i.e., $d_{\mathrm{KL}}(p_\Phi||q_\Phi) \geq C \cdot D_{\mathrm{KL}}(P||Q)$, where p_Φ and q_Φ are the probability laws of the random variables $\Phi(X)$ and $\Phi(Y)$, respectively, and C is some universal constant.**

I. Introduction

Many networks of interest display community structure, i.e., their vertices (denoting the objects) are organized into groups, called communities, and edges capturing their pairwise dependencies. For example, in social network analysis, these groups can be seen as communities with higher edge dependencies than the rest of the network. Generally speaking, the main goal in community detection is to identify these communities. We consider the following probabilistic definition of the hidden community model [1].

Definition 1 (Hidden Community Model). *Let $\mathcal{C}^*$ be drawn uniformly at random from all subsets of $\{1, 2, \ldots, n\}$ of cardinality K. Given probability measures P and Q on a common measurable space, let A be an $n \times n$ symmetric matrix with zero diagonal where for all $1 \leq i < j \leq n$, A_{ij} are mutually independent, $A_{ij} \sim P$ if $i, j \in \mathcal{C}^*$, and $A_{ij} \sim Q$ otherwise.*

Observing A, the main task is to accurately (or approximately) recover the underlying community $\mathcal{C}^*$. The distributions P and Q as well as the community size K depend on n. It is then reasonable that, for a fixed network size n, as the community size K decreases, or the distributions P and Q get closer (in some sense), the recovery problem becomes harder.

The community detection problem was extensively studied in the literature (see, e.g., [1-14], and many references therein). The information theoretic limits for exact recovery have become increasingly well-understood in the literature. For example, in the Bernoulli case, it was shown [8] that if $K \cdot d_{\mathrm{KL}}(q||p) - c \cdot \log K \to \infty$ and $K \cdot d_{\mathrm{KL}}(q||p) \geq c \cdot \log n$, for some large constant $c > 0$, then exact recovery is achievable via the maximum likelihood estimator (MLE), otherwise, exact recovery is impossible for any algorithms. Similar results were proved in the Gaussian case [7].

II. Recent Results

Recently, [9] derived the information limits with sharp constants for a broad class of distributions P and Q. In a recent work [14], among other things, we complement [9] by showing that for the same class of distributions P and Q one may *reduce* to the case $P = \mathrm{Bernoulli}(p)$ and $Q = \mathrm{Bernoulli}(q)$. In other words, we show that there exists a map (or, quantizer) which takes as an input the matrix A in Definition 1, and outputs a (p, q)-binary matrix in a way which "preserves the information". Specifically, given $X \sim P$ and $Y \sim Q$, we show that there exists some map (or, qunatizer) $\Phi : \mathbb{R} \to \{0,1\}$, preserving the KL-divergence, i.e., $d_{\mathrm{KL}}(p_\Phi||q_\Phi) \geq C \cdot D_{\mathrm{KL}}(P||Q)$, where p_Φ and q_Φ are the probability laws of the random variables $\Phi(X)$ and $\Phi(Y)$, respectively, and C is some universal constant. This result, together with [8] provides an alternative and constructive proof of the achievability part in [9] for general P and Q.

References

[1] Y. Deshpande and A. Montanari, "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time," *Foundations of Computational Mathematics*, vol. 15, no. 4, pp. 1069–1128, Aug. 2015.

[2] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, vol. 3, no. 3, p. 9851012, Oct. 2009. [Online]. Available: arxiv.org/pdf/1510.09219

[3] C. Butucea and Y. I. Ingster, "Detection of a sparse submatrix of a high-dimensional noisy matrix," *Bernoulli*, vol. 19, no. 5B, pp. 2652–2688, 2013.

[4] C. Butucea, Y. I. Ingster, and I. Suslina, "Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix," *ESAIM: Probability and Statistics*, vol. 19, pp. 115–134, 2013.

[5] A. Montanari, "Finding one community in a sparse random graph," Feb. 2015. [Online]. Available: arxiv.org/pdf/1502.05680

[6] E. Arias-Castro and N. Verzelen, "Community detection in dense random networks," *Ann. Statist.*, vol. 42, no. 3, pp. 940–969, 2014.

[7] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh, "Minimax localization of structural information in large noisy matrices," *Advances in Neural Information Processing Systems*, 2011.

[8] J. Chen, Y. Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," in *ICML 2014*, 2014.

[9] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4729–4745, Aug 2017.

[10] ——, "Achieving exact cluster recovery threshold via semidefinite programming," Nov. 2014. [Online]. Available: arxiv.org/pdf/1412.6156

[11] ——, "Computational lower bounds for community detection on random graphs," *In Proceedings COLT 2015*, June 2015.

[12] ——, "Submatrix localization via message passing," Oct. 2015. [Online]. Available: arxiv.org/pdf/1510.09219

[13] T. T. Cai, T. Liang, and A. Rakhlin, "Computational and statistical boundaries for submatrix localization in a large noisy matrix," *Ann. Statist.*, vol. 45, no. 4, pp. 1403–1430, 2017.

[14] M. Brennan, G. Bresler, and W. Huleihel, "Universality in the computational and statistical limits of submatrix detection," *In preparation.*

Information-Distilling Quantizers

Bobak Nazer
BU
bobak@bu.edu

Or Ordentlich
HUJI
or.ordentlich@mail.huji.ac.il

Yury Polyanskiy
MIT
yp@mit.edu

Let X and Y be a pair of random variables with a given distribution P_{XY}. This work deals with the problem of quantizing Y into M values, under the objective of maximizing the mutual information between the quantizer's output and X. We denote the value of the mutual information attained by the optimal M-ary quantizer by

$$I(X;[Y]_M) \triangleq \sup_{\tilde{Y}\in[Y]_M} I(X;\tilde{Y}). \qquad (1)$$

where $[Y]_M$ is the set of all (deterministic) M-ary quantizations of Y,

$$[Y]_M \triangleq \{f(Y)\ :\ f:\mathcal{Y}\to[M]\}.$$

When X and Y are thought of as the input and output of a channel, respectively, the problem boils down to designing the M-level quantizer that maximizes the information rate, whereas (1) is the highest information rate attainable. It is therefore not surprising that this problem has received considerable attention [1]–[5]. For example, it is well known [6, Section 2.11] that when X is a BPSK input to an AWGN channel with output Y it holds that $I(X;[Y]_2)\geq 2I(X;Y)/\pi$ and this is achieved by taking $f(\cdot)$ to be the maximum a posteriori (MAP) estimator of X from Y.

Our focus is studying the fundamental properties of the function $I(X;[Y]_M)$, and in particular, identifying the joint distributions P_{XY} that are the most difficult to quantize, and characterizing the behavior of $I(X;[Y]_M)$ for these cases. Special attention is given to the symmetric binary case where $X\sim\text{Bernoulli}(1/2)$. In this setting, it may seem that the optimal binary quantizer should always retain a significant fraction of $I(X;Y)$, and that the MAP quantizer should be sufficient to this end. For large $I(X;Y)$, it is not difficult to see that this is indeed the case [7, Proposition 5]. However, if $Y\in\{0,1,?\}$ is the output of a binary erasure channel with input X, for large erasure probabilities the MAP quantizer may be arbitrarily inferior to the asymmetric quantizer $f(0)=0$, $f(1)=f(?)=1$ [7, Section III.c].

Furthermore, in certain cases, no binary quantizer can retain a significant fraction of $I(X;Y)$. Our main result is the following [7, Theorem 1]. Logarithms are taken w.r.t. base 2, with the exception of the $\ln$ function that is taken w.r.t. base e.

Theorem 1: If $X\sim\text{Bernoulli}(1/2)$ and $I(X;Y)=\beta>0$, we have for binary quantization

$$I(X;[Y]_2)\geq\frac{1}{3e}\frac{\beta}{1+\ln\left(\frac{1}{\beta}\right)}. \qquad (2)$$

Furthermore, for any $\eta\in(0,1)$ and any natural $M<\frac{12\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}{(1-\eta)^2}$

$$I(X;[Y]_M)\geq(M-1)\frac{\beta}{\max\{\log\left(\frac{1}{\beta}\right),1\}}\frac{\eta(1-\eta)^2}{12}. \qquad (3)$$

Finally, for any $0<\beta\leq 1$, there exist distributions P_{XY} with $X\sim\text{Bernoulli}(1/2)$ and $I(X;Y)=\beta$, for which

$$I(X;[Y]_M)\leq 2M\frac{\beta}{\ln\left(\frac{e\log(e)}{2\beta}\right)}, \qquad (4)$$

for every natural M.

Note that this is in stark contrast to the intuition from the binary AWGN channel. While for the former, two quantization levels suffice for retaining a $2/\pi$ fraction of $I(X;Y)$, Theorem 1 shows that there exist distributions for which at least $\Omega(\log(1/I(X;Y)))$ quantization levels are needed in order to retain a fixed fraction of $I(X;Y)$.

References

[1] R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar, "On the construction of polar codes," in *2011 IEEE International Symposium on Information Theory Proceedings*, July 2011, pp. 11–15.

[2] I. Tal, A. Sharov, and A. Vardy, "Constructing polar codes for non-binary alphabets and macs," in *2012 IEEE International Symposium on Information Theory Proceedings*, July 2012, pp. 2132–2136.

[3] B. M. Kurkoski and H. Yagi, "Quantization of binary-input discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4544–4552, Aug 2014.

[4] I. Tal, "On the construction of polar codes for channels with moderate input alphabet sizes," in *IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1297–1301.

[5] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," *arXiv preprint arXiv:1701.02119*, 2017.

[6] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*. Courier Corporation, 2013.

[7] B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 96–100.

Analog Source Coding and Robust Frames

Marina Haikin
Tel Aviv University
Email: mkokotov@gmail.com

Ram Zamir
Tel Aviv University
Email: zamir@eng.tau.ac.il

Matan Gavish
The Hebrew University
Email: gavish@cs.huji.ac.il

Abstract—**Analog coding is a low-complexity method to combat erasures, based on linear redundancy in the signal space domain. Previous work examined ”bandlimited discrete Fourier transform (DFT)” codes for Gaussian channels with erasures or impulses. We extend this concept to source coding with ”erasure side-information” at the encoder. Furthermore, we show that the performance of bandlimited DFT can be significantly improved using irregular spectrum, and more generally, using equiangular tight frames.**

Key words: Distortion side information, erasures, equiangular tight frames, Welch bound, random matrix theory.

I. Distortion side information at the encoder

Consider encoding a source X under a side-information dependent distortion measure $d(x, \hat{x}, s)$, where the side information S is statistically independent of the source X and is available at the encoder. It is shown in [5] that if an optimal conditional distribution $p(\hat{x}|x, s)$ satisfies $I(S; \hat{X}) = 0$, then the rate-distortion performance is the same as if S was available also at the decoder. Specifically, this condition holds for the case of an “erasure distortion measure” $d(x, \hat{x}, s) = s \cdot d(x, \hat{x})$, for $s \in \{0, 1\}$, where only source samples for which $S = 1$ are “important”.

II. Analog coding of a source with erasures

Analog coding decouples the tasks of protecting against erasures and noise [7]. For erasure correction, it creates an analog redundancy by means of band-limited discrete Fourier transform (DFT) interpolation, or more generally, by an over-complete expansion based on a frame. In [2] we examine the analog coding paradigm for the dual setup of a source with erasure side-information (SI) at the encoder [5]. The excess rate of analog coding above the rate-distortion function (RDF) is associated with the energy of the inverse of submatrices of the frame, where each submatrix corresponds to a possible erasure pattern. We show that by selecting the DFT frequencies from a *difference set*, or more generally, by using equiangular tight frames (ETF), we minimize the excess rate over all possible frames (although do not achieve the RDF); see Section III below.

III. Random subsets of deterministic frames

Suppose we draw a random subset of k rows from a frame with n rows (vectors) and m columns (dimensions), where k and m are proportional to n. Consider the distribution of singular values of the k-subset matrix. For a variety of important ETFs and tight non-ETFs, we observe in [3] that, for large n, the singular values can be precisely described by a known probability distribution: Wachter’s MANOVA (multivariate ANOVA) spectral distribution, a phenomenon that was previously known only for two types of random frames [1]. In terms of convergence to this limit, the k-subset matrix from all of these frames is shown to be empirically indistinguishable from the classical MANOVA (Jacobi) random matrix ensemble. Thus, empirically, the MANOVA ensemble offers a universal description of the spectra of randomly selected k subframes, even those taken from deterministic frames.

IV. Welch Bounds with Erasures

The Welch Bound [6] is a lower bound on the root mean square cross correlation between n unit-norm vectors $f_1, \ldots, f_n$ in the m dimensional space (R^m or C^m), for $n > m$. Letting $F = [f_1| \ldots |f_n]$ denote the m-by-n matrix (frame) composed of the n vectors, the Welch bound can be viewed as a lower bound on the second moment of F, namely on the trace of the squared Gram matrix $(F'F)^2$. In [4] we extend the Welch Bound to a random selection of a subset from F, as well as to higher order moments of F. The extended lower bound holds with equality if and only if F is an ETF. Thus, it provides an analytical support for the results in [2], and sheds light on the superiority of ETFs for a variety of applications, such as spread spectrum communications, compressed sensing and analog coding [2].

References

[1] B. Farrell. Limiting empirical singular value distribution of restrictions of discrete fourier transform matrices. *Journal of Fourier Analysis and Applications*, 17(4):733–753, 2011.

[2] M. Haikin and R. Zamir. Analog coding of a source with erasures. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2074–2078, 2016.

[3] M. Haikin, R. Zamir, and M. Gavish. Random subsets of structured deterministic frames have manova spectra. *Proceedings of the National Academy of Sciences*, pages E5024–E5033, 2017.

[4] M. Haikin, R. Zamir, and M. Gavish. Frame moments and Welch bounds with erasures. In *Information Theory (ISIT), 2018 IEEE International Symposium on*, submitted.

[5] E. Martinian, G. W. Wornell, and R. Zamir. Source coding with distortion side-information. *IEEE Trans. Information Theory*, 54:4638–4665, Oct. 2008.

[6] L. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

[7] J Wolf. Redundancy, the discrete fourier transform, and impulse noise cancellation. *IEEE Transactions on Communications*, 31(3):458–461, 1983.

Wiretap and Gelfand-Pinsker Channels Analogy and its Applications

Ziv Goldfeld
Ben Gurion University
gziv@post.bgu.ac.il

Haim H. Permuter
Ben Gurion University
haimp@bgu.ac.il

***Abstract*—A framework of analogy between wiretap channels (WTCs) and state-dependent point-to-point channels with non-causal encoder channel state information (referred to as Gelfand-Pinker channels (GPCs)) is proposed. A good (reliable and secure) sequence of wiretap codes is shown to induce a good (reliable) sequence of codes for a corresponding GPC. Consequently, the framework enables exploiting existing results for GPCs to produce converse proofs for their wiretap analogs. The fundamental limits of communication of two analogous wiretap and GP models are characterized by the same rate bounds; the optimization domains may differ. The analogy readily extends to multiuser broadcasting scenarios, encompassing broadcast channels (BCs) with deterministic components, degradation ordering between users, and BCs with cooperative receivers. Given a wiretap BC (WTBC) with two receivers and one eavesdropper, an analogous Gelfand-Pinsker BC (GPBC) is constructed by converting the eavesdropper's observation sequence to a state sequence with an appropriate product distribution, and non-causally revealing the states to the encoder. The transition matrix of the (state-dependent) GPBC is the appropriate conditional marginal of the WTBC's transition law, with the eavesdropper's output playing the role of the channel state. The analogy is exploited to characterize the secrecy-capacity regions of the SD-WTBC, which was an open problem until this work, based on the corresponding solution of the SD-GPBC.**

I. Introduction

Two fundamental, but seemingly unrelated, information-theoretic models are that of the wiretap channel (WTC) and the state-dependent point-to-point channel with non-causal encoder channel state information (CSI). The discrete and memoryless (DM) WTC (Fig. 1(a)) was introduced by Wyner in his celebrated 1975 paper [1] that initiated the study of physical layer security. Csiszár and Körner characterized the secrecy-capacity of the WTC as

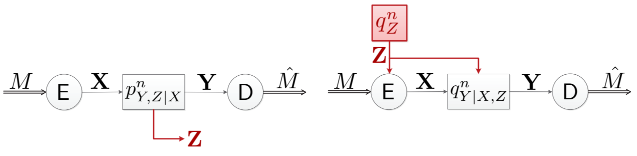
$$C_{\mathsf{WT}}(p_{Y,Z|X}) = \max_{p_{U,X}} \Big[I(U;Y) - I(U;Z) \Big], \qquad (1)$$

where $p_{Y,Z|X}$ is the WTC's transition matrix and the underlying distribution is $p_{U,X}p_{Y,Z|X}$. The state-dependent channel with non-causal encoder CSI is due to Gelfand and Pinsker (GP) [2], and is henceforth referred to as the GP channel (GPC). A single-letter capacity formula for any GPC $q_{Y|X,Z}$ with state distribution q_Z was derived in [2]:

$$C_{\mathsf{GP}}(q_Z, q_{Y|X,Z}) = \max_{q_{U,X|Z}} \Big[I(U;Y) - I(U;Z) \Big], \qquad (2)$$

where the joint distribution is $q_Z q_{U,X|Z} q_{Y|X,Z}$. An interesting question is whether the resemblance of (1) and (2) is coincidental or is there an inherent relation between these problems.

This paper shows that an inherent relation is indeed the case, by proposing a rigorous framework that links the WTC and the GPC, establishing these two problems as analogous to one another. Specifically, we prove that any good (reliable and secure) sequence of codes for the WTC induces a good (reliable) sequence of codes of the same rate for a corresponding GPC. This observation enables exploiting known outer bounds on the GPC capacity to outer bound the secrecy-capacity of an analogous WTC. While the solutions to the base cases from Fig. 1 have been known for decades, many multiuser extensions of these models remain open problems. Through the analogy we derive a converse proof for the semi-deterministic (SD) wiretap broadcast channel (WTBC), an open problem until this work, thus characterizing its secrecy-capacity region.



(a) (b)

Fig. 1: (a) The WTC with transition probability $p_{Y,Z|X}$, where X is the channel input and Y and Z are the channel outputs observed by the legitimate receiver and the eavesdropper, respectively; (b) The GPC with state distribution $Z \sim q_Z$, and channel transition probability $q_{Y|X,Z}$, where X is the input and Y is the output.

To this end we extend the wiretap-GP analogy to multiuser broadcasting scenarios. Given a WTBC $p_{Y_1,Y_2,Z|X}$ (Fig. 2(a)), with two legitimate receivers observing Y_1 and Y_2 and one eavesdropper that intercepts Z, an analogous GP broadcast channel (GPBC), shown in Fig. 2(b), is constructed as follows:

1) Converting the eavesdropper's observation sequence Z^n to an independently and identically distributed (i.i.d.) state sequence with some appropriate distribution;
2) Revealing the state sequence in a non-causal manner to the encoder;
3) Setting the state-dependent BC $p_{Y_1,Y_2|X,Z}$ (the conditional marginal of the WTBC's transition probability) with Z in the role of the state.

The aforementioned relation between good sequences of codes for analogous WTBCs and GPBCs remains valid, which allows capitalizing on known GPBC capacity results to derive converse proofs for their analogous WTBC.

The GPBC has been widely studied in the literature and the capacity region is known for various cases [3]–[5]. Of particular interest is the capacity derivation of the SD-GPBC from [4]. WTBC also received considerable attention in the literature [6]–[8]; however, solutions are known only for some

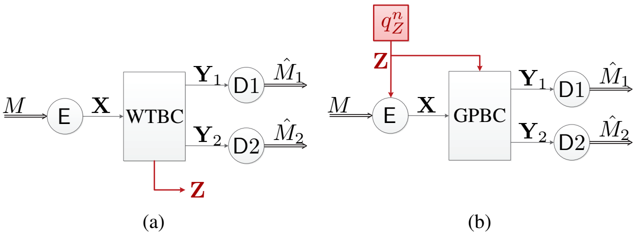


Fig. 2: (a) The WTBC with transition probability $p_{Y_1,Y_2,Z|X}$, where X is the channel input and Y_1, Y_2 and Z are the channel outputs observed by the legitimate receivers and the eavesdropper, respectively; (b) An analogous GPBC is obtained from the WTBC by replacing the eavesdropper's observation with a state random variable $Z \sim q_Z$, revealing $\mathbf{Z}$ in a non-causal manner to the encoder and setting the state-dependent BC $p_{Y_1,Y_2|X,Z}$ as the conditional marginal distribution of the WTBC's transition probability $p_{Y_1,Y_2,Z|X}$.

special cases. To the best of our knowledge, the widest framework of DM WTBCs for which tight secrecy-capacity results are available is due to [8], where, in particular, the region for the SD-WTBC was derived under a further assumption that the eavesdropper is less noisy than the stochastic receiver. The coding scheme therein remains feasible without this less-noisy property; the converse proofs, however, relies on it. Since no corresponding assumption was imposed while deriving the SD-GPBC result from [4], our analogy-based proof method characterizes the SD-WTBC secrecy-capacity regions without assuming this ordering between the sub-channels. As a natural extension to the analogy for the base case (WTCs versus GPCs), the obtained secrecy-capacity regions are described by the same rate bounds as their GPBC counterparts.

An important ingredient in proving the analogy is to adopt the definition of WTC achievability from, e.g., [7], [9], [10], that merges the reliability and security requirements into a single demand on the joint distribution induced by a wiretap code. Specifically, we require that a good sequence of wiretap codes induces a sequence of joint distributions (on the message, its estimate and the eavesdropper's observation) that is asymptotically indistinguishable in total variation from a target measure under which:

1) The message M and its estimate $\hat{M}$ are almost surely equal (a reliability requirement);
2) The eavesdropper's observation is independent of the message and is distributed according to some product measure, say q_Z (a security requirement).

Denoting by $P^{(c_n)}_{M,\hat{M},\mathbf{Z}}$ the joint distribution of M, $\hat{M}$ and $\mathbf{Z}$ induced by a wiretap code c_n, the above requirements mean that for large block lengths $P^{(c_n)}_{M,\hat{M},\mathbf{Z}} \approx P^{(c_n)}_M \mathbb{1}_{\{\hat{M}=M\}} q^n_Z$, where the approximation is in total variation.

With that notion of achievability, we then use distribution approximation arguments to show that such a sequence of wiretap codes induces a sequence of reliable codes for the analogous GPC. The GP encoder and decoder(s) are distilled from the joint distribution induced by the wiretap code by appropriately inverting it. Under this inversion, the asymptotic i.i.d. distribution of the eavesdropper's observation $\mathbf{Z}$ becomes the state distribution in the corresponding GPC. The asymptotic independence of $\mathbf{Z}$ and the message(s) in the WTC's target distribution corresponds to the independence of the message(s) and the state in a GP coding scenario. The performance metric described above strongly related to the more standard notion of achievability used in [11], where performance of a wiretap code was measured via the error probability and the effective secrecy metric. We show that under mild conditions (namely, a super-linear decay of the involved quantities), our definition of achievability and the one from [11] are equivalent.

II. Preliminary Definitions

We set up the problem of a WTBC, which is used in the next section for developing the analogy paradigm. The notations we use are from [12, Section II]. Let $\mathcal{X}$, $\mathcal{Y}_1$, $\mathcal{Y}_2$ and $\mathcal{Z}$ be finite sets (all alphabets throughout this work are assumed to be finite) and let $p_{Y_1,Y_2,Z|X} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Z})$ be a transition probability distribution from $\mathcal{X}$ to $\mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Z}$. The $\left(\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Z}, p_{Y_1,Y_2,Z|X}\right)$ DM-WTBC is illustrated in Fig. 2(a). The sender chooses a pair of messages (m_1, m_2) uniformly at random from product set $\left[1:2^{nR_1}\right] \times \left[1:2^{nR_2}\right]$ and maps it onto a sequence $\mathbf{x} \in \mathcal{X}^n$ (the mapping may be random). The sequence $\mathbf{x}$ is transmitted over the DM-WTBC with transition probability $p_{Y_1,Y_2,Z|X}$. The output sequences $\mathbf{y}_1 \in \mathcal{Y}_1^n$, $\mathbf{y}_2 \in \mathcal{Y}_2^n$ and $\mathbf{z} \in \mathcal{Z}^n$ are observed by Receiver 1, Receiver 2 and the eavesdropper, respectively. Based on $\mathbf{y}_j$, $j = 1, 2$, Receiver j produces an estimate $\hat{m}_j$ of m_j. The eavesdropper tries to glean whatever it can about the transmitted messages (m_1, m_2) from $\mathbf{z}$.

Definition 1 (WTBC Code) *An (n, R_1, R_2)-code c_n for the WTBC with a product message set $\mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)}$, where for $j = 1, 2$ we set $\mathcal{M}_1^{(n)} \triangleq [1 : 2^{nR_j}]$, is a triple of functions $\left(f_n, \phi_1^{(n)}, \phi_2^{(n)}\right)$ such that $f_n : \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)} \to \mathcal{P}(\mathcal{X}^n)$ is a stochastic encoder, and $\phi_j^{(n)} : \mathcal{Y}_j^n \to \mathcal{M}_j^{(n)}$ is the decoding function for Receiver j, for $j = 1, 2$.*

For any (n, R_1, R_2)-code $c_n = \left(f_n, \phi_1^{(n)}, \phi_2^{(n)}\right)$, the induced joint distribution is:

$$P^{(c_n)}(m_{[1:2]}, \mathbf{x}, \mathbf{y}_{[1:2]}, \mathbf{z}, \hat{m}_{[1:2]}) = \frac{1}{\left|\mathcal{M}_1^{(n)}\right|\left|\mathcal{M}_2^{(n)}\right|} f_n(\mathbf{x}|m_{[1:2]}) \\ \times p^n_{Y_1,Y_2,Z|X}(\mathbf{y}_1, \mathbf{y}_2, \mathbf{z}|\mathbf{x}) \mathbb{1}_{\bigcap\limits_{j=1,2}\left\{\hat{m}_j=\phi_j^{(n)}(\mathbf{y}_j)\right\}}, \quad (3)$$

where $m_{[1:2]} \triangleq (m_1, m_2)$ and similarly for $\mathbf{y}_{[1:2]}$ and $\hat{m}_{[1:2]}$.

Our analogy relies on developing a unified perspective on two different problems. We arrive at the desired unification by defining achievability in a manner that is slightly different from typical definitions. Adopting the definition of achievability from [7], [9], [10], we merge the reliability and security requirements into a single requirement on the induced distribution from (3) phrased in terms of total variation.

Definition 2 (WTBC Achievability) *A pair of non-negative real numbers $(R_1, R_2) \in \mathbb{R}^2_+$ is called achievable if there exists a $\gamma > 0$, a probability distribution $q_Z \in \mathcal{P}(\mathcal{Z})$ and* *a sequence of (n, R_1, R_2)-codes $\{c_n\}_{n\in\mathbb{N}}$ such that for any sufficiently large n*

$$\Big|\Big|P^{(c_n)}_{M_{[1:2]},\hat{M}_{[1:2]},Z^n}-p^{(U)}_{\mathcal{M}_1^{(n)}\times\mathcal{M}_2^{(n)}}\mathbb{1}_{\{\hat{M}_{[1:2]}=M_{[1:2]}\}}q_Z^n\Big|\Big|_{\mathsf{TV}}\leq e^{-n\gamma}, \tag{4}$$

where $p^{(U)}_{\mathcal{A}}$ is the uniform distribution over a finite set $\mathcal{A}$.

Remark 1 (Rate of Convergence) *The exponential rate of convergence in* (4) *is not necessary. Any super-linear convergence rate is sufficient for the purposes of this work.*

Remark 2 (Equivalence to Standard Definitions) *The achievability definition in this work is equivalent to the more standard notion of achievability used in [11]. Therein, achievability was defined in terms of a vanishing average error probability and the* effective secrecy *metric that requires*

$$\begin{aligned}&\mathsf{D}\Big(P^{(c_n)}_{M_1,M_2,Z^n}\Big|\Big|p^{(U)}_{\mathcal{M}_1^{(n)}\times\mathcal{M}_2^{(n)}}q_Z^n\Big)\\&\quad=\underbrace{I_{P^{(c_n)}}(M_1,M_2;Z^n)}_{\textit{Strong secrecy measure}}+\underbrace{\mathsf{D}\Big(P^{(c_n)}_{Z^n}\Big|\Big|q_Z^n\Big)}_{\textit{Stealth measure}}\end{aligned} \tag{5}$$

is made arbitrarily small. See [12, Section III-B] for details.

Remark 3 (Target i.i.d. Distribution) *The exact identity of target i.i.d. distribution q_Z^n that approximates the $P^{(c_n)}_{Z^n|M_{[1:2]},\hat{M}_{[1:2]}}$ in* (4) *and* (5) *cannot always be a priori determined solely based on the WTBC's transition kernel $p_{Y_1,Y_2,Z|X}$. The structure of q_Z depends on the sequence of codes $\{c_n\}_{n\in\mathbb{N}}$, and, typically, it can be understood from the proof of achievability.*[1] *Accordingly, the definition of achievability (Definition 2) does not shoot for a specific q_Z; rather, it just requires the existence of any q_Z satisfying* (4).

As usual, the *secrecy-capacity region* $\mathcal{C}_{\mathsf{WT}}(p_{Y_1,Y_2,Z|X})$ is the convex closure of the set of achievable rate pairs.

III. Wiretap and Gelfand-Pinsker Analogy

We describe the analogy principle for the base case of the classic wiretap and GP channels. As a first simple example, the analogy is used to derive a converse proof for the WTC's secrecy-capacity theorem. Then, we outline extensions of this idea to multiuser (namely, broadcasting) scenarios. These extension are subsequently used to prove the main secrecy-capacity results of this work that are stated in Section IV.

A. The Base Case - A Unified Perspective

For simplicity of presentation consider the classic wiretap and GPCs. These problems are related through the fact that their target joint distributions share the same structure. To see this, consider the $p_{Y,Z|X}$ WTC, for which achievability is defined similarly to Definition 2, and the point-to-point GPC with state distribution q_Z and channel transition probability $q_{Y|X,Z}$.[2] The joint distribution induced by an (n,R)-code $c_n=(f_n,\phi_n)$ for the wiretap channel is (see (3))

$$\tilde{P}^{(c_n)}(m,\mathbf{x},\mathbf{y},\mathbf{z},\hat{m})=\frac{1}{|\mathcal{M}_n|}f_n(\mathbf{x}|m)p^n_{Y,Z|X}(\mathbf{y},\mathbf{z}|\mathbf{x})\mathbb{1}_{\{\hat{m}=\phi_n(\mathbf{y})\}} \tag{6}$$

while the induced distribution for the GPC with respect to an (n,R)-code $b_n=(g_n,\psi_n)$, where $g_n:\mathcal{M}_n\times\mathcal{Z}\to\mathcal{P}(\mathcal{X})$ is a stochastic encoder and $\phi_n:\mathcal{Y}^n\to\mathcal{M}_n$ is the decoder, is

$$\begin{aligned}\tilde{Q}^{(b_n)}(\mathbf{z},m,\mathbf{x},\mathbf{y},\hat{m})=q_Z^n(\mathbf{z})\frac{1}{|\mathcal{M}_n|}g_n(\mathbf{x}|\mathbf{z},m)q^n_{Y|X,Z}(\mathbf{y}|\mathbf{x},\mathbf{z})\\ \times\mathbb{1}_{\{\hat{m}=\psi_n(\mathbf{y})\}}.\end{aligned} \tag{7}$$

With respect to Definition 2, a non-negative real number R is achievable for the WTC if there exist a distribution $q_Z\in\mathcal{P}(\mathcal{Z})$ and a sequence of (n,R)-codes $\{c_n\}_{n\in\mathbb{N}}$, such that

$$\Big|\Big|\tilde{P}^{(c_n)}_{M,\hat{M},Z^n}-p^{(U)}_{\mathcal{M}_n}\mathbb{1}_{\{\hat{M}=M\}}q_Z^n\Big|\Big|_{\mathsf{TV}}\xrightarrow[n\to\infty]{}0. \tag{8}$$

For the GPC, it can be shown that under mild conditions,[3] a vanishing error probability is equivalent to

$$\Big|\Big|\tilde{Q}^{(c_n)}_{M,\hat{M},Z^n}-p^{(U)}_{\mathcal{M}_n}\mathbb{1}_{\{\hat{M}=M\}}q_Z^n\Big|\Big|_{\mathsf{TV}}\xrightarrow[n\to\infty]{}0. \tag{9}$$

For details, see [12, Section IV-A-1].

Having (8) and (9), it is evident that while each problem has its own induced joint distribution, their target measures share the same structure. In both problems, a "good" sequence of codes induces a sequence of distributions ($\big\{\tilde{P}^{(c_n)}\big\}_{n\in\mathbb{N}}$ or $\big\{\tilde{Q}^{(b_n)}\big\}_{n\in\mathbb{N}}$ for the WTC or the GPC, respectively) that approximates a target distribution where: (i) $M=\hat{M}$ almost surely; (ii) $\mathbf{Z}$ is independent of M. The first item is a consequence of the reliability requirement in both problems. For the second item, note that, while the independence of $\mathbf{Z}$ and M is the security requirement in the WTC scenario, it is actually part of the problem definition for the GPC. The above described correspondence between the WTC and the GPC stands at the heart of the analogy between them.

B. Analogy Between Multiuser Setups

As a natural extension to the ideas from Section III-A, we now describe the analogy between WTBCs and GPBCs. Consider a WTBC $\big(\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{Z},p_{Y_1,Y_2,Z|X}\big)$ as defined in Section II. An analogous GPBC is constructed in three steps (see Fig. 2):

1) Replace the eavesdropper of the WTBC with a state sequence $\mathbf{Z}\sim q_Z^n$, where q_Z^n is the target product measure from the definition of WTBC achievability (see Definition 2);
2) Non-causally reveal $\mathbf{Z}$ to the encoder;
3) Set the GPBC's transition probability as the conditional marginal distribution $p_{Y_1,Y_2|X,Z}$.

The produced analogous $\big(\mathcal{Z},\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,q_Z,p_{Y_1,Y_2|X,Z}\big)$ GPBC inherits the properties the WTBC possesses (e.g.,

[1] For instance, for the degraded binary symmetric WTBC with crossover probabilities p_L and p_E for the legitimate and eavesdropper channels, respectively, where $p_L<p_B$, one may verify that q_Z may be chosen as a product $\mathsf{Ber}\left(\frac{1}{2}\right)$ measure. This is a consequence of the optimal input distribution that attains that secrecy-capacity $h(p_E)-h(p_L)$ being $\big(\mathsf{Ber}\left(\frac{1}{2}\right)\big)^n$.

[2] We adhere to the standard definitions for GPCs, see, e.g., [13, Setion 7.6].

[3] namely, a super-linear decay of the error probability

deterministic components, order of degradeness, etc). For example, if the WTBC is SD $p_{Y_1,Y_2,Z|X} = \mathbb{1}_{\{Y_1=y_1(X)\}} p_{Y_2,Z|X}$, then so is the GPBC since $p_{Y_1,Y_2|X,Z} = \mathbb{1}_{\{Y_1=y_1(X)\}} p_{Y_2|X,Z}$. If one of the observed signals of the legitimate receivers is a degraded version of the other, then the same ordering applies for the signal intercepted by the receivers of the GPBC. The analogy also accounts for WTBC settings with cooperative components. Namely, if the receivers of the WTBC are connected by, e.g., a finite-capacity bit-pipe, then the same applies for the receivers of the analogous GPBC.

As for the base case, the capacity regions of two analogous wiretap and GP BCs are described by rate bounds of the same structure. The underlying distribution and the part thereof over which we take the union is, however, different. This relation between the regions is emphasized in Section IV.

Since GPBCs have been extensively treated in the literature and capacity results are available for numerous cases [3]–[5], the analogy allows leveraging these results to study corresponding WTBCs. This is done by relating the performance of two analogous models as follows. Due to lack of space, the proof of the proposition is omitted; the reader is referred to [12] for details.

Proposition 1 (Good Wiretap Codes and Good GP Codes) *Consider a $\big(\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{Z},p_{Y_1,Y_2,Z|X}\big)$ WTBC. Let $(R_1,R_2) \in \mathbb{R}_+^2$ be an achievable rate pair for the WTBC, with a corresponding sequence of (n,R_1,R_2)-codes $\{c_n\}_{n\in\mathbb{N}}$, where $c_n = \Big(f_n,\phi_1^{(n)},\phi_2^{(n)}\Big)$, for each $n \in \mathbb{N}$. For every $n \in \mathbb{N}$, define $g_n \triangleq P^{(c_n)}_{\mathbf{X}|\mathbf{Z},M_{[1:2]}}$ and $\psi_j^{(n)} \triangleq \phi_j^{(n)}$, for $j = 1,2$, where $P^{(c_n)}_{\mathbf{X}|\mathbf{Z},M_{[1:2]}}$ is the conditional marginal distribution of $\mathbf{X}$ given $(\mathbf{Z},M_1,M_2)$ with respect to $P^{(c_n)}$ from* (3) *induced by the n-th wiretap code c_n. Then:*

1) $b_n \triangleq \Big(g_n,\psi_1^{(n)},\psi_2^{(n)}\Big)$ is an (n,R_1,R_2)-code for the $\big(\mathcal{Z},\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,q_Z,p_{Y_1,Y_2|X,Z}\big)$ GPBC.

2) The distribution $Q^{(b_n)}_{\mathbf{Z},M_{[1:2]},\mathbf{X},\mathbf{Y}_{[1:2]},\hat{M}_{[1:2]}}$ induced by b_n (analogous to $\tilde{Q}^{(b_n)}$ from (7) *with $M_{[1:2]}$, $\mathbf{Y}_{[1:2]}$ and $\hat{M}_{[1:2]}$ in the roles of M, $\mathbf{Y}$ and $\hat{M}$ therein, respectively) satisfies $\big|\big|P^{(c_n)} - Q^{(b_n)}\big|\big|_{\mathsf{TV}} \leq e^{-n\gamma}$, for any n large enough.*

3) The sequence of codes $\{b_n\}_{n\in\mathbb{N}}$ attains $\mathsf{P_e}(b_n) \xrightarrow[n\to\infty]{} 0$, and consequently, (R_1,R_2) is an achievable rate pair for the aforementioned GPBC.

Proof: *For simplicity of notation, throughout the proof we denote $M_{12} \triangleq M_{[1:2]}$, $m_{12} \triangleq m_{[1:2]}$, $\hat{M}_{12} \triangleq \hat{M}_{[1:2]}$, $\hat{m}_{12} \triangleq \hat{m}_{[1:2]}$ and $\mathcal{M}_{12} \triangleq \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)}$. The first claim is straightforward as for each $n \in \mathbb{N}$, $P^{(c_n)}_{\mathbf{X}|\mathbf{Z},M_{12}}$ and $\psi_j^{(n)}$, for $j = 1,2$, are valid (stochastic) encoder and decoders for the GPBC. For (2), fix $n \in \mathbb{N}$, and first observe*

$$
\begin{aligned}
&P^{(c_n)}_{M_{12},\mathbf{X},\mathbf{Y}_{[1:2]},\mathbf{Z},\hat{M}_{12}}\\
&\quad\overset{(a)}{=} P^{(c_n)}_{M_{12},\mathbf{Z}} \cdot g_n \cdot p^n_{Y_1,Y_2|X,Z} \cdot \mathbb{1}_{\bigcap_{j=1,2}\left\{\hat{M}_j=\psi_j^{(n)}(\mathbf{Y}_j)\right\}}\\
&\quad\overset{(b)}{=} P^{(c_n)}_{M_{12},\mathbf{Z}} \cdot Q^{(b_n)}_{\mathbf{X},\mathbf{Y}_{[1:2]},\hat{M}_{12}|M_{12},\mathbf{Z}}
\end{aligned}
\tag{10}
$$

where (a) follows by the factorization of $P^{(c_n)}$ from (3)*, while (b) is because $b_n = \Big(g_n,\psi_1^{(n)},\psi_2^{(n)}\Big)$ and due to the structure of $Q^{(b_n)}$. Recalling that $Q^{(c_n)}_{\mathbf{Z},M_{12}} = q_Z^n \cdot p^{(U)}_{\mathcal{M}_{12}}$, we have*

$$
\big|\big|P^{(c_n)} - Q^{(b_n)}\big|\big|_{\mathsf{TV}} = \Big|\Big|P^{(c_n)}_{M_{12},\mathbf{Z}} - p^{(U)}_{\mathcal{M}_{12}} \cdot q_Z^n\Big|\Big|_{\mathsf{TV}} \xrightarrow[n\to\infty]{} 0.
\tag{11}
$$

Claim (3) follows because $\mathsf{P_e}(b_n)$ is upper bounded as

$$
\begin{aligned}
\mathsf{P_e}(b_n) &= \sum_{\substack{m_{12},\hat{m}_{12}:\\ m_{12}\neq\hat{m}_{12}}} \Big[Q^{(c_n)}(m_{12},\hat{m}_{12}) - p^{(U)}_{\mathcal{M}_{12}}(m_{12})\mathbb{1}_{\{\hat{m}_{12}=m_{12}\}}\Big]\\
&\overset{(a)}{=} \Big|\Big|Q^{(c_n)}_{M_{12},\hat{M}_{12}} - p^{(U)}_{\mathcal{M}_1^{(n)}\times\mathcal{M}_2^{(n)}}\mathbb{1}_{\left\{\hat{M}_{12}=M_{12}\right\}}\Big|\Big|_{\mathsf{TV}}\\
&\overset{(b)}{\leq} \Big|\Big|Q^{(b_n)}_{M_{12},\hat{M}_{12}} - P^{(c_n)}_{M_{12},\hat{M}_{12}}\Big|\Big|_{\mathsf{TV}}\\
&\qquad\qquad + \Big|\Big|P^{(c_n)}_{M_{12},\hat{M}_{12}} - p^{(U)}_{\mathcal{M}_{12}}\mathbb{1}_{\{\hat{M}_{12}=M_{12}\}}\Big|\Big|_{\mathsf{TV}}\\
&\overset{(c)}{\leq} \Big|\Big|Q^{(b_n)} - P^{(c_n)}\Big|\Big|_{\mathsf{TV}} + \Big|\Big|P^{(c_n)}_{M_{12},\hat{M}_{12},\mathbf{Z}} - p^{(U)}_{\mathcal{M}_{12}}\mathbb{1}_{\{\hat{M}_{12}=M_{12}\}}q_Z^n\Big|\Big|_{\mathsf{TV}}
\end{aligned}
$$

where (a) is because $||p-q||_{\mathsf{TV}} = \sum_{x:\ p(x)>q(x)}\big[p(x)-q(x)\big]$ and since $m_{12} \neq \hat{m}_{12}$ if and only if $Q^{(c_n)}(m_{12},\hat{m}_{12}) \geq p^{(U)}_{\mathcal{M}_{12}}(m_{12})\mathbb{1}_{\left\{\hat{m}_{12}=m_{12}\right\}}$; (b) is the triangle inequality; (c) uses Property (3-a) from [12, Lemma 1]. Finally, the RHS above vanishes to 0 as $n \to \infty$ by (11) *and our hypothesis.* ■

IV. The Secrecy-Capacity Region of the SD-WTBC

We give a single-letter characterization of the secrecy-capacity region of the SD-WTBC. A WTBC is SD if $p_{Y_1,Y_2,Z|X} = \mathbb{1}_{\{Y_1=y_1(X)\}} p_{Y_2,Z|X}$, where $y_1 : \mathcal{X}\times\mathcal{Z} \to \mathcal{Y}_1$ and $p_{Y_2,Z|X} : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_2\times\mathcal{Z})$. Until now, the secrecy-capacity region of this setup was known only under the assumption that the stochastic channel is less noisy than the channel to the eavesdropper [8, Theorem 5]. Our analogy-based converse proof makes this assumption unnecessary.

Theorem 1 (Secrecy-Capacity) *The secrecy-capacity region of the $\big(\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{Z},\mathbb{1}_{\{Y_1=y_1(X)\}}p_{Y_2,Z|X}\big)$ SD-WTBC is given by the union of rate pairs $(R_1,R_2) \in \mathbb{R}_+^2$ satisfying:*

$$R_1 \leq H(Y_1|Z), \tag{12a}$$

$$R_2 \leq I(U;Y_2) - I(U;Z), \tag{12b}$$

$$R_1 + R_2 \leq H(Y_1|Z) + I(U;Y_2) - I(U;Y_1,Z) \tag{12c}$$

where the union is over all $p_{U,X} \in \mathcal{P}(\mathcal{U}\times\mathcal{X})$, each inducing a joint distribution $p_{U,X}\mathbb{1}_{\{Y_1=y_1(X)\}}p_{Y_2,Z|X}$. Furthermore, one may restrict the auxiliary random variable U to take values in a set $\mathcal{U}$ whose cardinality is bounded by $|\mathcal{U}| \leq |\mathcal{X}| + 1$.

The direct part of Theorem 1 relies on a specialization of the inner bound on the secrecy-capacity region of the WTBC derived in [7, Theorem 3]. As the performance criterion in that work corresponds to the definition of achievability used herein (Definition 2), the result from [7] applies for our setup. Setting $Q = U_0 = 0$, $U_1 = Y_1$ and recasting U_2 as U reduces the rate bounds from [7, Theorem 3] to those from (12). Since $Y_1 = y_1(X)$, this choice of the auxiliaries $\big(Q,U_{[0:2]}\big)$ is feasible. The analogy-based converse proof is given next.

Converse Proof: Let $(R_1, R_2) \in \mathbb{R}_+^2$ be an achievable rate pair for the SD-WTBC and $\{c_n\}_{n\in\mathbb{N}}$ be the corresponding sequence of (n, R_1, R_2)-codes satisfying (4) for some $\gamma > 0$ and $q_Z \in \mathcal{P}(\mathcal{Z})$, and any n large enough. By Proposition 1, $\{c_n\}_{n\in\mathbb{N}}$ gives rise to a sequence of (n, R_1, R_2)-codes $\{b_n\}_{n\in\mathbb{N}}$ for the $\big(\mathcal{Z}, \mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, q_Z, p_{Y_1,Y_2|X,Z}\big)$ GPBC, each inducing a joint distribution $Q^{(b_n)}$, such that:

1) $\big|\big|P^{(c_n)} - Q^{(b_n)}\big|\big|_{\mathsf{TV}} \leq e^{-n\gamma}$, for any large enough n, where $P^{(c_n)}$ is the distribution from (3) induced by c_n.
2) $\mathsf{P_e}(b_n) \xrightarrow[n\to\infty]{} 0$.

Furthermore, note that since the WTBC is SD, i.e., its transition probability factors as $p_{Y_1,Y_2,Z|X} = \mathbb{1}_{\{Y_1=y_1(X)\}} p_{Y_2,Z|X}$, the obtained GPBC is also SD. Namely, the GPBC's transition probability decomposes as $p_{Y_1,Y_2|X,Z} = \mathbb{1}_{\{Y_1=y_1(X)\}} p_{Y_2|X,Z}$, which falls under the framework of [4, Theorem 1].

The converse proof of [4, Theorem 1] for the SD-GPBC shows that if $\{b_n\}_{n\in\mathbb{N}}$ is a sequence of (n, R_1, R_2)-codes with a vanishing error probability, then

$$R_1 \leq \frac{1}{n}\sum_{i=1}^{n} H_Q(Y_{1,i}|Z_i) + \epsilon_n \tag{13a}$$

$$R_2 \leq \frac{1}{n}\sum_{i=1}^{n} \Big[I_Q\big(M_2, Y_2^{i-1}, Z_{i+1}^n; Y_{2,i}\big) - I_Q\big(M_2, Y_2^{i-1}, Z_{i+1}^n; Z_i\big)\Big] + \epsilon_n \tag{13b}$$

$$R_1 + R_2 \leq \frac{1}{n}\sum_{i=1}^{n} \Big[I_Q\big(M_2, Y_2^{i-1}, Z_{i+1}^n, Y_{1,i+1}^n; Y_{2,i}\big) + H_Q(Y_{1,i}|Z_i) - I_Q\big(M_2, Y_2^{i-1}, Z_{i+1}^n, Y_{1,i+1}^n; Z_i, Y_{2,i}\big)\Big] + \epsilon_n, \tag{13c}$$

where the subscript Q indicates that the underlying distribution is $Q^{(b_n)}$ and $\epsilon_n \triangleq \frac{2}{n} + \mathsf{P_e}(b_n)\sum_{j=1,2} R_j$. Since the total variation of two distribution upper bounds the total variation between their marginals [12, Property (3-a), Lemma 1]),

$$\bigg|\bigg|P^{(c_n)}_{M_2,Y^i,Z_i^n} - Q^{(b_n)}_{M_2,Y^i,Z_i^n}\bigg|\bigg|_{\mathsf{TV}} \leq e^{-n\gamma} \tag{14}$$

for large n, uniformly in $i \in [1:n]$. Recall that over finite probability spaces an exponentially decaying total variation dominates the difference between two corresponding mutual information terms (see [12, Lemma 3]). Combining this observation with (14), we may replace the information measures from the RHS of (13) that are taken with respect to $Q^{(b_n)}$ with the same terms, but with an underlying distribution $P^{(c_n)}$ (which we denote by a subscript P) plus a vanishing term. Namely, there exists a $\delta > 0$, such that for n large enough

$$R_1 \leq \frac{1}{n}\sum_{i=1}^{n} H_P(Y_{1,i}|Z_i) + \epsilon_n + e^{-n\delta} \tag{15a}$$

$$R_2 \leq \frac{1}{n}\sum_{i=1}^{n} \Big[I_P(V_i; Y_{2,i}) - I_P(V_i; Z_i)\Big] + \epsilon_n + 2e^{-n\delta} \tag{15b}$$

$$R_1 + R_2 \leq \frac{1}{n}\sum_{i=1}^{n} \Big[H_P(Y_{1,i}|Z_i) + I_P(V_i, T_i; Y_{2,i}) - I_P(V_i, T_i; Y_{1,i}, Z_i)\Big] + \epsilon_n + 3e^{-n\delta} \tag{15c}$$

where, for every $i \in [1:n]$, we have defined $V_i \triangleq \big(M_2, Y_2^{i-1}, Z_{i+1}^n\big)_P$ and $T_i \triangleq \big(Y_{1,i+1}^n\big)_P$, with the subscript P indicating that the underlying distribution is $P^{(c_n)}$.

Letting n tend to infinity in (15), we see that any achievable rate pair (R_1, R_2) must be contained in the convex closure of the union of rate pairs satisfying:

$$R_1 \leq H_p(Y_1|Z) \tag{16a}$$

$$R_2 \leq I_p(V; Y_2) - I_p(V; Z) \tag{16b}$$

$$R_1 + R_2 \leq H_p(Y_1|Z) + I_p(V, T; Y_2) - I_p(V, T; Y_1, Z) \tag{16c}$$

where the union is over all $p_{V,T,X} \in \mathcal{P}(\mathcal{V} \times \mathcal{T} \times \mathcal{X})$, each inducing a joint distribution $p \triangleq p_{V,T,X}\, p_{Y_1,Y_2,Z|X}$, i.e., $(Y_1, Y_2, Z) -\!\!\circ\!\!- X -\!\!\circ\!\!- (V, T)$ forms a Markov chain. This Markov relation follows because $(Y_{1,i}, Y_{2,i}, Z_i) -\!\!\circ\!\!- X_i -\!\!\circ\!\!- \big(M_2, Y_{1,i+1}^n, Y_2^{i-1}, Z_{i+1}^n\big)$, for all $i \in [1:n]$, under $P^{(c_n)}$.

To conclude the proof it remains to show that there exists an auxiliary random variable U, such that for any (V, T):

$$I_p(V; Y_2) - I_p(V; Z) \leq I_p(U; Y_2) - I_p(U; Z) \tag{17a}$$

$$H_p(Y_1|Z) + I_p(V, T; Y_2) - I_p(V, T; Y_1, Z) \leq H_p(Y_1|Z) + I_p(U; Y_2) - I_p(U; Y_1, Z). \tag{17b}$$

This is established by closely following the arguments from the end of the converse proof of the analogous SD-GPBC [4, Section III], as outlined next. Setting $U = V$ if p is such that $I_p(T; Y_2|V) - I_p(T; Y_1, Z|V) \leq 0$, and $U = (V, T)$ if $I_p(T; Y_2|V) - I_p(T; Z|V) \geq 0$ suffices. Finally, noting that every distribution p must satisfy at least one of these information inequalities concludes the converse.

References

[1] A. D. Wyner. The wire-tap channel. *Bell Sys. Techn.*, 54(8):1355–1387, Oct. 1975.

[2] S. I. Gelfand and M. S. Pinsker. Coding for channel with random parameters. *Problemy Pered. Inform. (Problems of Inf. Trans.)*, 9(1):19–31, 1980.

[3] Y. Steinberg. Coding for the degraded broadcast channel with random parameters, with causal and noncausal side information. *IEEE Trans. Inf. Theory*, 51(8):2867–2877, Aug. 2005.

[4] A. Lapidoth and L. Wang. The state-dependent semideterministic broadcast channel. *IEEE Trans. Inf. Theory*, 59(4):2242–2251, 2013.

[5] L. Dikstein, H. H. Permuter, and Y. Steinberg. On state-dependent broadcast channels with cooperation. *IEEE Trans. Inf. Theory*, 62(5):2308–2323, May 2016.

[6] E. Ekrem and S. Ulukus. Multi-receiver wiretap channel with public and confidential messages. *IEEE Trans. Inf. Theory*, 59(4):2165–2177, Apr. 2013.

[7] M. H. Yassaee, M. R. Aref, and A. Gohari. Achievability proof via output statistics of random binning. *IEEE Trans. Inf. Theory*, 60(11):6760–6786, Nov. 2014.

[8] M. Benammar and P. Piantanida. Secrecy capacity region of some classes of wiretap broadcast channels. *IEEE Trans Inf. Theory*, 61(10):5564–5582, Oct. 2015.

[9] H. Tyagi and S. Watanabe. Converses for secret key agreement and secure computing. *IEEE Trans. Inf. Theory*, 61(9):4809–4827, 2015.

[10] M. H. Yassaee. One-shot achievability via fidelity. In *Proc. Int. Symp. Inf. Theory (ISIT-2015)*, pages 301–305, Hong Kong, China, Jun. 2015.

[11] J. Hou and G. Kramer. Effective secrecy: Reliability, confusion and stelth. In *IEEE Int. Symp. Inf. Theory (ISIT-2014)*, Honolulu, HI, USA, Jun.-Jul. 2014.

[12] Z. Goldfeld, , and H. H. Permuter. Wiretap and gelfand-pinsker channels analogy and its applications. *Submitted for publication to IEEE Trans. Inf. Theory*, 2017.

[13] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2011.

Optical Wiretap Channel with Input-Dependent Gaussian Noise Under Peak Intensity Constraint

Morteza Soltani and Zouheir Rezki
University of Idaho
Department of Electrical and Computer Engineering
83844 Moscow, Idaho
Email: solt8821@vandals.uidaho.edu, zrezki@uidaho.edu

Abstract—**This paper studies the optical wiretap channel with input-dependent Gaussian noise, in which the main distortion is caused by an additive Gaussian noise whose variance depends on the current signal strength. Subject to non-negativity and peak-intensity constraints on the channel input, we first evaluate the conditions under which this wiretap channel is stochastically degraded. We then study the secrecy-capacity-achieving input distribution of this wiretap channel and prove it to be discrete with a finite number of mass points. Moreover, we show that the entire rate-equivocation region of this wiretap channel is also obtained by discrete input distributions with a finite support. Similar to the case for the Gaussian wiretap channel under a peak-power constraint, here too, we observe that under non-negativity and peak-intensity constraints, there is a tradeoff between the secrecy capacity and the capacity in the sense that both may not be achieved simultaneously.**

I. Introduction

Exchanging confidential information over a communication medium (wired, wireless or optical) in the presence of unauthorized eavesdroppers has been always a challenging problem for system designers. This problem has been conventionally addressed by cryptographic encryption [1] without considering the imperfections introduced by the communication channel. In this model, using *secret keys* are the main approach for having secure communications. Wyner [2], on the other hand, proved the possibility of secure communications without relying on encryption by introducing the notion of a stochastically degraded wiretap channel.

For the class of degraded wiretap channels, it has been established in [2] that there exists a single-letter characterization for the rate-equivocation region. Authors in [3] studied the Gaussian wiretap channel under an average power constraint and obtained a single-letter expression for the entire rate-equivocation region. Particularly, they showed that under an average power constraint, the Gaussian distribution is the optimal input distribution for attaining both the capacity and secrecy capacity with no compromise between the communication rate and the equivocation rate at the eavesdropper. On the other hand, under a peak-power constraint, the work in [4] proved that the entire rate-equivocation region of the Gaussian wiretap channel is achieved by discrete input distributions with finite support. More specifically, the secrecy-capacity achieving input distribution may not be identical to the capacity-achieving counterpart in general, resulting in a tradeoff between the rate and its equivocation.

This work considers an *optical* wiretap channel based on intensity modulation with input-dependent Gaussian noise which consists of a transmitter, a legitimate user and an eavesdropper. We first evaluate the conditions under which the optical wiretap channel with input-dependent Gaussian noise is stochastically degraded. We then use the results in [2] to conclude that there exists a single-letter expression for the entire rate-equivocation region. Next, we employ the functional optimization problem addressed in [5] to obtain necessary and sufficient conditions, also known as Karush-Kuhn-Tucker (KKT) conditions, for the optimal input distribution. Finally, we prove by contradiction that the secrecy capacity as well as the entire rate-equivocation region of this wiretap channel, are obtained by discrete input distributions with a finite number of mass points. We provide numerical results which demonstrate that similar to the case of the Gaussian wiretap channel under a peak-power constraint, here too, the secrecy capacity and the capacity are not achieved by the same distribution in general. This, in turn, implies that for this wiretap channel, there is a tradeoff between the rate and its equivocation.

Due to the existence of input-dependent noise components, our technical proofs differ from those of [4]. Our analysis for showing the analyticity of the mutual information densities is more challenging. Additionally, our contradiction statements for proving the discreteness of the optimal input distribution are different.

II. System Model

In the considered optical wiretap channel, the channel input X is a nonnegative random variable representing the intensity of the optical signal. Since intensity is constrained due to practical and safety restrictions by a peak constraint in general, the input has to satisfy $X \leq A$ [6]. Therefore, the channel input is constrained as

$$0 \leq X \leq A. \tag{1}$$

This work has been supported by King Abdullah University of Science and Technology (KAUST), under a competitive research grant (CRG) OSR-2016-CRG5-2958-01.

In this setup, each link is a memoryless channel and is defined by [7]

$$Y = X + \sqrt{X} N_{B,1} + N_{B,0}, \tag{2}$$

$$Z = X + \sqrt{X} N_{E,1} + N_{E,0}, \tag{3}$$

where Y and Z denote the legitimate user's and the eavesdropper's observations, respectively. $N_{B,0}$ and $N_{E,0}$ are independent identically distributed (i.i.d.) zero-mean Gaussian random variables with variances σ_B^2 and σ_E^2, describing the input-independent noise components at the legitimate user and the eavesdropper, respectively. $N_{B,1}$ and $N_{E,1}$ are i.i.d. zero-mean Gaussian random variables with variances $\eta_B^2\sigma_B^2$ and $\eta_E^2\sigma_E^2$, describing the input-dependent noise components at the legitimate user and the eavesdropper, respectively, where η_B^2 and η_E^2 are the ratios of the input-dependent noise variances to the input-independent noise variances of the legitimate user's and the eavesdropper's channels, respectively. Furthermore, $N_{B,0}$ and $N_{B,1}$ are assumed to be independent and so are $N_{E,0}$ and $N_{E,1}$.

In this optical wiretap channel, since the input-dependent distortion is caused by the laser diode at the transmitter side [7], we consider the input-dependent noise components in both legitimate user's and wiretap channels to be statistically equivalent, i.e., $\sigma_B^2\eta_B^2 = \sigma_E^2\eta_E^2$. However, the variance of the input-independent noise of the wiretap channel is assumed to be strictly greater than that of the legitimate user's channel, i.e., $\sigma_E^2 > \sigma_B^2$ (otherwise the secrecy capacity defined later in this Section is zero). Therefore, under the conditions $\sigma_B^2\eta_B^2 = \sigma_E^2\eta_E^2$ and $\sigma_E^2 > \sigma_B^2$, the random variables X, Y and Z form a Markov chain $X \to Y \to Z$ and consequently the optical wiretap channel becomes stochastically degraded. As a result, the rate-equivocation region of such an optical wiretap channel can be expressed in a single-letter form due to [2].

An $(n, 2^{nR})$ code for the peak intensity-constrained optical wiretap channel with input-dependent Gaussian noise consists of the random variable W (message set) uniformly distributed over the set $\mathcal{W} = \{1, 2, \cdots, 2^{nR}\}$, an encoder at the transmitter $f_n : \mathcal{W} \to [0, A]^n$ satisfying the non-negativity and peak-intensity constraints, and a decoder at the legitimate user $g_n : \mathbb{R}^n \to \mathcal{W}$. Equivocation of a code is measured by the normalized conditional entropy $\frac{1}{n} H(W|Z^n)$. The probability of error for such a code is defined as $P_e^n = \Pr\left\{g_n(Y^n) \neq W\right\}$. A rate-equivocation pair (R, R_e) is said to be achievable if there exists an $(n, 2^{nR})$ code satisfying

$$\lim_{n\to\infty} P_e^n = 0, \tag{4}$$

$$R_e \leq \lim_{n\to\infty} \frac{1}{n} H(W|Z^n). \tag{5}$$

The rate-equivocation region consists of all achievable rate-equivocation pairs, and is denoted by $\mathcal{E}$. A rate R is said to be perfectly secure if we have $R_e = R$, i.e., if there exists an $(n, 2^{nR})$ code satisfying $\lim_{n\to\infty} \frac{1}{n} I(W; Z^n) = 0$. The supremum of such rates is defined to be the secrecy capacity and denoted by C_S.

The entire rate-equivocation region of the optical wiretap channel is given by the union of the rate-equivocation pairs (R, R_e) such that [2]

$$R \leq I(X; Y), \tag{6}$$

$$R_e \leq I(X; Y) - I(X; Z), \tag{7}$$

for some input distribution $F_X \in \mathcal{A}^+$, where $I(X; Y)$ and $I(X; Z)$ are the mutual information of the legitimate user's and the eavesdropper's channels, respectively, and the feasible set $\mathcal{A}^+$ is given by

$$\mathcal{A}^+ \triangleq \left\{F_X : \int_0^A dF_X(x) = 1\right\}. \tag{8}$$

III. Main Results

This section presents the main results about the optical wiretap channel with input-dependent Gaussian noise when non-negativity and peak-intensity constraints are imposed on the channel input.

A. Results on the Secrecy Capacity

The secrecy capacity of the optical wiretap channel with input-dependent Gaussian noise under non-negativity and peak-intensity constraints is given by the solution of the following optimization problem

$$\max_{F_X \in \mathcal{A}^+} g_0(F_X), \tag{9}$$

where $g_0(F_X) = I(X; Y) - I(X; Z)$ is the objective function of the optimization problem.

Under the constraints (1), the solution of (9) is discrete with a finite support as stated by Theorem 1.

Theorem 1. *There exists a unique input distribution that attains the secrecy capacity of the optical wiretap channel with input-dependent Gaussian noise under non-negativity and peak-intensity constraints. Furthermore, the support set of this optimal input distribution is a finite set.*

Proof. The proof is provided in Section IV. ■

B. Results on the Rate-Equivocation Region

By a time-sharing argument, the rate-equivocation region of the optical wiretap channel with input-dependent Gaussian noise is convex. Therefore, the region can be characterized by finding tangent lines to $\mathcal{E}$, which are given by the solutions of

$$\max_{F_X \in \mathcal{A}^+} g_\lambda(F_X), \tag{10}$$

where $g_\lambda(F_X) = \lambda I(X; Y) + (1 - \lambda)\left[I(X; Y) - I(X; Z)\right]$ for all $\lambda \in [0, 1]$. Next, we establish that the entire rate-equivocation region of the optical wiretap channel with input-dependent Gaussian noise under constraints(1) is also obtained by discrete input distributions with finite supports.

Theorem 2. *There exists a unique input distribution that achieves the boundary of the rate-equivocation region of the optical wiretap channel with input-dependent Gaussian* *noise under non-negativity and peak-intensity constraints. This optimal input distribution is discrete with a finite support.*

Proof. Due to length constraint, the proof is given in [8, Section IV-D]. ■

It is worth mentioning that for the case when η_B^2 and η_E^2 are 0 (i.e., the optical wiretap channel with input-independent Gaussian noise), similar approaches to those presented in [4] can be used to prove the discreteness of the optimal solutions of (9) and (10). An interesting observation is that our approach for proving the discreteness of the optimal solutions of (9) and (10) when $\eta_B^2, \eta_E^2 \neq 0$ cannot be generalized to the case when $\eta_B^2 = \eta_E^2 = 0$. This can also be observed in [9].

IV. Proof of the Main Results

A. Preliminaries and Notation

Since both channels are AWGN with input-dependent noise, the output densities for Y and Z exist for any input distribution F_X, and are given by

$$P_Y(y; F_X) = \int_0^A p(y|x)\, dF_X(x), \;\; y \in \mathbb{R} \tag{11}$$

$$P_Z(z; F_X) = \int_0^A p(z|x)\, dF_X(x), \;\; z \in \mathbb{R} \tag{12}$$

where $p(y|x)$ and $p(z|x)$ are given by [7]

$$p(y|x) = \frac{1}{\sqrt{2\pi\, \sigma_{B,X}^2(x)}} \exp\Big(-\frac{(y-x)^2}{2\,\sigma_{B,X}^2(x)}\Big), \tag{13}$$

$$p(z|x) = \frac{1}{\sqrt{2\pi\, \sigma_{E,X}^2(x)}} \exp\Big(-\frac{(z-x)^2}{2\,\sigma_{E,X}^2(x)}\Big), \tag{14}$$

where $\sigma_{B,X}^2(x) = \sigma_B^2(1+\eta_B^2 x)$ and $\sigma_{E,X}^2(x) = \sigma_E^2(1+\eta_E^2 x)$. We define the rate-equivocation density $r_e(x; F_X)$ as

$$r_e(x; F_X) = i_B(x; F_X) - i_E(x; F_X), \tag{15}$$

where $i_B(x; F_X)$ and $i_E(x; F_X)$ are the mutual information densities for the legitimate user's and eavesdropper's channel, respectively and are given by

$$\begin{aligned} i_B(x; F_X) = &- \int_{\mathbb{R}} p(y|x) \log\left(P_Y(y; F_X)\right) dy \\ &- \frac{1}{2}\log\left(2\pi e \sigma_{B,X}^2(x)\right), \end{aligned} \tag{16}$$

$$\begin{aligned} i_E(x; F_X) = &- \int_{\mathbb{R}} p(z|x) \log\left(P_Z(z; F_X)\right) dz \\ &- \frac{1}{2}\log\left(2\pi e \sigma_{E,X}^2(x)\right). \end{aligned} \tag{17}$$

The mutual information and the mutual information density are related through

$$I(X;Y) = \int_0^A i_B(x; F_X)\, dF_X(x), \tag{18}$$

$$I(X;Z) = \int_0^A i_E(x; F_X)\, dF_X(x). \tag{19}$$

One can show that the conditional densities in (13) and (14) are bounded as [9, Lemma 3]

$$\exp(-\alpha - \beta' y^2) \leq p(y|x) \leq \exp(\alpha - \beta y^2), \tag{20}$$

$$\exp(-\mu - \xi' z^2) \leq p(z|x) \leq \exp(\mu - \xi z^2), \tag{21}$$

for all $x \in [0, A]$, $y \in \mathbb{R}$, where $\alpha, \beta, \beta', \mu, \xi$ and ξ' are positive constants. Hence, for all $F_X \in \mathcal{A}^+$

$$\exp(-\alpha - \beta' y^2) \leq P_Y(y; F_X) \leq \exp(\alpha - \beta y^2), \tag{22}$$

$$\exp(-\mu - \xi' z^2) \leq P_Z(z; F_X) \leq \exp(\mu - \xi z^2). \tag{23}$$

Thus, we can write

$$|\log\left(P_Y(y; F_X)\right)| \leq \alpha + \beta' y^2, \tag{24}$$

$$|\log\left(P_Z(z; F_X)\right)| \leq \mu + \xi' z^2. \tag{25}$$

Next, we prove Theorem 1 using the preliminaries provided in this section.

B. Proof of Theorem 1

To prove Theorem 1, we first prove that the set of input distributions $\mathcal{A}^+$ satisfying (8), is compact and convex. We then show that the objective function in (9) is continuous, strictly concave and weakly differentiable in the input distribution F_X and hence we conclude that the solution to the optimization problem (9) exists and is unique. We continue the proof by deriving the necessary and sufficient conditions (KKT conditions) for the optimality of the optimal input distribution F_X^* and finally by means of contradiction we show that this optimal input distribution is discrete with a finite number of mass points.

Throughout the paper, we occasionally refer the reader to the technical report [8] where we have presented details that we can not provide here due to length constraint.

1) The feasible set $\mathcal{A}^+$ is compact and convex: The proof follows along similar lines as in [10, Appendix A.1].

2) $g_0(F_X)$ is continuous in input distribution F_X: It is established in [8, Section IV-B-2] that $g_0(F_X)$ is continuous in F_X.

3) $g_0(F_X)$ is strictly concave in F_X: The proof is given in [8, Section IV-B-Lemma 1].

4) $g_0(F_X)$ is weakly differentiable: We provide the proof in [8, Section IV-B-4].

Since the feasible set $\mathcal{A}^+$ is compact and convex and the objective function $g_0(F_X)$ is continuous, strictly concave and weakly differentiable, steps analogous to [5, Corollary 1] yield the following necessary and sufficient conditions for the optimality of the distribution F_X^*

$$r_e(x; F_X^*) \leq C_S, \;\; \forall x \in [0, A] \tag{26}$$

$$r_e(x; F_X^*) = C_S, \;\; \forall x \in S_{F_X^*} \tag{27}$$

where $S_{F_X^*}$ is the support set of F_X^* and the secrecy capacity C_S is expressed as

$$\begin{aligned} C_S &= I_B(F_X^*) - I_E(F_X^*) = \\ &h_Y(F_X^*) - h_Z(F_X^*) + \frac{1}{2}\mathbb{E}_{F_X^*}\left[\log\left(\frac{\sigma_{E,X}^2(x)}{\sigma_{B,X}^2(x)}\right)\right], \end{aligned} \tag{28}$$

where $I_B(F_X^*)$ and $I_E(F_X^*)$ are the mutual information for Bob and Eve, respectively, generated by the optimal input distribution F_X^*. Similarly, $h_Y(F_X^*)$ and $h_Z(F_X^*)$ are the differential entropies of Y and Z, respectively, generated by the input distribution F_X^*. Moreover, $\mathbb{E}_{F_X^*}$ denotes the expectation operator with respect to optimal distribution F_X^*.

We now prove by contradiction that the secrecy-capacity-achieving input distribution F_X^* has a finite number of mass points. To reach a contradiction, we use the KKT conditions in (26) and (27). To this end, we first show that both $i_B(x; F_X)$ and $i_E(x; F_X)$ have analytic extensions over some open connected set $\mathcal{D} = \left\{w : \Re(w) > -1/\eta_B^2\right\}$ in the complex plane $\mathbb{C}$ that includes the positive real line $\mathbb{R}_0^+$, where $\Re(\cdot)$ denote the real part of a complex variable.

5) The rate-equivocation density $r_e(x; F_X)$ is an analytic function on $\mathcal{D}$: Due to space limitations, we present the proof in [8, Section IV-B-5].

6) The secrecy-capacity-achieving input distribution is discrete: To prove the discreteness of the optimal input distribution F_X^*, we use a contradiction approach. To this end, let us assume that $S_{F_X^*}$ has an infinite number of elements. In view of the optimality condition (27), analyticity of $r_e(w; F_X)$ over the open connected set $\mathcal{D}$ and the identity theorem of complex analysis along with the Bolzano-Weierstrass Theorem, if $S_{F_X^*}$ has an infinite number of mass points, we get $r_e(w; F_X^*) = C_S$ for all $w \in \mathcal{D}$, which results in

$$r_e(x; F_X^*) = C_S, \quad \forall x \in \left(-1/\eta_B^2, +\infty\right). \tag{29}$$

Next, we show that (29) results in a contradiction. By observing the bounds given in (20)–(25), one can easily show that

$$\int_{\mathbb{R}} \exp\left(-\alpha-\beta' y^2\right)\left[-\alpha-\beta' y^2\right] dy \le \int_{\mathbb{R}} p(y|x)\times \log\left(P_Y(y; F_X^*)\right) dy \le \int_{\mathbb{R}} \exp\left(\alpha-\beta y^2\right)\left[\alpha+\beta' y^2\right] dy, \tag{30}$$

for all $x \in \left(-1/\eta_B^2,\, A\right) \subset \left(-1/\eta_B^2,\, +\infty\right)$. Similarly,

$$\int_{\mathbb{R}} \exp\left(-\mu-\xi' z^2\right)\left[-\mu-\xi' z^2\right] dz \le \int_{\mathbb{R}} p(z|x)\times \log\left(P_Z(z; F_X^*)\right) dz \le \int_{\mathbb{R}} \exp\left(\mu-\xi z^2\right)\left[\mu+\xi' y^2\right] dz, \tag{31}$$

for all $x \in \left(-1/\eta_B^2,\, A\right)$. Therefore, we can write

$$L_B \le -\int_{\mathbb{R}} p(y|x) \log\left(P_Y(y; F_X^*)\right) dy + \int_{\mathbb{R}} p(z|x)\times \log\left(P_Z(y; F_X^*)\right) dz \le U_B, \tag{32}$$

where the lower bound L_B and the upper bound U_B are given respectively as

$$L_B = \int_{\mathbb{R}} \left[-\mu-\xi' z^2\right] \exp\left(-\mu-\xi' z^2\right) dz + \int_{\mathbb{R}} \left[-\alpha-\beta' y^2\right] \exp\left(\alpha-\beta y^2\right) dy, \tag{33}$$

$$U_B = \int_{\mathbb{R}} \left[\mu+\xi' z^2\right] \exp\left(\mu-\xi z^2\right) dz + \int_{\mathbb{R}} \left[\alpha+\beta' y^2\right] \exp\left(-\alpha-\beta' y^2\right) dy. \tag{34}$$

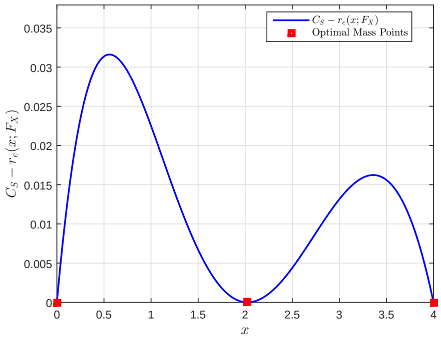


Fig. 1. Illustration of $C_S - r_e(x; F_X)$ yielded by the optimal input distribution when $\sigma_B^2 = 1$, $\sigma_E^2 = 2$, $\eta_B^2 = 0.25$, $\eta_E^2 = 0.125$ and $A = 4$.

We note that since the constants β, β', ξ and ξ' are all positive, L_B and U_B are finite values. Substituting (16) and (17) into (29) and using the bounds in (32), we can write

$$L_B \le C_S + \frac{1}{2} \log\left(\frac{\sigma_{B,X}^2(x)}{\sigma_{E,X}^2(x)}\right) \le U_B. \tag{35}$$

Now, let $\left\{x^{(n)}\right\}_{n=1}^{\infty}$ be a convergent sequence in $\mathbb{S} \triangleq \left(-1/\eta_B^2, A\right)$ with a limit point $x^{(0)} = -1/\eta_B^2$. It is clear that 1) $x^{(n)}$ and $\sigma_{B,X}^2\left(x^{(n)}\right)$ are real for all positive integers n, and 2) $\lim_{n\to\infty} \sigma_{B,X}^2\left(x^{(n)}\right) = 0$. Following the results in [9, Theorem 3] and using (35) we can write

$$\lim_{n\to\infty} (L_B - C_S) \le \lim_{n\to\infty} \frac{1}{2} \log\left(\frac{\sigma_{B,X}^2(x^{(n)})}{\sigma_{E,X}^2(x^{(n)})}\right) \le \lim_{n\to\infty} (U_B - C_S). \tag{36}$$

Since $\lim_{n\to\infty} \frac{1}{2} \log\left(\frac{\sigma_{B,X}^2(x^{(n)})}{\sigma_{E,X}^2(x^{(n)})}\right) = -\infty$ (due to the fact that $\sigma_{E,X}^2(x^{(0)})$ is a finite value) and the $\lim_{n\to\infty}(L_B - C_S)$ is a finite value, thus a contradiction occurs. This, in turn, implies that the support set $S_{F_X^*}$ cannot have an infinite number of elements and therefore the optimal input distribution F_X^* is discrete with a finite number of mass points.

V. Numerical Results

Fig. 1 provides a plot of the equivocation density for an optimal input distribution for $A = 4$, $\sigma_B^2 = 1$, $\sigma_E^2 = 2$, $\eta_B^2 = 0.25$, and $\eta_E^2 = 0.125$. We numerically found that for these parameters, the optimal input distribution is ternary with mass points located at $x = 0$, 2.025 and 4 with probability masses 0.2862, 0.3045 and 0.4093, respectively. We observe that $C_S - r_e(x; F_X)$ is generally nonnegative and is equal to zero at the optimal mass points; verifying the optimality conditions in (26) and (27).

Fig 2 illustrates the secrecy capacity C_S and the difference $C_B - C_E$ versus the peak-intensity constraint A, where C_B and C_E are the legitimate user's and the eavesdropper's capacities, respectively. We observe that this difference is in general a

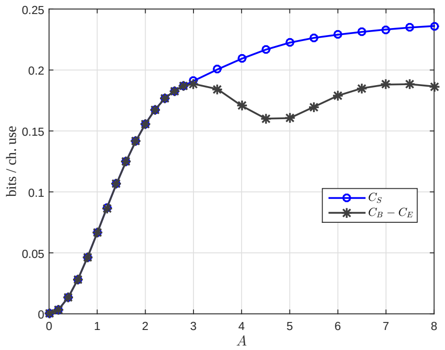


Fig. 2. The secrecy capacity for $\sigma_B^2 = 1$, $\sigma_E^2 = 2$, $\eta_B^2 = 0.25$ and $\eta_E^2 = 0.125$ versus the peak-intensity constraint A.

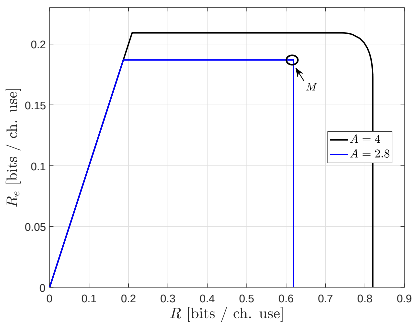


Fig. 3. The rate-equivocation region for $\sigma_B^2 = 1$, $\sigma_E^2 = 2$, $\eta_B^2 = 0.25$, and $\eta_E^2 = 0.125$ under peak-intensity constraints $A = 2.8$ and $A = 4$. Point M refers to the case when secrecy capacity and capacity are achieved simultaneously.

lower bound for the secrecy capacity C_S which can be easily proven. We also observe that, for small values of A, $C_B - C_E$ and C_S are identical. However, as A increases, $C_B - C_E$ and C_S become different. Similar to the secrecy capacity results of the Gaussian wiretap channel under a peak-power constraint provided in [4], here too, $I(X;Y)$ and $I(X;Z)$ are maximized by the same discrete distribution, however, $I(X;Y) - I(X;Z)$ is maximized by a different distribution. As a specific example, when $A = 4$, while both $I(X;Y)$ and $I(X;Z)$ are maximized by the same *binary* distribution with mass points at $x = 0$ and 4 with probability masses 0.5088 and 0.4912, respectively, $I(X;Y) - I(X;Z)$ is maximized by a *ternary* distribution with mass points at $x = 0$, 2.025 and 4 with probability masses 0.2862, 0.3045 and 0.4093, respectively. This explains the difference between C_S and $C_B - C_E$ at $A = 4$ in this figure.

Fig. 3 depicts the entire rate-equivocation region of the optical wiretap channel with input-dependent Gaussian noise under non-negativity and peak-intensity constraints when $\sigma_B^2 = 1$, $\sigma_E^2 = 2$, $\eta_B^2 = 0.25$, and $\eta_E^2 = 0.125$ for two different values of A. When $A = 2.8$, it is clear from the figure that both the secrecy capacity and the capacity can be attained simultaneously (Point "M" in the figure). In particular, for $A = 2.8$, the binary input distribution with mass points located at $x = 0$ and 2.8 with probabilities 0.5183 and 0.4817, respectively, achieves both the capacity and the secrecy capacity. This implies that, when $A = 2.8$, the transmitter can communicate with the legitimate user at the capacity while achieving the maximum equivocation at the eavesdropper. On the other hand, when $A = 4$, the secrecy capacity and the capacity cannot be achieved simultaneously (notice the curved shape in the figure). More specifically, for $A = 4$, the binary input distribution with mass points located at $x = 0$ and 4 with probabilities 0.5088 and 0.4912 achieves the capacity, while a ternary distribution with mass points located at $x = 0$, 2.025, 4 with probability masses 0.2862, 0.3045 and 0.4093, respectively, achieves the secrecy capacity, i.e., the optimal input distributions for the secrecy capacity and the capacity are different. In other words, there is a tradeoff between the rate and its equivocation.

VI. Conclusions

This paper studies the optical wiretap channel with input-dependent Gaussian noise under non-negativity and peak-intensity constraints. It is shown that the secrecy capacity and the boundary of the entire rate-equivocation region is achieved by discrete input distributions with a finite support. An interesting result that this paper reveals is that under such constraints, the secrecy capacity and the capacity of this optical wiretap channel cannot be obtained simultaneously in general, i.e., there is a tradeoff between the rate and its equivocation in the sense that, to increase the communication rate, one must compromise from the equivocation, and conversely to increase the achieved equivocation, one must compromise from the communication rate.

References

[1] C. E. Shannon, "Communication theory of secrecy systems," *j-BELL-SYST-TECH-J*, vol. 28, no. 4, pp. 656–715, 1949.

[2] A. D. Wyner, "The Wire-tap Channel," *Bell Systems Technical Journal*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[3] S. Leung-Yan-Cheong and M. Hellman, "The Gaussian wire-tap channel," *IEEE Transactions on Information Theory*, vol. 24, no. 4, pp. 451–456, Jul 1978.

[4] O. Ozel, E. Ekrem, and S. Ulukus, "Gaussian Wiretap Channel With Amplitude and Variance Constraints," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5553–5563, Oct 2015.

[5] J. G. Smith, "The Information Capacity of Amplitude- and Variance-Constrained Scalar Gaussian Channels," *Information and Control*, vol. 18, no. 3, pp. 203–219, April 1971.

[6] S. Arnon, J. Barry, G. Karagiannidis, R. Schober, and M. Uysal, *Advanced Optical Wireless Communication Systems*, 1st ed. New York, NY, USA: Cambridge University Press, 2012.

[7] S. M. Moser, "Capacity Results of an Optical Intensity Channel With Input-Dependent Gaussian Noise," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 207–223, Jan 2012.

[8] M. Soltani and Z. Rezki, "Optical Wiretap Channel with Input-Dependent Gaussian Noise Under Peak and Average Intensity Constraints," *Technical Report*, Apr. 2017. [Online]. Available: https://sites.google.com/site/zouheirrezki/publications

[9] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-Achieving Probability Measure for Conditionally Gaussian Channels with Bounded Inputs," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2073–2088, June 2005.

[10] C. Luo., *Communication for Wideband Fading Channels: On Theory and Practice*. PhD Thesis, Massachusetts Institute of Technology, Feb. 2006.

State-Dependent Parallel Gaussian Channels With a State-Cognitive Helper

Michael Dikshtein*, Ruchen Duan[†], Yingbin Liang [‡], and Shlomo Shamai (Shitz)[§]
*Department of EE, Technion, Haifa 32000, Israel, michaeldic@campus.technion.ac.il
[†]Samsung Semiconductor Inc, San Diego, CA 92121 USA, r.duan@samsung.com
[‡]Department of ECE, The Ohio State University, Columbus, OH 43210 USA, liang.889@osu.edu
[§]Department of EE, Technion, Haifa 32000, Israel, sshlomo@ee.technion.ac.il

***Abstract*—The state-dependent parallel channel with differently scaled states and a common state-cognitive helper is studied, in which two transmitters wish to send two messages to their corresponding receivers respectively over two parallel Gaussian subchannels. The two Gaussian channels are corrupted by the same but differently scaled states. The state is not known to the transmitters nor to the receivers, but known to a helper noncausally, which assists the receivers to cancel the state. Differently from previous studies that characterized the capacity region only in the infinite state power regime and under independent state corruption at the two receivers, this paper investigates the case under arbitrary state power and with the same but differently scaled states. An inner bound on the capacity region is derived and is compared to an outer bound. Then the channel parameters are partitioned into various cases, and segments on the capacity region boundary are characterized for each case.**

***Index Terms*—Dirty paper coding, , Gelf'and-Pinsker scheme, noncausal channel state information, parallel channel.**

I. Introduction

With the development of cellular systems, to support more users and higher transmission rates, non-orthogonal multi-user access (NOMA) has been intensively investigated, where interference cancellation is the key issue for the non-orthogonal transmission. In this paper, we investigate a type of state-dependent channels with helper, in which the state is not known to either transmitters or receivers, but is noncausally known to a state-cognitive helper. This model captures interference cancelation in various practical scenarios. For example, users in a multi-cell systems may be interfered by a base station located in other cells. Such a base station, being as the source that causes the interference, clearly knows the information of the interference (modeled by state) and can serve as a helper to help to cancel the interference. Alternatively, that base station can also convey the interference information to other base stations via the back haul network so that other base stations can serve as helpers to cancel the interference. As a comparison, this type of state-dependent models differ from the original state-dependent channels studied in e.g., [1] and [2], in that the state-cognitive helper is not informed of the transmitters' messages, and hence its state cancellation strategies are necessarily independent from message encoding at the transmitters.

The basic state-dependent Gaussian channel with a helper was introduced by [3], in which the capacity in the infinite power regime was characterized and was shown to be achievable by lattice coding. The capacity under arbitrary state power was established for some special cases in [4]. As more models, some state-dependent MACs also fall into the type of state-dependent models with state-cognitive helpers. The state-dependent asymmetric multiple access channel (MAC) was studied in [5], in which an inner bound was derived using Gelfand-Pinsker coding for the state-cognitive user, and using the regular MAC scheme for the uninformed user. In [6], the MAC with two states and with each state is known at one transmitter was studied, and the lattice coding was used to derive achievable regions. In [7], this channel was further studied with an additional common message shared between the informed and the uninformed user. New lower and upper bounds were derived. In a recent work [8] on the state-dependent MAC, a new outer bound was derived which is tighter than the previous bounds. In [4], the state-dependent MAC with an additional helper was studied, and the partial/full capacity region was characterized under various channel parameters. In [9], the state-dependent multicast channel was introduced and capacity bounds were derived for the independent and differently scaled states scenarios. Moreover, some state-dependent relay channel models can also be viewed as an extension of the state-dependent channel with a helper, where the relay serves the role of the helper by knowing the state information. In [10], the state-dependent relay channel with state non-causally available at the relay is considered. An achievable rate was derived using a combination of decode-and-forward, Gelfand-Pinsker binning and codeword splitting. And in [11], additional noiseless cooperation links with finite capacity were assumed between the transmitter and the relay, and various coding techniques were explored.

The most relevant work to this paper is [12], in which the state-dependent parallel channel with a helper was studied, for the regime with infinite state power and with two receivers being corrupted by two independent states. A time-sharing scheme was proved to be capacity achieving under certain channel parameters. In contrast, in this paper, the two receivers of the parallel channel are corrupted by the same but differently scaled states, and the state can take arbitrary power. In this case, the time-sharing scheme is no longer optimal.

Thus, in this paper, we derive an inner bound on the capacity region using an achievability scheme that integrates single-bin dirty paper coding and direct state subtraction. We

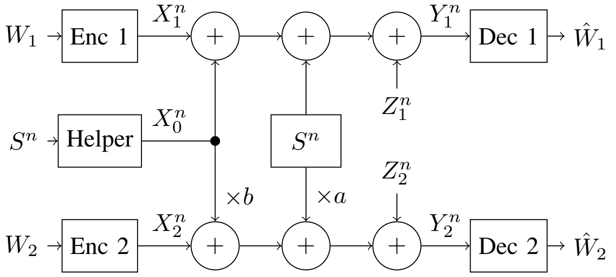


Fig. 1: The state-dependent parallel channel with a helper.

then compare such an inner bound with an outer bound that consists of the capacity of point-to-point channel without state and an outer bound developed in [3] for the point-to-point state-dependent channel with a helper. The comparison yields the capacity region for certain ranges of channel parameters. More specifically, when the helper's power is above a certain threshold (which is less than the state power and hence direct state cancellation cannot be applied) and the helper's signal is scaled the same as the state, we show that the state interference can be fully cancelled for both channels, and thus the capacity region is the same as that of the corresponding channel without state.

II. Channel Model

In this paper, we study the state-dependent parallel network with a state-cognitive helper, in which two transmitters communicate with two corresponding receivers over a state-dependent parallel channel. The two receivers are corrupted by two differently scaled states, respectively. The state information is not know to either the transmitters or the receivers, but to a helper noncausally. Hence, the helper assists these receivers to cancel the state interference (see Figure 1).

More specifically, the encoder at transmitter i, $f_i : M_i \to \mathcal{X}_i^n$, maps a message $m_i \in \{1, \ldots, 2^{nR_i}\}$ to a codeword x_i^n, for $i = 1, 2$. The inputs x_1^n and x_2^n are sent respectively over the two subchannels of the parallel channel. The two receivers are corrupted by an independent and identically distributed (i.i.d.) state sequence $s^n \in \mathcal{S}^n$, which is known to a common helper noncausally. Hence, the encoder at the helper, $f_0 : \mathcal{S}^n \to \mathcal{X}_0^n$, maps the state sequence $s^n \in \mathcal{S}^n$ into a codeword $x_0^n \in \mathcal{X}_0^n$. The channel transition probability is given by $P_{Y_1|X_0X_1S} \cdot P_{Y_2|X_0X_2S}$. The decoder at receiver i, $g_i : \mathcal{Y}_i^n \to \mathcal{M}_i$, maps a received sequence y_i^n into a message $\hat{m}_i \in \mathcal{M}_i$, for $i = 1, 2$. We assume that the messages are uniformly distributed over the sets $\mathcal{M}_1$ and $\mathcal{M}_2$. We define the average probability of error for a length-n code as follows:

$$P_e = \frac{1}{|\mathcal{M}_1||\mathcal{M}_2|} \sum_{m_1=1}^{\mathcal{M}_1} \sum_{m_2=1}^{\mathcal{M}_2} \mathbb{P}\left\{\hat{m}_1 \neq m_1, \hat{m}_2 \neq m_2\right\}. \quad (1)$$

Definition 1. A rate pair (R_1, R_2) is said to be *achievable* if there exist a sequence of message sets $\mathcal{M}_1^{(n)}$ and $\mathcal{M}_2^{(n)}$ with $\left|\mathcal{M}_1^{(n)}\right| = 2^{nR_1}$ and $\left|\mathcal{M}_2^{(n)}\right| = 2^{nR_2}$, and encoder-decoder tuples $\left(f_0^{(n)}, f_1^{(n)}, f_2^{(n)}, g_1^{(n)}, g_2^{(n)}\right)$ such that the average probability of error $P_e^{(n)} \to 0$ as $n \to \infty$.

Definition 2. We define the *capacity region* of the channel as the closure of the set of all achievable rate pairs (R_1, R_2).

In this paper, we focus on the Gaussian channel, with the outputs at the two receivers for one channel use given by

$$Y_1 = X_0 + X_1 + S + Z_1 \quad (2a)$$

$$Y_2 = bX_0 + X_2 + aS + Z_2 \quad (2b)$$

where Z_1 and Z_2 are noise variables with Gaussian distributions $Z_1 \sim \mathcal{N}(0, 1)$ and $Z_2 \sim \mathcal{N}(0, 1)$, and S is the state variable with Gaussian distribution $S \sim \mathcal{N}(0, Q)$. Both the noise variables and the state variable are i.i.d. over channel uses. The channel inputs X_0, X_1, and X_2 are subject to the average power constraints $\frac{1}{n}\sum_{i=1}^{n} X_{ki}^2 \leq P_k$, $k \in \{0, 1, 2\}$. The constant a represents the channel gain of the state sequence in the second subchannel compared to the first subchannel. Similarly, the constant b is the gain of the helper signal in the second subchannel compared to that in the first subchannel. Thus our model presents a general scenario, where the helper's power and the state power can be arbitrary.

Our goal is to characterize the capacity region of the Gaussian channel under various channel parameters (a, b, P_0, P_1, P_2, Q).

III. Main Results

In this section, we first derive inner and outer bounds on the capacity region for the state-dependent parallel channel with a helper. Then by comparing the inner and outer bounds, we characterize the segments on the capacity region boundary under various different channel parameters. Some proofs are omitted due to space limitations.

A. Inner and Outer Bounds

We start by deriving an inner bound on the capacity region for the discrete memoryless channel based on single bin Gel'fand-Pinsker binning scheme.

Proposition 1. *For the discrete memoryless state-dependent parallel channel with a helper under the same but differently scaled states at the two receivers, an inner bound on the capacity region consists of rate pairs (R_1, R_2) satisfying:*

$$R_1 \leq \min\left\{I(U, X_1; Y_1) - I(U; S), I(X_1; Y_1|U)\right\} \quad (3a)$$

$$R_2 \leq \min\left\{I(U, X_2; Y_2) - I(U; S), I(X_2; Y_2|U)\right\} \quad (3b)$$

for some distribution $P_{U|S}P_{X_0|US}P_{X_1}P_{X_2}$.

Proof: Fix the conditional pmf $P_{X_0U|S}P_{X_1}P_{X_2}$. Generate $2^{n\tilde{R}}$ sequences $u^n(v)$ based on P_U and 2^{nR_k} codewords $x_k^n(w_k)$ based on P_{X_k}, where $k \in \{1, 2\}$

Given s^n, the encoder at the helper finds $\tilde{v}$, such that $(u^n(\tilde{v}), s^n) \in T_{\epsilon'}^n$. It can be shown that for large n, such $\tilde{v}$ exists with high probability if $\tilde{R} \geq I(U; S)$. Then given

$(u^n(\tilde{v}), s^n)$, generate x_0^n with i.i.d. components based on $P_{X_0|SU}$ for transmission.

Given w_k, the encoder at transmitter k transmits $x_k^n(w_k)$.

Given y_k^n, the decoder at receiver k finds $(\hat{v}, \hat{w_k})$ such that $(u^n(\hat{v}), x_k^n(w_k), y_k^n) \in T_\epsilon^n$. If no or more than one $\hat{w}_k$ can be found, error is declared.

It can be shown that for sufficiently large n, decoding is correct with high probability if

$$\begin{aligned} R_k &\leq I(X_k; Y_k|U) \\ \tilde{R} + R_k &\leq I(U, X_k; Y_k) \end{aligned}$$

This completes the proof. ■

Based on the above inner bound for the discrete memoryless case, we derive the following inner bound for the Gaussian channel.

Proposition 2. *An inner bound on the capacity region for the state-dependent parallel Gaussian channel with a helper and under the same but differently scaled state consists of rate pairs (R_1, R_2) satisfying:*

$$R_1 \leq \min\left\{f_{1,1}(\alpha,\beta,P_1), g_{1,1}(\alpha,\beta,P_1)\right\} \tag{4a}$$

$$R_2 \leq \min\left\{f_{a,b}(\alpha,\beta,P_2), g_{a,b}(\alpha,\beta,P_2)\right\} \tag{4b}$$

where, α and β are real constants satisfying $|\beta| \leq \sqrt{P_0/Q}$, and

$$f_{a,b}(\alpha,\beta,P) = \frac{1}{2}\log\frac{P_0'\left(b^2P_0' + (a+b\beta)^2Q + P + 1\right)}{P_0'Q(b\alpha - a - b\beta)^2 + P_0' + \alpha^2 Q}, \tag{5a}$$

$$g_{a,b}(\alpha,\beta,P) = \frac{1}{2}\log\left(1 + \frac{P\left(P_0' + \alpha^2 Q\right)}{P_0'Q(b\alpha - a - b\beta)^2 + P_0' + \alpha^2 Q}\right), \tag{5b}$$

where $P_0' = P_0 - \beta^2 Q$.

Proof: The proof follows from Proposition 1 by choosing the following joint Gaussian distribution for the random variables:

$$\begin{aligned} &X_0 = X_0' + \beta S, \quad U = X_0' + \alpha S \\ &X_0' \sim \mathcal{N}(0, P_0'), \quad X_1 \sim \mathcal{N}(0, P_1) \quad X_2 \sim \mathcal{N}(0, P_2) \end{aligned}$$

where X_0', S, X_1, X_2 are independent. P_0' is chosen such that the power constraint on X_0 is satisfied with equality

$$P_0 \geq \mathbb{E}X_0^2 = P_0' + \beta^2 Q.$$

■

We note that the above choice of the helper's signal incorporates two parts with X_0' designed using single-bin dirty paper coding, and βS acting as direct state subtraction.

We next present an outer bound which applies the point-to-point channel capacity and the upper bound derived for the point-to-point channel with a helper in [3].

Lemma 1. *An outer bound on the capacity region of the states-dependent parallel Gaussian channel with a helper consists of rate pairs (R_1, R_2) satisfying:*

$$\begin{aligned} R_1 \leq \min\bigg\{&\frac{1}{2}\log\left(1 + \frac{P_1}{P_0 + 2\rho_{0S}\sqrt{P_0 Q} + Q + 1}\right) \\ &+\frac{1}{2}\log\left((1-\rho_{0S}^2)P_0 + 1\right), \frac{1}{2}\log\left(1 + P_1\right)\bigg\} \end{aligned} \tag{6a}$$

$$\begin{aligned} R_2 \leq \min\bigg\{&\frac{1}{2}\log\left(1 + \frac{P_2}{b^2P_0 + 2ab\rho_{0S}\sqrt{P_0 Q} + a^2 Q + 1}\right) \\ &+\frac{1}{2}\log\left((1-\rho_{0S}^2)b^2P_0 + 1\right), \frac{1}{2}\log\left(1 + P_2\right)\bigg\} \end{aligned} \tag{6b}$$

for some ρ_{0S} that satisfies $-1 \leq \rho_{0S} \leq 1$.

B. Capacity Region Characterization

In this section, we optimize α and β in Proposition 2, and compare the rate bounds with the outer bounds in Lemma 1 to characterize the points or segments on the capacity region boundary.

We first define $\phi_{a,b}(\rho_{0S}, P)$ and $\theta_{a,b}(\rho_{0S}, P)$ as in (7) for notational convenience.

Since the inner bound in Proposition 2 is not convex, it is difficult to provide a close form for the jointly optimized bounds. Therefore, we first optimize the bounds for R_1 and R_2 respectively, and then provide conditions on channel parameters such that these bounds match the outer bound. Based on the conditions, we partition the channel parameters into the sets, in which different segments of the capacity region boundary can be obtained.

We first consider the rate bound for R_1 in (4a). By setting

$$\alpha_1 \triangleq \frac{(1+\beta_1)P_0'}{P_0'+1}, \quad \beta_1 \triangleq \rho_{0S}^*\sqrt{\frac{P_0}{Q}}$$

$f_{1,1}(\alpha,\beta,P_1)$ and $g_{1,1}(\alpha,\beta,P_1)$ take the following form

$$\begin{aligned} f_{1,1}(\alpha_1,\beta_1,P_1) &= \phi_{1,1}(\rho_{0S}^*, P_1) \\ g_{1,1}(\alpha_1,\beta_1,P_1) &= \theta_{1,1}(\rho_{0S}^*, P_1) \end{aligned}$$

where $\rho_{0S}^* \in [-1, 1]$ maximizes $\phi_{1,1}(\rho_{0S}, P_1)$. In fact, α_1 maximizes $f_{1,1}(\alpha,\beta,P_1)$ for fixed β, and β_1 maximizes the function with $\alpha = \alpha_1$.

If $\phi_{1,1}(\rho_{0S}^*, P_1) \leq \theta_{1,1}(\rho_{0S}^*, P_1)$, $R_1 = \phi_{1,1}(\rho_{0S}^*, P_1)$ is achievable, and this matches the upper bound in (6a). Thus, one segment of the capacity region is specified by

$$R_1 = \phi_{1,1}(\rho_{0S}^*, P_1) \tag{8a}$$

$$R_2 \leq \min\{f_{a,b}(\alpha_1,\beta_1,P_2), g_{a,b}(\alpha_1,\beta_1,P_2)\} \tag{8b}$$

We further observe that the second term $g_{1.1}(\alpha,\beta,P_1)$ in (4a) is optimized by setting $\alpha = 1+\beta$, and hence

$$g_{1,1}(\alpha, \alpha-1, P_1) = 0.5\log(1+P_1).$$

If $g_{1,1}(\alpha,\alpha-1,P_1) \leq f_{1,1}(\alpha,\alpha-1,P_1)$, i.e.,

$$P_0'^2 \geq \alpha^2 Q(P_1 + 1 - P_0'), \tag{9}$$

$$\phi_{a,b}(\rho_{0S}, P) = \frac{1}{2}\log\left(1 + \frac{P}{b^2P_0 + 2ab\rho_{0S}\sqrt{P_0Q} + a^2Q + 1}\right) + \frac{1}{2}\log\left((1-\rho_{0S}^2)b^2P_0 + 1\right) \tag{7a}$$

$$\theta_{a,b}(\rho_{0S}, P) = \frac{1}{2}\log\left(1 + \frac{P\left((1 + b^2P_0(1-\rho_{0S}^2))^2 + (1-\rho_{0S}^2)b^2P_0(a\sqrt{Q} + b\rho_{0S}\sqrt{P_0})^2\right)}{\left(a^2Q + 2ab\rho_{0S}\sqrt{P_0Q} + b^2P_0 + 1\right)\left(b^2(1-\rho_{0S}^2)P_0 + `1\right)}\right) \tag{7b}$$

then the inner bound for R_1 becomes $R_1 = 0.5\log(1+P_1)$, which is the capacity of the point-to-point channel without state and matches the outer bound in (6a). Thus one segment of the capacity is specified by

$$R_1 = 0.5\log(1+P_1) \tag{10a}$$

$$R_2 \leq \min\{f_{a,b}(\alpha, \alpha-1, P_2), g_{a,b}(\alpha, \alpha-1, P_2)\}. \tag{10b}$$

We then consider the rate bound for R_2. Similarly, the following segments on the capacity boundary can be obtained. If $\phi_{a,b}(\rho_{0S}^{**}, P_2) \leq \theta_{a,b}(\rho_{0S}^{**}, P_2)$, one segment of the capacity region boundary is specified by

$$R_1 \leq \min\{f_{1,1}(\alpha_2, \beta_2, P_1), g_{1,1}(\alpha_2, \beta_2, P_1)\} \tag{11a}$$

$$R_2 = \phi_{a,b}(\rho_{0S}^{**}, P_2) \tag{11b}$$

where

$$\alpha_2 \triangleq \frac{(a+b\beta_2)bP_0'}{b^2P_0'+1}, \quad \beta_2 \triangleq \rho_{0S}^{**}\sqrt{\frac{P_0}{Q}}$$

and $\rho_{0S}^{**} \in [-1, 1]$ maximizes $\phi_{a,b}(\rho_{0S}, P_2)$.

Furthermore, if $g_{a,b}(\alpha, \alpha - a/b, P_2) \leq f_{a,b}(\alpha, \alpha - a/b, P_2)$, one segment of the capacity region boundary is specified by

$$R_1 \leq \min\left\{f_{1,1}\left(\alpha, \alpha - \frac{a}{b}, P_1\right), g_{1,1}\left(\alpha, \alpha - \frac{a}{b}, P_1\right)\right\} \tag{12a}$$

$$R_2 = 0.5\log(1+P_2). \tag{12b}$$

Summarizing the above analysis, we obtain the following characterization of segments of the capacity region boundary.

Theorem 1. *The channel parameters* (a, b, P_0, P_1, P_2, Q) *can be partitioned into the sets* $\mathcal{A}_1, \mathcal{B}_1, \mathcal{C}_1$*, where*

$$\begin{aligned}
\mathcal{A}_1 &= \{(a, b, P_0, P_1, P_2, Q) : \phi_{1,1}(\rho_{0S}^*, P_1) \leq \theta_{1,1}(\rho_{0S}^*, P_1)\} \\
\mathcal{C}_1 &= \{(a, b, P_0, P_1, P_2, Q) : P_0'^2 \geq \alpha^2 Q(P_1 + 1 - P_0') \\
&\qquad \textit{where } P_0' = P_0 - (\alpha-1)^2Q, \textit{ for some } \alpha \in \Omega_\alpha\} \\
\mathcal{B}_1 &= (\mathcal{A}_1 \cup \mathcal{C}_1)^c.
\end{aligned}$$

If $(a, b, P_0, P_1, P_2, Q) \in \mathcal{A}_1$*, then* (8a) − (8b) *captures one segment of the capacity region boundary, where the state cannot be fully cancelled. If* $(a, b, P_0, P_1, P_2, Q) \in \mathcal{C}_1$*, then* (10a) − (10b) *captures one segment of the capacity region boundary where the state is fully cancelled. If* $(a, b, P_0, P_1, P_2, Q) \in \mathcal{B}_1$*, then the* R_1 *segment of the capacity region boundary is not characterized.*

The channel parameters (a, b, P_0, P_1, P_2, Q) *can also be partitioned into the sets* $\mathcal{A}_2, \mathcal{B}_2, \mathcal{C}_2$*, where*

$$\begin{aligned}
\mathcal{A}_2 &= \{(a, b, P_0, P_1, P_2, Q) : \phi_{a,b}(\rho_{0S}^{**}, P_2) \leq \theta_{a,b}(\rho_{0S}^{**}, P_2)\} \\
\mathcal{C}_2 &= \{(a, b, P_0, P_1, P_2, Q) : b^2P_0'^2 \geq \alpha^2 Q(P_2 + 1 - b^2P_0') \\
&\qquad \textit{where } P_0' = P_0 - (\alpha - a/b)^2 Q, \textit{ for some } \alpha \in \Omega_\alpha\} \\
\mathcal{B}_2 &= (\mathcal{A}_2 \cup \mathcal{C}_2)^c.
\end{aligned}$$

If $(a, b, P_0, P_1, P_2, Q) \in \mathcal{A}_2$*, then* (11a) − (11b) *captures one segment of the capacity region boundary, where the state cannot be fully cancelled. If* $(a, b, P_0, P_1, P_2, Q) \in \mathcal{C}_2$*, then* (12a) − (12b) *captures one segment of the capacity boundary where the state is fully cancelled. If* $(a, b, P_0, P_1, P_2, Q) \in \mathcal{B}_2$*, then the* R_2 *segment of the capacity region boundary is not characterized.*

The above theorem describes two partitions of the channel parameters, respectively under which segments on the capacity region boundary corresponding to R_1 and R_2 can be characterized. Intersection of two sets, each from one partition, collectively characterizes the entire segments on the capacity region boundary.

Figure 2 lists all possible intersection of sets that the channel parameters can belong to. For each case in Figure 2, we use red solid line to represent the segments on the capacity region that are characterized in Theorem 1, and we also mark the value of the capacity that each segment corresponds to as characterized in Theorem 1.

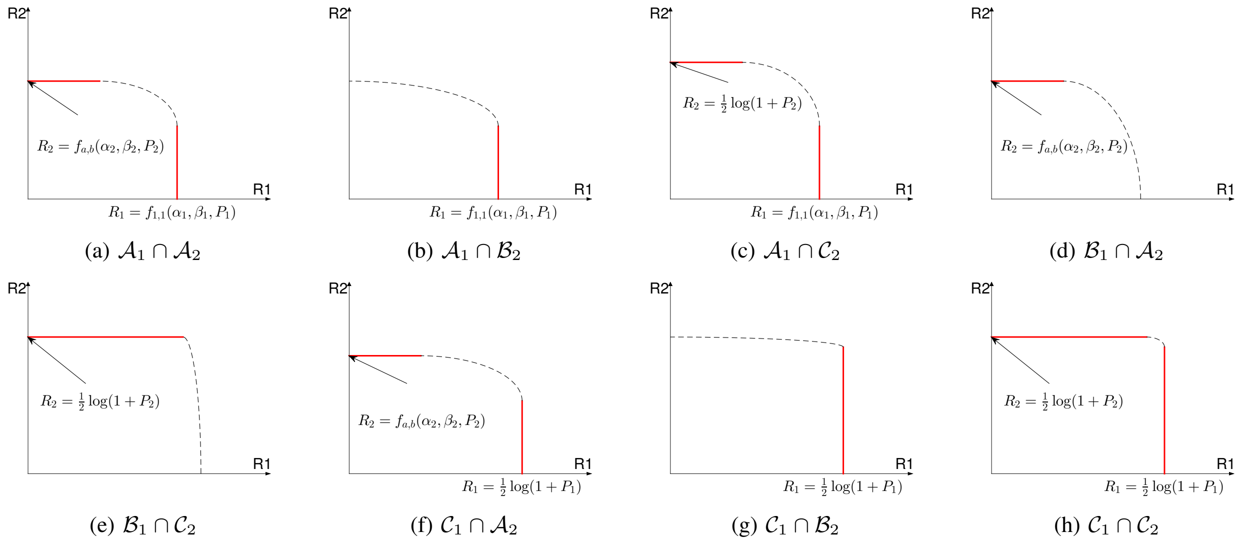
One interesting example in Theorem 1 is the case with $a = b$, in which R_1 and R_2 are optimized with the same set of coefficients α and β when $P_0'^2 \geq \alpha^2 Q(P_1 + 1 - P_0')$ and $a^2P_0'^2 \geq \alpha^2 Q(P_2 + 1 - a^2P_0')$. Thus, the point-to-point channel capacity is obtained for both R_1 and R_2, with state being fully cancelled. We state this result in the following theorem.

Theorem 2. *If* $a = b$, $P_0'^2 \geq \alpha^2 Q(P_1 + 1 - P_0')$ *and* $a^2P_0'^2 \geq \alpha^2 Q(P_2 + 1 - a^2P_0')$ *where* $P_0' = P_0 - (\alpha - 1)^2 Q$*, for some* $\alpha \in \Omega_\alpha$ *then the capacity region of the state-dependent parallel Gaussian channel with a helper and under the same but differently scaled states contains* (R_1, R_2) *satisfying*

$$\begin{aligned}
R_1 &\leq 0.5\log(1+P_1) \\
R_2 &\leq 0.5\log(1+P_2).
\end{aligned}$$

C. Numerical Example

We now examine our results via simulations. We set $P_0 = 6$, $P_1 = P_2 = 5$, $Q = 12$, and $b = 0.8$, and plot the inner and outer bounds for the capacity region (R_1, R_2) for two values of a. It can be observed from Figure 3 that the upper


R2
R1
$R_2 = f_{a,b}(\alpha_2, \beta_2, P_2)$
$R_1 = f_{1,1}(\alpha_1, \beta_1, P_1)$
$R_2 = \frac{1}{2}\log(1+P_2)$
$R_1 = \frac{1}{2}\log(1+P_1)$

(a) $\mathcal{A}_1 \cap \mathcal{A}_2$ (b) $\mathcal{A}_1 \cap \mathcal{B}_2$ (c) $\mathcal{A}_1 \cap \mathcal{C}_2$ (d) $\mathcal{B}_1 \cap \mathcal{A}_2$

(e) $\mathcal{B}_1 \cap \mathcal{C}_2$ (f) $\mathcal{C}_1 \cap \mathcal{A}_2$ (g) $\mathcal{C}_1 \cap \mathcal{B}_2$ (h) $\mathcal{C}_1 \cap \mathcal{C}_2$

Fig. 2: Segments of the capacity region for all cases of channel parameters.

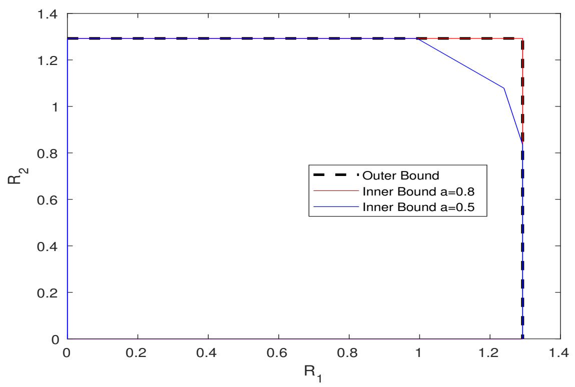


Fig. 3: Capacity bounds for channel parameters $P_0 = 6$, $P_1 = P_2 = 5$, $Q = 12$, $b = 0.8$ and various state gain a.

bound is defined by the rectangular region of channel without state. The inner bound, in the contrary, is susceptible to the value of a, such that in the case where $a = b$, our inner and outer bounds coincide everywhere, while in the case $a \neq b$ they coincide only on some segments. Both observations corroborate the characterization of the capacity in Theorem 1.

IV. Conclusion

In this paper, we have studied the parallel state-dependent Gaussian channel with a state-cognitive helper and with the same but differently scaled states. An inner bound was derived and was compared to an upper bound, and the segments of the capacity region boundary were characterized for various channel parameters. Furthermore, if the helper's signal and the state are equally scaled, the full rectangular capacity region of the two point-to-point channels without state can be achieved. As future work, we will analyze the case with channels being corrupted by independent states, and characterize the capacity region for various channel parameters.

Acknowledgment

The work of M. Dikshtein and S. Shamai was supported by the European Union's Horizon 2020 Research And Innovation Programme with grant agreement no. 694630, and by the Heron consortium via the Israel minister of economy and science. The work of Y. Liang was supported by NSF Grant CCF-1618127.

References

[1] S. Gel'fand and M. Pinsker. Coding for channels with ramdom parameters. *Probl. Contr. Inf. Theory*, 9(1):19–31, January 1980.

[2] M. H. M. Costa. Writing on dirty paper. *IEEE Trans. Inf. Theory*, 29(3):439–441, May 1983.

[3] S. Mallik and R. Koetter. Helpers for cleaning dirty papers. In *7th International ITG Conference on Source and Channel Coding*, pages 1–5, Jan 2008.

[4] Y. Sun, R. Duan, Y. Liang, A. Khisti, and S. Shamai (Shitz). Capacity characterization for state-dependent gaussian channel with a helper. *IEEE Trans. Inf. Theory*, 62(12):7123–7134, Dec 2016.

[5] S. P. Kotagiri and J. N. Laneman. Achievable rates for multiple access channels with state information known at one encoder. In *Proc. Allerton Conf. Communications, Control, and Computing*, 2004.

[6] T. Philosof, A. Khisti, U. Erez, and R. Zamir. Lattice strategies for the dirty multiple access channel. In *2007 IEEE ISIT*, pages 386–390, June 2007.

[7] A. Zaidi, S. P. Kotagiri, J. N. Laneman, and L. Vandendorpe. Multiaccess channels with state known to one encoder: Another case of degraded message sets. In *2009 IEEE ISIT*, pages 2376–2380, June 2009.

[8] W. Yang, Y. Liang, S. S. Shitz, and H. V. Poor. Outer bounds for gaussian multiple access channels with state known at one encoder. In *2017 IEEE ISIT*, pages 869–873, June 2017.

[9] A. Khisti, U. Erez, A. Lapidoth, and G. W. Wornell. Carbon copying onto dirty paper. *IEEE Trans. Inf. Theory*, 53(5):1814–1827, May 2007.

[10] A. Zaidi, S. P. Kotagiri, J. N. Laneman, and L. Vandendorpe. Cooperative relaying with state available noncausally at the relay. *IEEE Trans. Inf. Theory*, 56(5):2272–2298, May 2010.

[11] M. Li, O. Simeone, and A. Yener. Message and state cooperation in a relay channel when only the relay knows the state. *CoRR*, abs/1102.0768, 2011.

[12] R. Duan, Y. Liang, A. Khisti, and S. Shamai (Shitz). State-dependent parallel gaussian networks with a common state-cognitive helper. *IEEE Trans. Inf. Theory*, 61(12):6680–6699, Dec 2015.

Capacity of a Dual Enrollment System with Two Keys Based on an SRAM-PUF

Lieneke Kusters and Frans M.J. Willems
Eindhoven University of Technology, Eindhoven, The Netherlands. Contact: c.j.kusters@tue.nl

***Abstract*—We investigate the capacity of an SRAM-PUF based secrecy system that produces two secret keys during two consecutive enrollments. We determined the region of secret-key rates that are achievable and show that the total secret-key capacity is larger than for a single enrollment system. In our achievability proofs we focussed on linear codes.**

I. Introduction

An SRAM-PUF has a binary response that is unpredictable but reliable, and that is unique to the specific SRAM. Therefore, SRAM-PUF observation vectors are considered as a digital fingerprint, and may be used to generate and reconstruct secret keys [1]. Such secret keys can be used to authenticate a device or to secure data. It is important that the secret key is hard to guess by an attacker, and at the same time perfectly reconstructable by the user who can observe the SRAM-PUF.

The problem of (re-)generating a secret key from SRAM-PUF observations can be directly mapped to the problem of secret-key agreement [2]. In this case, an encoder and a decoder observe dependent sequences with some known joint distribution and need to agree on a secret key. It is known that the maximum achievable secret-key rate for this scenario is equal to the mutual information between the observed random variables by the encoder and decoder respectively [3], [4]. This rate can be achieved by one-way communication between the encoder and the decoder. The encoder generates a secret key and corresponding helper message, based on its input. We refer to this process as enrollment. The helper message is send over a public channel to the decoder. The decoder uses the helper message and his own observation to reconstruct the same key.

Clearly, the mutual information and thus the achievable secret-key rate may increase when more observations are considered. However, can we also increase the secret-key rate when additional input is observed by the encoder after an enrollment has already been completed? In [5], we have studied the case when enrollment of the same SRAM-PUF is repeated multiple times. The encoder regenerates the secret key and a corresponding helper message after observing an additional input from the SRAM-PUF. We have shown that given certain symmetry properties of the SRAM-PUF no leakage results from the additional helper messages. However, all previous keys are considered invalid and the secret-key rate remains the same. In the current work, we allow the encoder to generate a second key and corresponding helper message after observing an additional response from the SRAM-PUF. We show that the secret-key rate can be sequentially increased in this case while ensuring no leakage about any of the keys.

In the following, we first introduce the SRAM-PUF model and the notation that is used in the paper. Then, we analyze the 1-enrollment scheme, where a single key and helper message are generated by the encoder. We give the converse and show achievability of the secret-key rates with linear codes for the 1-enrollment scheme. Then, we continue to the 2-enrollment scheme, where a second key and helper message are generated, and we derive the achievable rates for this scheme.

II. Notation and SRAM-PUF statistical model

We use capitals to refer to random variables and lowercase symbols for realizations of random variables. All vectors in this paper are printed in bold and are binary. The SRAM-PUF observation vector has lenght n and corresponds to the values of the n cells in an SRAM cell array. We assume that the values of the SRAM-PUF cells are independent of each other and identically distributed. Moreover the observations of an SRAM cell are permutation invariant, hence for three consecutive observations x, y, and z of an SRAM cell, $p(x, y, z) = p(x, z, y) = \cdots = p(z, y, x)$. This leads to $H(X) = H(Y) = H(Z)$ and $H(XY) = H(YZ) = H(ZX)$. We focus here on three subsequent observation vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, of the same SRAM-PUF. Then e.g. $H(\boldsymbol{X}) = nH(X)$ etc. by the fact that observation vectors are i.i.d. . Given that the SRAM-PUF is permutation invariant we obtain

$$H(\boldsymbol{X}) = H(\boldsymbol{Y}) = H(\boldsymbol{Z}),$$
$$H(\boldsymbol{XY}) = H(\boldsymbol{XZ}) = H(\boldsymbol{YZ}).$$

III. 1-enrollment setting

In the 1-enrollment setting shown in Figure 1, an encoder constructs a secret key k and helper message m after observing observation vector $\boldsymbol{x}$. A decoder observing observation vector $\boldsymbol{z}$, should be able to reconstruct k when given helper message m. Furthermore, the helper message m should not reveal any information about k to an attacker who can not observe $\boldsymbol{x}$ nor $\boldsymbol{z}$. The secret key assumes values in $\{1, 2, \cdots, |\mathcal{K}|\}$ where $|\mathcal{K}| \leq 2^n$ since $\boldsymbol{x} \in \{0, 1\}^n$.

Definition 1: A secret-key rate R is called achievable in the 1-enrollment setting, if for all $\delta > 0$ and for all n large enough, there exist encoders and decoders such that

$$\Pr(\widehat{K} \neq K) \leq \delta,$$

This work was funded by the Eurostars-2 joint programme with co-funding from the EU Horizon 2020 programme under the E! 9629 PATRIOT project.

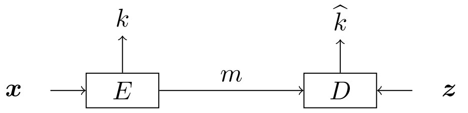


Fig. 1: Single enrollment scenario.

$$\frac{1}{n}H(K)+\delta \geq \frac{1}{n}\log_2|\mathcal{K}| \geq R-\delta,$$
$$\frac{1}{n}I(K;M) < \delta.$$

The secret-key capacity is defined as the maximum achievable rate $C \triangleq \max R$.

Theorem 1: Achievable secret-key rates for the 1-enrollment setting satisfy

$$R \leq I(X;Z),$$

and the secret-key capacity $C = I(X;Z)$.

The 1-enrollment setting is the same as the secret-key generation scenario studied by Ahlswede and Csiszar [4] and Maurer [3].

A proof of achievability and the converse of Theorem 1 can be found e.g. in [6]. There, a random coding argument was used to show that codes exist that can achieve the secret-key capacity C. Here, we will extend their argument by showing that the secret-key capacity can be achieved using linear codes.

A. Converse

First, we derive an upper bound for the achievable secret-key rate. We use Fano's inequality to obtain

$$\begin{aligned} H(K|\boldsymbol{Z}M) &= H(K|\boldsymbol{Z}M\widehat{K}) \\ &\leq H(K|\widehat{K}) \\ &\leq 1 + P_{ek}\log_2|\mathcal{K}| \leq 1 + nP_{ek}, \end{aligned}$$

where $P_{ek} = \Pr(\widehat{K} \neq K)$. For achievable rates we get

$$\begin{aligned} H(K) &= I(K;\boldsymbol{Z}M) + H(K|\boldsymbol{Z}M) \\ &\leq I(K;M) + I(K;\boldsymbol{Z}|M) + 1 + nP_{ek} \\ &\leq I(K;M) + H(\boldsymbol{Z}) - H(\boldsymbol{Z}|MK\boldsymbol{X}) + 1 + nP_{ek} \\ &= I(K;M) + H(\boldsymbol{Z}) - H(\boldsymbol{Z}|\boldsymbol{X}) + 1 + nP_{ek} \\ &\leq n\delta + nI(X;Z) + 1 + n\delta. \end{aligned}$$

This leads to

$$R - \delta \leq \frac{1}{n}H(K) + \delta \leq I(X;Z) + 3\delta + \frac{1}{n}.$$

For $n \to \infty$ and $\delta \downarrow 0$ we obtain that achievable rates R must satisfy $R \leq I(X;Z)$.

B. Linear codes for secret-key capacity

Next, we show that linear codes exist that achieve the secret-key capacity $I(X;Z)$ for the 1-enrollment scheme.

Fix an $\epsilon > 0$. Now $\mathcal{A}_\epsilon^{(n)}(X)$ and $\mathcal{A}_\epsilon^{(n)}(XZ)$ are the sets of typical and jointly typical sequences as defined in Cover and Thomas [7], based on the joint distribution of the XZ-source.

A linear coding strategy S_n is specified by the parity-matrices $\boldsymbol{H}_m$ and $\boldsymbol{H}_k$, with dimensions $(n\rho_m \times n)$ and $(n\rho_k \times n)$ respectively. The encoder observes a sequence $\boldsymbol{x}$ of length n and generates a helper message $\boldsymbol{m} = \boldsymbol{H}_m\boldsymbol{x}^T$ and secret key $\boldsymbol{k} = \boldsymbol{H}_k\boldsymbol{x}^T$ and sends the helper message over a public channel to the decoder. The secret key is a binary vector of length $n\rho_k$, and $|\mathcal{K}| = 2^{n\rho_k}$. The decoder observes a sequence $\boldsymbol{z}$ and reconstructs the unique sequence $\widehat{\boldsymbol{x}}$ such that $\boldsymbol{H}_m\widehat{\boldsymbol{x}}^T = \boldsymbol{m}$ and $(\widehat{\boldsymbol{x}},\boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XZ)$. If the reconstruction of $\boldsymbol{x}$ is successful the decoder can also reconstruct the secret $\widehat{\boldsymbol{k}} = \boldsymbol{H}_k\widehat{\boldsymbol{x}}^T$.

Finally, we introduce a virtual decoder that observes both the secret $\boldsymbol{k}$ and the helper message $\boldsymbol{m}$ and reconstructs $\widetilde{\boldsymbol{x}}$ such that $\boldsymbol{H}_m\widetilde{\boldsymbol{x}}^T = \boldsymbol{m}$ and $\boldsymbol{H}_k\widetilde{\boldsymbol{x}}^T = \boldsymbol{k}$ and $\widetilde{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X)$.

We measure the reliability of a linear coding strategy S_n in terms of the error probability

$$P_e(S_n) = \Pr(\widehat{\boldsymbol{X}} \neq \boldsymbol{X} \text{ or } \widetilde{\boldsymbol{X}} \neq \boldsymbol{X}|S_n).$$

Next we assume that all matrix elements are chosen uniformly from $\{0,1\}$ and we bound the error probability averaged over all randomly generated linear codes $\boldsymbol{H}_m$, $\boldsymbol{H}_k$ as $E[P_e(S_n)] \leq \Pr(E_0) + \Pr(E_1) + \Pr(E_2)$, with

$$\begin{aligned} E_0 &= \{(\boldsymbol{X},\boldsymbol{Z}) \notin \mathcal{A}_\epsilon^{(n)}(XZ)\}, \\ E_1 &= \{\exists\widehat{\boldsymbol{x}} \neq \boldsymbol{X} : \boldsymbol{H}_m\widehat{\boldsymbol{x}}^T = \boldsymbol{H}_m\boldsymbol{X}^T \text{ and} \\ &\qquad (\widehat{\boldsymbol{x}},\boldsymbol{Z}) \in \mathcal{A}_\epsilon^{(n)}(XZ)\}, \\ E_2 &= \{\exists\widetilde{\boldsymbol{x}} \neq \boldsymbol{X} : \boldsymbol{H}_m\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_m\boldsymbol{X}^T \text{ and} \\ &\qquad \boldsymbol{H}_k\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_k\boldsymbol{X}^T \text{ and } \widetilde{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X)\}. \end{aligned}$$

By the properties of typical sequences $\Pr(E_0) < \epsilon$.

$$\begin{aligned} \Pr(E_1) &= \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z}) \Pr(\exists\widehat{\boldsymbol{x}} \neq \boldsymbol{x} : \boldsymbol{H}_m\widehat{\boldsymbol{x}}^T = \boldsymbol{H}_m\boldsymbol{x}^T \text{ and} \\ &\qquad (\widehat{\boldsymbol{x}},\boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XZ)) \\ &\leq \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z}) \sum_{\widehat{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X|\boldsymbol{z}), \widehat{\boldsymbol{x}} \neq \boldsymbol{x}} \Pr(\boldsymbol{H}_m(\widehat{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0}) \\ &\leq \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z}) |\mathcal{A}_\epsilon^{(n)}(X|\boldsymbol{z})| 2^{-n\rho_m} \\ &\leq 2^{n(H(X|Z)+2\epsilon)} 2^{-n\rho_m}. \end{aligned}$$

Note that vector $\boldsymbol{e} = \widehat{\boldsymbol{x}} \oplus \boldsymbol{x}$ has at least one non-zero component (since $\boldsymbol{x} \neq \widehat{\boldsymbol{x}}$), say at position j. Then for a randomly generated $\boldsymbol{H}_m$ of dimension $(n\rho_m \times n)$

$$\begin{aligned} \Pr(\boldsymbol{H}_m\boldsymbol{e}^T = \boldsymbol{0}) &= \Pr\left(\sum_{i=1}^{n} \boldsymbol{H}_m(:,i)\boldsymbol{e}(i) = \boldsymbol{0}\right) \\ &= \Pr\left(\sum_{i=1, i \neq j}^{n} \boldsymbol{H}_m(:,i)\boldsymbol{e}(i) = \boldsymbol{H}_m(:,j)\right) \\ &= \Pr(\boldsymbol{H}_m(:,j) = \boldsymbol{c}) = 2^{-n\rho_m}, \end{aligned}$$

where $\boldsymbol{H}(:,i)$ corresponds to the i^{th} column of the matrix and $\boldsymbol{e}(i)$ corresponds to the value at the i^{th} position of vector $\boldsymbol{e}$ and $\boldsymbol{c}$ corresponds to some column vector. Next

$$\Pr(E_2) = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \Pr(\exists\widetilde{\boldsymbol{x}} \neq \boldsymbol{x} : \boldsymbol{H}_m\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_m\boldsymbol{x}^T \text{ and}$$

$$
\begin{aligned}
&\boldsymbol{H}_k\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_k\boldsymbol{x}^T \text{ and } \widetilde{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X))\\
&\le \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \sum_{\widetilde{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X), \widetilde{\boldsymbol{x}} \neq \boldsymbol{x}} \Pr(\boldsymbol{H}_m(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0}, \boldsymbol{H}_k(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0})\\
&\le \sum_{\boldsymbol{x}} p(\boldsymbol{x}) |\mathcal{A}_\epsilon^{(n)}(X)| 2^{-n(\rho_m+\rho_k)}\\
&\le 2^{n(H(X)+\epsilon)} 2^{-n(\rho_m+\rho_k)}.
\end{aligned}
$$

We conclude that $E[P_e(S_n)] \le 3\epsilon$ as long as

$$
\begin{aligned}
\rho_m &> H(X|Z) + 2\epsilon,\\
\rho_m + \rho_k &> H(X) + \epsilon.
\end{aligned}
$$

From this it follows that linear codes exist with $\rho_m = H(X|Z)+3\epsilon$ and $\rho_k = I(X;Z) - \epsilon$, such that decoding (both for the helper message decoder and the virtual decoder) has an acceptable error probability, as long as n is large enough.

C. Leakage, Uniformity of the keys, and Rate

Now, we investigate for the linear codes found before, having dimensions $n\rho_m$ and $n\rho_k$, the resulting leakage about secret $\boldsymbol{K}$ from syndrome $\boldsymbol{m}$.

$$
\begin{aligned}
I(\boldsymbol{K};\boldsymbol{M}) &= H(\boldsymbol{K}) + H(\boldsymbol{M}) - H(\boldsymbol{MK})\\
&\le H(\boldsymbol{K}) + H(\boldsymbol{M}) - H(\boldsymbol{XMK}) + H(\boldsymbol{X}|\boldsymbol{MK})\\
&\le nI(X;Z) - n\epsilon + n(H(X|Z) + 3\epsilon) - nH(X) +\\
&\quad 1 + nE[P_e(S_n)])\\
&\le 2n\epsilon + 1 + nE[P_e(S_n)] \le 5n\epsilon + 1,
\end{aligned}
$$

thus for n large enough and some appropriate choice for ϵ we conclude that $\frac{1}{n}I(\boldsymbol{K};\boldsymbol{M}) \le \delta$ for any $\delta > 0$, and the leakage requirement is satisfied.

Next we find that

$$
\begin{aligned}
H(\boldsymbol{X}) &= H(\boldsymbol{XMK})\\
&\le H(\boldsymbol{K}) + H(\boldsymbol{M}) + H(\boldsymbol{X}|\boldsymbol{MK})\\
&\le H(\boldsymbol{K}) + n(H(X|Z) + 3\epsilon) +\\
&\quad 1 + nE[P_e(S_n)], \text{ and thus}\\
\frac{1}{n}H(\boldsymbol{K}) &\ge I(X;Z) - 6\epsilon - \frac{1}{n}.
\end{aligned}
$$

Therefore, for any $\delta > 0$ we obtain that $\frac{1}{n}H(\boldsymbol{K}) + \delta \ge I(X;Z) - \epsilon = \frac{1}{n}\log_2|\mathcal{K}| \ge R - \delta = I(X;Z) - \delta$ by suitable choice of ϵ and large enough n. Now the uniformity/rate condition of Definition 1 is satisfied. It follows that rate $R = I(X;Z)$ is achievable, and that the secret-key capacity is $I(X;Z)$. This concludes the proof.

IV. 2-ENROLLMENT SETTING

In the 2-enrollment setting shown in Figure 2, we assume that a first enrollment is performed by encoder 1. A secret key k_1 and helper message m_1 are generated based on observation of $\boldsymbol{x}$. A decoder that observes $\boldsymbol{z}$ and m_1 has sufficient information to form an estimate $\widehat{k}_1$ of secret k_1.

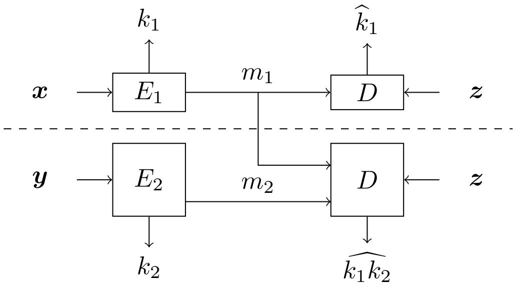


Fig. 2: Two enrollments scenario.

Encoder 2 observes observation vector $\boldsymbol{y}$ and performs a second enrollment, generating a secret key k_2 and corresponding helper message m_2. A decoder that observes both helper messages m_1 and m_2, and an observation vector $\boldsymbol{z}$ should be able to form the estimate $\widehat{k_1k_2}$. Furthermore, both helper messages should not reveal any information about the secret keys to an attacker who can not observe any observation vector $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{z}$. Finally the secret keys k_1 and k_2 should be uniformly distributed and independent of each other.

Definition 2: A secret-key rate pair (R_1, R_2) is called achievable in the 2-enrollment setting, if for all $\delta > 0$ and for all n large enough, there exist encoders and decoders such that

$$
\begin{aligned}
\Pr(\widehat{K}_1 \neq K_1 \vee \widehat{K_1K_2} \neq K_1K_2) &\le \delta,\\
\frac{1}{n}H(K_1K_2) + \delta &\ge \frac{1}{n}\log_2|\mathcal{K}_1||\mathcal{K}_2|,\\
\frac{1}{n}\log_2|\mathcal{K}_1| &\ge R_1 - \delta,\\
\frac{1}{n}\log_2|\mathcal{K}_2| &\ge R_2 - \delta,\\
\frac{1}{n}I(K_1K_2; M_1M_2) &\le \delta.
\end{aligned}
$$

The secret-key capacity is the maximum achievable total rate $C = R_1 + R_2$.

Theorem 2: The secret-key capacity $C = I(XY;Z)$ and the achievable secret-key rate pairs (R_1, R_2) satisfy

$$
\begin{aligned}
R_1 &\le I(X;Z)\\
R_1 + R_2 &\le I(XY;Z).
\end{aligned}
$$

In the following, we first show the converse of Theorem 2, that is no secret-key rate pairs (R_1, R_2) exist for which $R_1 > I(X;Z)$ or $R_1 + R_2 > I(XY;Z)$.

A. Converse for 2-enrollment secret-key capacity

The upper bound to the achievable rate for the first key R_1 follows from our results for the 1-enrollment scheme. We continue with

$$
\begin{aligned}
H(K_1K_2|\boldsymbol{Z}M_1M_2) &= H(K_1K_2|\boldsymbol{Z}M_1M_2\widehat{K_1K_2})\\
&\le H(K_1K_2|\widehat{K_1K_2})\\
&\le 1 + P_{ek}\log_2|\mathcal{K}_1||\mathcal{K}_2| \le 1 + 2nP_{ek},
\end{aligned}
$$

with $P_{ek} = \Pr(\widehat{K_1K_2} \neq K_1K_2)$. For achievable rates

$$
H(K_1K_2) = I(K_1K_2; \boldsymbol{Z}M_1M_2) + H(K_1K_2|\boldsymbol{Z}M_1M_2)
$$

$$
\begin{aligned}
&\le I(K_1K_2;M_1M_2) + I(K_1K_2;\boldsymbol{Z}|M_1M_2) + 1 + 2nP_{ek}\\
&\le I(K_1K_2;M_1M_2) + H(\boldsymbol{Z}) -\\
&\quad H(\boldsymbol{Z}|M_1M_2K_1K_2\boldsymbol{XY}) + 1 + 2nP_{ek}\\
&= I(K_1K_2;M_1M_2) + H(\boldsymbol{Z}) -\\
&\quad H(\boldsymbol{Z}|\boldsymbol{XY}) + 1 + 2nP_{ek}\\
&\le n\delta + nI(XY;Z) + 1 + 2n\delta.
\end{aligned}
$$

This results in

$$
\begin{aligned}
R_1 - \delta + R_2 - \delta &\le \frac{1}{n} H(K_1K_2) + \delta\\
&\le 3\delta + I(XY;Z) + \frac{1}{n} + \delta.
\end{aligned}
$$

Now with $\delta \downarrow 0$ and $n \to \infty$ we obtain the bound $R_1 + R_2 \le I(XY;Z)$ for achievable rate pairs.

B. Linear codes for 2-enrollment setting

Next, we demonstrate that linear codes exist that achieve the rates specified in Theorem 2, for the 2-enrollment scheme.

Fix an $\epsilon > 0$. Now $\mathcal{A}_\epsilon^{(n)}(XYZ)$ is the set of jointly typical sequences as defined in Cover and Thomas [7], based on the joint distribution of the XYZ-source.

We specify a linear coding strategy S_n by four parity check matrices. Two parity-check matrices $\boldsymbol{H}_{m1}$ and $\boldsymbol{H}_{k1}$ for the first encoder, with dimensions $(n\rho_{m1} \times n)$ and $(n\rho_{k1} \times n)$ respectively, and two further parity-check matrices $\boldsymbol{H}_{m2}$ and $\boldsymbol{H}_{k2}$ for the second encoder, with dimensions $(n\rho_{m2} \times n)$ and $(n\rho_{k2} \times n)$ respectively.

Encoder 1 observes a sequence $\boldsymbol{x}$ of length n, generates a helper message $\boldsymbol{m}_1 = \boldsymbol{H}_{m1}\boldsymbol{x}^T$ of length $n\rho_{m1}$ and a secret key $\boldsymbol{k}_1 = \boldsymbol{H}_{k1}\boldsymbol{x}^T$ of length $n\rho_{k1}$, and sends this helper message over a public channel to the decoders. A one-step decoder that has received $\boldsymbol{m}_1$ and did observe a sequence $\boldsymbol{z}$, reconstructs the unique sequence $\widehat{\boldsymbol{x}}$ such that $\boldsymbol{m}_1 = \boldsymbol{H}_{m1}\widehat{\boldsymbol{x}}^T$ and $(\widehat{\boldsymbol{x}}, \boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XZ)$. If the reconstruction of $\boldsymbol{x}$ is successful, this one-step decoder can reconstruct the secret $\widehat{\boldsymbol{k}} = \boldsymbol{H}_{k1}\widehat{\boldsymbol{x}}^T$.

Encoder 2 observes a sequence $\boldsymbol{y}$ of length n, generates a helper message $\boldsymbol{m}_2 = \boldsymbol{H}_{m2}\boldsymbol{y}^T$ of length $n\rho_{m2}$ and a secret key $\boldsymbol{k}_2 = \boldsymbol{H}_{k2}\boldsymbol{y}^T$ of length $n\rho_{k2}$, and sends the helper message $\boldsymbol{m}_2$ over a public channel to the two-step decoder. This decoder has already processed the first step, see above. In addition, since it has obtained $\boldsymbol{m}_2$ this two-step decoder reconstructs the unique $\widehat{\boldsymbol{y}}$ such that $\boldsymbol{m}_2 = \boldsymbol{H}_{m2}\widehat{\boldsymbol{y}}^T$ and $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XYZ)$, where $\widehat{\boldsymbol{x}}$ was determined in the first step. Note that the secrets are binary vectors with $|\mathcal{K}_1| = 2^{n\rho_{k1}}$ and $|\mathcal{K}_2| = 2^{n\rho_{k2}}$.

Finally, we define a virtual decoder that observes both secrets $\boldsymbol{k}_1, \boldsymbol{k}_2$ and the helper messages $\boldsymbol{m}_1, \boldsymbol{m}_2$ and reconstructs $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ such that $\boldsymbol{k}_1 = \boldsymbol{H}_{k1}\widetilde{\boldsymbol{x}}^T$, $\boldsymbol{m}_1 = \boldsymbol{H}_{m1}\widetilde{\boldsymbol{x}}^T$ and $\boldsymbol{k}_2 = \boldsymbol{H}_{k2}\widetilde{\boldsymbol{y}}^T$ and $\boldsymbol{m}_2 = \boldsymbol{H}_{m2}\widetilde{\boldsymbol{y}}^T$ and $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \in \mathcal{A}_\epsilon^{(n)}(XY)$.

We measure the reliability of a linear coding strategy S_n in terms of the average error probability

$$
P_e(S_n) = \Pr(\widehat{\boldsymbol{X}} \ne \boldsymbol{X} \text{ or } \widehat{\boldsymbol{Y}} \ne \boldsymbol{Y} \text{ or } \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{Y}} \ne \boldsymbol{XY}|S_n).
$$

Matrix elements are chosen uniformly. We now bound the error probability averaged over all possible (randomly generated) linear codes $\boldsymbol{H}_{m1}$, $\boldsymbol{H}_{k1}$, $\boldsymbol{H}_{m2}$, $\boldsymbol{H}_{k2}$ as $E[P_e(Sn)] \le \Pr(E_0) + \Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \Pr(E_4) + \Pr(E_5)$, with

$$
\begin{aligned}
E_0 &= \{(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \notin \mathcal{A}_\epsilon^{(n)}(XYZ)\},\\
E_1 &= \{\exists \widehat{\boldsymbol{x}} \ne \boldsymbol{X} : \boldsymbol{H}_{m1}\widehat{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{X}^T \text{ and}\\
&\qquad (\widehat{\boldsymbol{x}}, \boldsymbol{Z}) \in \mathcal{A}_\epsilon^{(n)}(XZ)\},\\
E_2 &= \{\exists \widehat{\boldsymbol{y}} \ne \boldsymbol{Y} : \boldsymbol{H}_{m2}\widehat{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{Y}^T \text{ and}\\
&\qquad (\boldsymbol{X}, \widehat{\boldsymbol{y}}, \boldsymbol{Z}) \in \mathcal{A}_\epsilon^{(n)}(XYZ)\},\\
E_3 &= \{\exists \widetilde{\boldsymbol{x}} \ne \boldsymbol{X} : \boldsymbol{H}_{k1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{k1}\boldsymbol{X}^T \text{ and}\\
&\qquad \boldsymbol{H}_{m1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{X}^T \text{ and } (\widetilde{\boldsymbol{x}}, \boldsymbol{Y}) \in \mathcal{A}_\epsilon^{(n)}(XY)\},\\
E_4 &= \{\exists \widetilde{\boldsymbol{y}} \ne \boldsymbol{Y} : \boldsymbol{H}_{k2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{k2}\boldsymbol{Y}^T \text{ and}\\
&\qquad \boldsymbol{H}_{m2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{Y}^T \text{ and } (\boldsymbol{X}, \widetilde{\boldsymbol{y}}) \in \mathcal{A}_\epsilon^{(n)}(XY)\},\\
E_5 &= \{\exists (\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \ne (\boldsymbol{X}, \boldsymbol{Y}) : \boldsymbol{H}_{k1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{k1}\boldsymbol{X}^T \text{ and}\\
&\qquad \boldsymbol{H}_{m1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{X}^T \text{ and } \boldsymbol{H}_{k2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{k2}\boldsymbol{Y}^T \text{ and}\\
&\qquad \boldsymbol{H}_{m2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{Y}^T \text{ and } (\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \in \mathcal{A}_\epsilon^{(n)}(XY)\}.
\end{aligned}
$$

By the properties of typical sequences $\Pr(E_0) < \epsilon$.

$$
\begin{aligned}
\Pr(E_1) &= \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) \Pr(\exists \widehat{\boldsymbol{x}} \ne \boldsymbol{x} :\\
&\qquad \boldsymbol{H}_{m1}\widehat{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{x}^T \text{ and } (\widehat{\boldsymbol{x}}, \boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XZ))\\
&\le \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) \sum_{\widehat{\boldsymbol{x}} \in \mathcal{A}_\epsilon^{(n)}(X|\boldsymbol{z}), \widehat{\boldsymbol{x}} \ne \boldsymbol{x}} \Pr(\boldsymbol{H}_{m1}(\widehat{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0})\\
&\le \sum_{\boldsymbol{x},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) |\mathcal{A}_\epsilon^{(n)}(X|\boldsymbol{z})| 2^{-n\rho_{m1}}\\
&\le 2^{n(H(X|Z)+2\epsilon)} 2^{-n\rho_{m1}}.
\end{aligned}
$$

$$
\begin{aligned}
\Pr(E_2) &= \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \Pr(\exists \widehat{\boldsymbol{y}} \ne \boldsymbol{y} :\\
&\qquad \boldsymbol{H}_{m2}\widehat{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{y}^T \text{ and } (\boldsymbol{x}, \widehat{\boldsymbol{y}}, \boldsymbol{z}) \in \mathcal{A}_\epsilon^{(n)}(XYZ))\\
&\le \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \sum_{\widehat{\boldsymbol{y}} \in \mathcal{A}_\epsilon^{(n)}(Y|\boldsymbol{xz}), \widehat{\boldsymbol{y}} \ne \boldsymbol{y}} \Pr(\boldsymbol{H}_{m2}(\widehat{\boldsymbol{y}} \oplus \boldsymbol{y})^T = \boldsymbol{0})\\
&\le \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) |\mathcal{A}_\epsilon^{(n)}(Y|\boldsymbol{xz})| 2^{-n\rho_{m2}}\\
&\le 2^{n(H(Y|XZ)+2\epsilon)} 2^{-n\rho_{m2}}.
\end{aligned}
$$

$$
\begin{aligned}
\Pr(E_3) &= \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}) \Pr(\exists \widetilde{\boldsymbol{x}} \ne \boldsymbol{x} : \boldsymbol{H}_{k1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{k1}\boldsymbol{x}^T \text{ and}\\
&\qquad \boldsymbol{H}_{m1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{x}^T \text{ and } (\widetilde{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{A}_\epsilon^{(n)}(XY))\\
&\le \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}) \sum_{(\widetilde{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{A}_\epsilon^{(n)}(XY), \widetilde{\boldsymbol{x}} \ne \boldsymbol{x}} \Pr(\boldsymbol{H}_{k1}(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0},\\
&\qquad \boldsymbol{H}_{m1}(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0})\\
&\le \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}) |\mathcal{A}_\epsilon^{(n)}(X|\boldsymbol{y})| 2^{-n(\rho_{m1}+\rho_{k1})}
\end{aligned}
$$

$$\leq 2^{n(H(X|Y)+2\epsilon)}2^{-n(\rho_{m1}+\rho_{k1})}.$$

$$\begin{aligned}
\Pr(E_4) &= \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) \Pr(\exists \widetilde{\boldsymbol{y}} \neq \boldsymbol{y} : \boldsymbol{H}_{k2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{k2}\boldsymbol{y}^T \text{ and} \\
&\quad \boldsymbol{H}_{m2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{y}^T \text{ and } (\boldsymbol{x},\widetilde{\boldsymbol{y}}) \in \mathcal{A}_\epsilon^{(n)}(XY)) \\
&\leq \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) \sum_{(\boldsymbol{x},\widetilde{\boldsymbol{y}})\in\mathcal{A}_\epsilon^{(n)}(Y|\boldsymbol{x}),\widetilde{\boldsymbol{y}}\neq\boldsymbol{y}} \Pr(\boldsymbol{H}_{k2}(\widetilde{\boldsymbol{y}} \oplus \boldsymbol{y})^T = \boldsymbol{0}, \\
&\quad \boldsymbol{H}_{m2}(\widetilde{\boldsymbol{y}} \oplus \boldsymbol{y})^T = \boldsymbol{0}) \\
&\leq \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) |\mathcal{A}_\epsilon^{(n)}(Y|\boldsymbol{x})| 2^{-n(\rho_{m2}+\rho_{k2})} \\
&\leq 2^{n(H(Y|X)+2\epsilon)}2^{-n(\rho_{m2}+\rho_{k2})}.
\end{aligned}$$

$$\begin{aligned}
\Pr(E_5) &= \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) \Pr(\exists (\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \neq (\boldsymbol{x},\boldsymbol{y}) : \\
&\quad \boldsymbol{H}_{k1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{k1}\boldsymbol{x}^T \text{ and } \boldsymbol{H}_{m1}\widetilde{\boldsymbol{x}}^T = \boldsymbol{H}_{m1}\boldsymbol{x}^T \text{ and} \\
&\quad \boldsymbol{H}_{k2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{k2}\boldsymbol{y}^T \text{ and } \boldsymbol{H}_{m2}\widetilde{\boldsymbol{y}}^T = \boldsymbol{H}_{m2}\boldsymbol{y}^T \text{ and} \\
&\quad (\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \in \mathcal{A}_\epsilon^{(n)}(XY)) \\
&\leq \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) \sum_{(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\in\mathcal{A}_\epsilon^{(n)}(XY),\widetilde{\boldsymbol{y}}\neq\boldsymbol{y}} \Pr(\boldsymbol{H}_{k1}(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0}, \\
&\quad \boldsymbol{H}_{m1}(\widetilde{\boldsymbol{x}} \oplus \boldsymbol{x})^T = \boldsymbol{0}, \boldsymbol{H}_{k2}(\widetilde{\boldsymbol{y}} \oplus \boldsymbol{y})^T = \boldsymbol{0}, \\
&\quad \boldsymbol{H}_{m2}(\widetilde{\boldsymbol{y}} \oplus \boldsymbol{y})^T = \boldsymbol{0}) \\
&\leq \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}) |\mathcal{A}_\epsilon^{(n)}(XY)| 2^{-n(\rho_{m1}+\rho_{k1}+\rho_{m2}+\rho_{k2})} \\
&\leq 2^{n(H(XY)+\epsilon)}2^{-n(\rho_{m1}+\rho_{k1}+\rho_{m2}+\rho_{k2})}.
\end{aligned}$$

We conclude that for n large enough $E(P_e(S_n)) \leq 6\epsilon$ as long as

$$\begin{aligned}
\rho_{m1} &> H(X|Z) + 2\epsilon \\
\rho_{m2} &> H(Y|XZ) + 2\epsilon \\
\rho_{k1} + \rho_{m1} &> H(X|Y) + 2\epsilon \\
\rho_{k2} + \rho_{m2} &> H(Y|X) + 2\epsilon \\
\rho_{k1} + \rho_{m1} + \rho_{k2} + \rho_{m2} &> H(XY) + \epsilon.
\end{aligned}$$

Therefore, we have shown that linear codes exist that have $\rho_{m1} = H(X|Z) + 3\epsilon$ and $\rho_{m2} = H(Y|XZ) + 3\epsilon$, such that the decoders can successfully decode the secrets as long as n is large enough. Furthermore, we choose $\rho_{k1} = \alpha$ and $\rho_{k2} = I(XY;Z) - \alpha$ with $0 \leq \alpha \leq I(X;Z)$. Note that this choice for the matrix dimensions satisfies the above inequalities for SRAM-PUF's, since $H(X|Z) = H(X|Y)$ in this case.

C. Zero-leakage and uniformity of the keys

Now, we focus on the linear codes with dimensions ρ_{m1}, ρ_{k1}, ρ_{m2}, and ρ_{k2} that we have found before. The resulting leakage about the secrets from syndrome $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ is

$$\begin{aligned}
&I(\boldsymbol{K}_1\boldsymbol{K}_2; \boldsymbol{M}_1\boldsymbol{M}_2) \\
&\quad = H(\boldsymbol{K}_1\boldsymbol{K}_2) + H(\boldsymbol{M}_1\boldsymbol{M}_2) - H(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{K}_1\boldsymbol{K}_2) \\
&\quad \leq H(\boldsymbol{K}_1\boldsymbol{K}_2) + H(\boldsymbol{M}_1\boldsymbol{M}_2) - \\
&\quad\quad H(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{K}_1\boldsymbol{K}_2) + H(\boldsymbol{X}\boldsymbol{Y}|\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{K}_1\boldsymbol{K}_2) \\
&\quad \leq H(\boldsymbol{K}_1\boldsymbol{K}_2) + n(H(X|Z) + H(Y|XZ) + 6\epsilon) - \\
&\quad\quad H(\boldsymbol{X}\boldsymbol{Y}) + 1 + 2nE[P_e(S_n)] \\
&\quad \leq 6n\epsilon + 1 + 2nE[P_e(S_n)] < 6n\epsilon + 1 + 12n\epsilon.
\end{aligned}$$

Now for n large enough and an appropriate choice for ϵ we conclude that $\frac{1}{n}I(\boldsymbol{K}_1\boldsymbol{K}_2;\boldsymbol{M}_1\boldsymbol{M}_2) \leq \delta$ for any $\delta > 0$, which satisfies the leakage requirement.

Next we find, from Fano's inequality for the virtual decoder, that

$$\begin{aligned}
H(\boldsymbol{X}\boldsymbol{Y}) &= H(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{K}_1\boldsymbol{M}_1\boldsymbol{K}_2\boldsymbol{M}_2) \\
&\leq H(\boldsymbol{K}_1\boldsymbol{K}_2) + H(\boldsymbol{M}_1) + H(\boldsymbol{M}_2) + \\
&\quad H(\boldsymbol{X}\boldsymbol{Y}|\boldsymbol{K}_1\boldsymbol{M}_1\boldsymbol{K}_2\boldsymbol{M}_2) \\
&\leq H(\boldsymbol{K}_1\boldsymbol{K}_2) + n(H(X|Z) + H(Y|XZ) + 6\epsilon) + \\
&\quad 1 + 2nE[P_e(S_n)], \text{and thus}
\end{aligned}$$

$$\frac{1}{n}H(\boldsymbol{K}_1\boldsymbol{K}_2) \geq I(XY;Z) - 18\epsilon - \frac{1}{n}.$$

Therefore, for any $\delta > 0$ we obtain that $\frac{1}{n}H(\boldsymbol{K}_1\boldsymbol{K}_2) + \delta \geq I(XY;Z) = \frac{1}{n}\log_2|\mathcal{K}_1||\mathcal{K}_2| \geq R_1 - \delta + R_2 - \delta = I(XY;Z) - \delta$, by suitable choice of ϵ and large enough n. Thus the uniformity and independence conditions of the secret keys in Definition 2 hold. We conclude that the achievable secret-key rates are

$$R_1 \leq I(X;Z), R_1 + R_2 \leq I(XY;Z).$$

V. Conclusion

We have shown that an encoder can generate a second key after observing additional input, while ensuring that both helper messages do not reveal information about the keys. The total achievable secret-key rate increases from $I(X;Z)$ for the first key, to $I(XY;Z)$ for both keys. Therefore, the same rate can be achieved sequentially by the 2-enrollment scheme, as would be achievable when the encoder would generate a single key and helper message based on two observations. Finally, we note that the encoder does not require any information about the first enrollment in order to realize the second enrollment.

References

[1] D. E. Holcomb, W. P. Burleson, and K. Fu, “Power-Up SRAM state as an identifying fingerprint and source of true random numbers,” *IEEE Trans. Comput.*, vol. 58, no. 9, pp. 1198–1210, 2009.

[2] M. Bloch and J. a. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*, 1st ed. New York, NY, USA: Cambridge University Press, 2011.

[3] U. Maurer, “Secret key agreement by public discussion from common information,” *IEEE Trans. Inf. Theory*, vol. 39, pp. 733–742, May 1993.

[4] R. Ahlswede and I. Csiszár, “Common randomness in information theory and cryptography - part I: Secret sharing,” *IEEE Trans. Inf. Theory*, vol. 39, pp. 1121–1132, July 1993.

[5] L. Kusters, T. Ignatenko, F. M. J. Willems, R. Maes, E. van der Sluis, and G. Selimis, “Security of helper data schemes for sram-puf in multiple enrollment scenarios,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1803–1807.

[6] T. Ignatenko and F. Willems, “On the Security of the XOR-Method in Biometric Authentication Systems,” in *27th Symp. Inform. Theory Benelux*, 2006, pp. 197–204.

[7] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley & Sons, 2006.

The Decentralized Structures of Capacity Achieving Distributions of Channels with Memory and Feedback

Charalambos D. Charalambous, Christos K. Kourtellaris, Ioannis Tzortzis and Sergey Loyka

***Abstract*— We consider extremum problems of feedback capacity for models with memory, subject to average cost constraints. We show the optimal input process that maximizes directed information consists of two parts, one responsible to control the output process, and one responsible to transmit new information that interact. Unlike [1], the decentralized structure of the optimal input process is demonstrated for Gaussian models with memory on past inputs and outputs. A semi-separation principle is shown that states, the optimal input process is generated from multiple strategies of a decentralized optimization problem, of control and information transmission. Further, it is shown that the derivation of directed information stability is semi-separable, in the sense that it separates into a statement about the ergodic properties of the stochastic optimal control problem with partial information, and a statement related to an information transmission problem.**

I. Introduction

Recently, it is shown that Shannon's coding capacity extends to unstable dynamic systems, irrespectively of whether these are communication channels or control systems [2] (see also [1], [3] for extensive analysis). Shannon's coding capacity is called *control-coding capacity* to emphasize the interaction of control and information transmission parts of the optimal input process, that achieves capacity.

MIMO G-RM. This paper utilizes some of the results found in the above references, to investigate Multiple-Input Multiple-Output (MIMO) Gaussian Recursive Models (G-RMs), with input process $A^n \triangleq \{A_0, A_1, \ldots, A_n\}$ and output process $Y^n \triangleq \{Y_0, Y_1, \ldots, Y_n\}$, described by

$$Y_i = C^{i-1}\, Y^{i-1} + D_{i,i}\, A_i + D_{i,i-1}A_{i-1} + V_i, \tag{1}$$

$$S \triangleq (Y^{-1}, A_{-1}) = (y^{-1}, a_{-1}) \equiv s,$$

$$\mathbf{P}_{V_i|V^{i-1},A^i,S} = \mathbf{P}_{V_i}, V_i \sim N(0, K_{V_i}), K_{V_i} \succ 0, \tag{2}$$

$$(Y^{-1}, A_{-1}) \sim N(0, K_{Y^{-1},A_{-1}}),\ K_{Y^{-1},A_{-1}} \succ 0, \tag{3}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}\langle A_i, R_i A_i\rangle + \langle Y_{i-1}, Q_{i,i-1}Y_{i-1}\rangle\Big\} \leq \kappa, \tag{4}$$

$$(D_{i,i}, D_{i,i-1}) \in \mathbb{R}^{p\times q} \times \mathbb{R}^{p\times q}, \tag{5}$$

$$R_i \in \mathbb{S}_{++}^{q\times q}, \quad Q_{i,i-1} \in \mathbb{S}_{+}^{p\times p}, \quad i = 0, \ldots, n. \tag{6}$$

Here S is the initial data, $V_i \sim N(0, K_{V_i}), i = 0, 1, \ldots, n$ denotes zero mean Gaussian process, $\langle \cdot, \cdot \rangle$ denotes inner product of elements of linear spaces, $\mathbb{S}_{+}^{q\times q}$ denotes the set of symmetric positive semi-definite $q \times q$ matrices, $\mathbb{S}_{++}^{q\times q}$ its subset of positive definite matrices, and κ is the power. The initial state $S = s$ is known to the encoder and the decoder.

Main Results. For the extremum problem of maximizing directed information from A^n to B^n given the initial state $S = s$, denoted by $I(A^n \rightarrow B^n|s)$, over conditional distributions $\mathbf{P}_{A_i|A^{i-1},B^{i-1},S}, i = 0, \ldots, n$, that satisfy the average cost constraint, it is shown that a *semi-separation principle* holds with the following consequences.

(a) Part of the optimal input process A^n is characterized by the solution of a stochastic optimal control problem with partial information,

(b) the rest is characterized by the solution of an information transmission problem that interacts with that of the stochastic control part, and

(c) their computation is directly related to the notion of Person-by-Person (PbP) optimality, and team or global optimality in problems of optimal control and games, where two or more strategies do not share the same information, and aim at optimizing a single pay-off.

(d) The derivation of directed information stability is semi-separable, into a statement related to the ergodic properties of the stochastic optimal control problem, and a statement related to an information transmission problem, that interact in a specific order.

The semi-separation principle and its consequences (a)-(d) are attributed to the property that a Gaussian input process $\{A_i = A_i^g : i = 0, \ldots, n\}$ with corresponding Gaussian ouput process $\{Y_i = Y_i^g : i = 0, \ldots, n\}$, maximizes directed information $I(A^n \rightarrow Y^n|s)$ (subject to the average cost constraint), and that such an optimal process is given by the following orthogonal decomposition.

$$A_i^g = \overline{e}_i(Y^{g,i-1}, A_{i-1}^g, Z_i^g),\ i = 0, \ldots, n,\ S = s, \tag{7}$$

$$= U_i^g + \Lambda_{i,i-1}A_{i-1}^g + Z_i^g,\ \ U_i^g \triangleq \Gamma^{i-1}Y^{g,i-1}, \tag{8}$$

$$\equiv e_i(Y^{g,i-1}) + \Lambda_{i,i-1}A_{i-1}^g + Z_i^g \tag{9}$$

where

$$e_i(y^{i-1}) \triangleq \Gamma^{i-1}y^{i-1} \ \text{ is the control strategy,} \tag{10}$$

$$Z_i^g \ \text{ is independent of } \Big(A^{g,i-1}, Y^{g,i-1}\Big),$$

$$Z^{g,i} \ \text{ is independent of } \ V^i, \ \ i = 0, \ldots, n, \tag{11}$$

$$Z_i^g \sim N(0, K_{Z_i}) : i = 0, 1, \ldots, n \ \text{ is an independent Gaussian process} \tag{12}$$

for some deterministic matrices $\{(\Gamma^{i-1}, \Lambda_{i,i-1}) : i = 0, \ldots, n\}$ of appropriate dimensions.

Indeed, the following properties hold.

C. D. Charalambous, C. K. Kourtellaris and I. Tzortzis are with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. E-mails: {chadcha,kourtellaris.christos,tzortzis.ioannis}@ucy.ac.cy.

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

(P1) The optimal strategies $(e_i^*(\cdot), \Lambda_{i,i-1}^*, K_{Z_i}^*) : i = 0, \ldots, n\}$ are characterized by the solution of a decentralized optimization problem, where $e_i^*(\cdot), i = 0, \ldots, n$ is the solution of a stochastic optimal control problem, for a fixed $(\Lambda_{i,i-1}, K_{Z_i}) : i = 0, \ldots, n\}$, while the optimal $(\Lambda_{i,i-1}^*, K_{Z_i}^*) : i = 0, \ldots, n\}$ is the solution of an information transmission problem, with $e_i(\cdot) = e_i^*(\cdot), i = 0, \ldots, n$.

(P2) The following holds.

$$\text{If } K_{Z_i} = 0, i = 0, \ldots, n \text{ then } I(A^{g,n} \rightarrow Y^{g,n}|s) = 0.$$

(P2) is expected and easily verified, because the initial state $S = s$ is known to the encoder.

(P1) is an application of problems of optimal control and games, where two or more strategies do not share the same information, and aim at optimizing a single pay-off [4].

Special Cases of MIMO G-RM. Before we illustrate that the MIMO G-RM is fundamentally different from past investigations by other authors, we should mention that the MIMO G-RM is an infinite impulse response (IIR) model, and includes the following degenerate cases.

(1) Finite Impulse Response Model. If $C^{i-1} = 0, i = 0, \ldots, n$ then the MIMO G-RM reduces to a finite impulse response (FIR) model.

(2) No Dependence on Past Channel Inputs. If $D_{i,i-1} = 0, i = 0, \ldots, n$ then the MIMO G-RM reduces to the IIR model investigated in [1], [3].

A. Literature on Gaussian Channels with Memory & Feedback

For scalar-valued, Additive Gaussian Noise (AGN) channels with nonstationary and nonergodic noise, described by $Y_i = A_i + V_i$, $\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^n |A_i|^2\Big\} \leq \kappa$, $\mathbf{P}_{V_i|V^{i-1},A^i} = \mathbf{P}_{V_i|V^{i-1}}, i = 0, \ldots, n, V^n \sim N(0, K_{V^n})$, then the feedback capacity is characterized by Cover and Pombra [5], via

$$C_{0,n}^{CP}(\kappa) \overset{\triangle}{=} \frac{1}{2n} \max_{(\Gamma^n, K_{Z^n})} \log \frac{|\big(\Gamma^n + I\big)K_{V^n}\big(\Gamma^n + I\big)^T + K_{Z^n}|}{|K_{V^n}|} \tag{13}$$

$$\text{subject to } \frac{1}{n+1} tr\Big(\Gamma^n K_{V^n} (\Gamma^n)^T + K_{Z^n}\Big) \leq \kappa \tag{14}$$

where Z^n is a Gaussian process $N(0, K_{Z^n})$, orthogonal to V^n, and Γ^n is lower diagonal time-varying matrix with deterministic entries. Note that although, Z^n is called an "innovations process" in [5], this is not an orthogonal process. Note also that if $K_{Z^n} = 0$, since Γ^n is lower diagonal, then $C_{0,n}^{CP}(\kappa) = 0$, as expected. The closed form solution to (13) remains to this date an open problem.

The per unit time limit $C^{CP}(\kappa) \overset{\triangle}{=} \lim_{n \longrightarrow \infty} \frac{1}{n+1} C_{0,n}^{CP}(\kappa)$, for the special case of stationary ergodic noise with finite memory, described by a power spectral density $S_V(\omega) = |H(e^{j\omega})|^2$, where the filter $H(\cdot)$ is rational with stable poles and marginally stable zeros, is analyzed in [6] and in [7]. Theorem 7 and Corollary 7.1 in [7] state that capacity is achieved, when the innovations part of the input processes is zero (i.e., eqn(125) in [7] with $e_t = 0, t = 0, \ldots,$).

We should mention that Theorem 3.1 (of our paper) cannot be obtained from [6]–[8], and that the methods applied in [6], [7] are not applicable. Our results are based on a semi-separation principle and its consequences (a)-(d).

II. Feedback Capacity and Decentralized Strategies

In this section we introduce a general channel or control model (CM), and we recall the decentralized structure of the input process, and its control and communication aspects.

Consider a CM model with input process $A^n \overset{\triangle}{=} \{A_i : i = 0, 1, \ldots, n\}$, taking values in arbitrary alphabet spaces $\mathbb{A}^n \overset{\triangle}{=} \times_{i=0}^n \mathbb{A}_i$, an output process $Y^n \overset{\triangle}{=} \{Y_i : i = 0, 1, \ldots, n\}$ taking values in arbitrary alphabet spaces, $\mathbb{Y}^n \overset{\triangle}{=} \times_{i=0}^n \mathbb{Y}_i$. The initial data is $S \overset{\triangle}{=} (A^{-1}, Y^{-1}) = s \in \mathbb{S} \overset{\triangle}{=} \mathbb{A}^{-1} \times \mathbb{Y}^{-1}$.

The channel or control model (CM) is a sequence of conditional distributions

$$\mathbf{P}_{Y_i|Y^{i-1},A^i,S} \equiv Q_i(dy_i|y^{i-1}, a^i, s), \quad i = 0, \ldots, n. \tag{15}$$

The conditional distributions of the input process are chosen from the set

$$\mathcal{P}_{[0,n]} \overset{\triangle}{=} \big\{P_i(da_i|a^{i-1}, y^{i-1}, s) : i = 0, \ldots, n\big\}.$$

The above definition means, the encoder (or controller-encoder to be precise) knows the initial data $s = (y^{-1}, a^{-1})$, and applies noiseless feedback. The conditional distributions of the input process are subject to a cost constraint[1]

$$\mathcal{P}_{[0,n]}(\kappa) \overset{\triangle}{=} \Big\{P_i(da_i|a^{i-1}, y^{i-1}, s), i = 0, \ldots, n : \tag{16}$$

$$\frac{1}{n+1}\mathbf{E}_s^P\Big(\ell_{0,n}(A^n, Y^n)\Big) \leq \kappa\Big\} \subset \mathcal{P}_{[0,n]} \tag{17}$$

where $\ell_{0,n}(\cdot, \cdot) : \mathbb{A}^n \times \mathbb{Y}^n \longmapsto (-\infty, \infty]$ is a measurable function, $\kappa \in [0, \infty]$ is the total power.

The pay-off is the directed information from $A^n \overset{\triangle}{=} \{A_0, \ldots, A_n\}$ to $Y^n \overset{\triangle}{=} \{Y_0, \ldots, Y_n\}$, conditioned on the initial data $S = s$, and defined by [9], [10]

$$I(A^n \rightarrow Y^n|s) \overset{\triangle}{=} \sum_{i=0}^n I(A^i; Y_i|Y^{i-1}, s)$$

To connect directed information to the feedback capacity of the CM we introduce the following assumption [10].

Assumption 2.1: (i) If the information process to be encoded is $\{X_i : i = 0, \ldots, k\}$, then the following holds.

$$\mathbf{P}_{Y_i|Y^{i-1},A^i,S,X^k} = \mathbf{P}_{Y_i|Y^{i-1},A^i,S}, \forall k, i = 0, \ldots, n \tag{18}$$

(ii) The initial data $S = s$ is known to the encoder and decoder.

The finite-time horizon (FTH) information capacity (under Assumptions 2.1) is defined by

$$J_{A^n \rightarrow Y^n|s}(P^*, \kappa) \overset{\triangle}{=} \sup_{\mathcal{P}_{[0,n]}(\kappa)} I(A^n \rightarrow Y^n|s). \tag{19}$$

[1] The notation $\mathbf{E}_s^P$ indicates the dependence of the joint distribution on elements of $\mathcal{P}_{[0,n]}$ and the initial state $S = s$.

Throughout we assume existence of a maximizing distribution (such conditions are extracted from [11]).

The information capacity is defined by

$$C(\kappa) \triangleq \lim_{n \longrightarrow \infty} \frac{1}{n+1} J_{A^n \rightarrow Y^n|s}(P^*, \kappa) \tag{20}$$

provided the limit exists and it is finite.

Coding Theorems. Recall [12], Appendix A (code definition and achievable rate). By the converse coding theorem [13], a tight upper bound on any achievable rate is $C(\kappa)$. Moreover, if the optimal joint process $\{(A_i, Y_i) : i = 0, \ldots, n\}$ is either asymptotically stationary and ergodic [14], [15], or it induces information stability of the directed information density (see [12], Appendix A), then any code rate below $C(\kappa)$ is achievable. In general, the rate may depend on the initial data $S = s$, i.e., $C(\kappa) = C_s(\kappa)$.

Dualities of Capacity and Stochastic Optimal Control. Let $\mathcal{P}^D_{[0,n]}$ denote the restriction of randomized strategies $\mathcal{P}_{[0,n]}$ to the set of deterministic strategies

$$\mathcal{P}^D_{[0,n]} \triangleq \Big\{ a_0 = g_0(s), \ldots, a_n = g_n(s, a^{n-1}, y^{n-1}) \Big\}. \tag{21}$$

By [11], for any finite n, it can be shown that $C_{0,n}(\kappa) \triangleq J_{A^n \rightarrow Y^n|s}(P^*, \kappa), \kappa \in (\kappa_{min}, \infty) \subset [0, \infty)$ is a concave strictly increasing in $\kappa \in (\kappa_{min}, \infty)$, and the inverse function of $C_{0,n}(\kappa)$ denoted by $\kappa_{0,n}(C)$ is a convex non-decreasing in $C \in [0, \infty)$. This implies the following duality.

Dual Extremum Problem.

$$\kappa_{0,n}(C) \triangleq \inf_{\frac{1}{n+1} I(A^n \rightarrow Y^n|s) \geq C} \mathbf{E}^P_s \Big\{ \ell_{0,n}(A^n, Y^n) \Big\} \tag{22}$$

$$\geq J^{SC}_{0,n}(P^*) \triangleq \inf_{\mathcal{P}_{[0,n]}} \mathbf{E}^P_s \Big\{ \ell_{0,n}(A^n, Y^n) \Big\} \equiv \kappa_{0,n}(0) \tag{23}$$

$$= \inf_{\mathcal{P}^D_{[0,n]}} \mathbf{E}^g_s \Big\{ \ell_{0,n}(A^n, Y^n) \Big\} \equiv J^{SC}_{0,n}(g^*). \tag{24}$$

Here (24) follows from classical stochastic optimal control theory, which states that minimizing $\mathbf{E}^P_s\{\ell_{0,n}(A^n, Y^n)\}$ over $\mathcal{P}_{[0,n]}$ does not incur a better performance than maximizing it over $\mathcal{P}^D_{[0,n]}$ [16]. The minimum cost of control is $J^{SC}_{0,n}(P^*)$, and for $C \geq 0$, the cost of communication is

$$\kappa(C) - \kappa(0) \triangleq \lim_{n \longrightarrow \infty} \frac{1}{n+1} \kappa_{0,n}(C) - \lim_{n \longrightarrow \infty} \frac{1}{n+1} \kappa_{0,n}(0)$$

provided the limits exists and they are finite. Hence, for rate $C > 0$, it is necessary that the total cost of the communication system exceeds the critical value is $\kappa_{min}(n+1) = J^{SC}_{0,n}(P^*) \equiv \kappa_{0,n}(0) = J^{SC}_{0,n}(g^*)$. This is precisely the minimum cost of control, when no communication occurs, i.e., $\kappa(C) \geq \kappa_{min}$, so power is allocated to the control process. For examples of the threshold effect see [1], [3].

Suppose the randomized strategies $\mathcal{P}_{[0,n]}$ are restricted to deterministic strategies, $\mathcal{P}^D_{[0,n]}$, then by recursive substitution, $g_j(s, y^{j-1}, a^{j-1}) \equiv \overline{g}_j(s, y^{j-1})$, we have $\mathbf{P}^P(dy_i|y^{i-1}, s)\Big|_{P \in \mathcal{P}^D_{[0,n]}} = Q_i(dy_i|y^{i-1}, \{\overline{g}_0(s), \ldots, \overline{g}_i(s, y^{j-1})\}^i_{j=0}, s)$. Hence,

$$J_{A^n \rightarrow Y^n}(P^*, \kappa)\Big|_{\Big\{P^*_i(\cdot|\cdot) : i=0,\ldots,n\Big\} \in \mathcal{P}^D_{[0,n]}} = 0. \tag{25}$$

By (22), then $\kappa_{0,n}(C)\Big|_{\mathcal{P}_{[0,n]} = \mathcal{P}^D_{[0,n]}} = \kappa_{0,n}(0)$, and any optimal input process consists of a control process, which controls the output process, and a process which is responsible for information transmission.

III. Gaussian Recursive Model

Consider the G-RM (1)-(6), with $S = (Y^{-1}, A_{-1})$ known to encoder/decoder. By [17], the optimal distribution of the input is of the form $P_0(da_0|s), P_i(da_i|a_{i-1}, y^{i-1}, s), i = 1, \ldots, n$. The directed information from $A^n \triangleq \{A_0, \ldots, A_n\}$ to $Y^n \triangleq \{Y_0, \ldots, Y_n\}$ conditioned on $S = s$ is

$$I(A^n \rightarrow Y^n|s) = \sum_{i=0}^{n} \Big\{ H(Y_i|Y^{i-1}, s) - H(V_i) \Big\}. \tag{26}$$

Let $\{(A^g_i, Y^g_i, Z^g_i) : i = 0, \ldots, n\}$ denote a jointly Gaussian process, given $S = s$. By the maximum entropy property of Gaussian distributions it follows that the process given by (7)-(12), and satisfies the average constraint is optimal. Now, we prepare to compute directed information using (7)-(12). We need the following definitions[2].

$$\widehat{Y}_{i|i-1} \triangleq \mathbf{E}_s \Big\{ Y^g_i \Big| Y^{g,i-1} \Big\}, \quad \widehat{A}_{i|i} \triangleq \mathbf{E}_s \Big\{ A^g_i \Big| Y^{g,i} \Big\},$$

$$K_{Y_i|Y^{i-1}} \triangleq \mathbf{E}_s \Big\{ \Big(Y^g_i - \widehat{Y}_{i|i-1} \Big) \Big(Y^g_i - \widehat{Y}_{i|i-1} \Big)^T \Big| Y^{g,i-1} \Big\}$$

$$P_{i|i} = \mathbf{E}_s \Big(A^g_i - \widehat{A}_{i|i} \Big) \Big(A^g_i - \widehat{A}_{i|i} \Big)^T, \quad i = 0, \ldots, n.$$

From [18], and using the independent properties of the noise process, i.e., (2), (8)-(12) then

$$\widehat{A}_{i|i} = \Lambda_{i,i-1} \widehat{A}_{i-1|i-1} + U^g_i + \Delta_{i|i-1} \Big(Y^g_i - \widehat{Y}_{i|i-1} \Big), \tag{27}$$

$$\widehat{Y}_{i|i-1} = C^{i-1} Y^{g,i-1} + D_{i,i} U^g_i + \overline{\Lambda}_{i,i-1} \widehat{A}_{i-1|i-1}, \tag{28}$$

$$K_{Y_i|Y^{i-1}} = \overline{\Lambda}_{i,i-1} P_{i-1|i-1} \overline{\Lambda}^T_{i,i-1} + D_{i,i} K_{Z_i} D^T_{i,i} \tag{29}$$

$$+ K_{V_i}, i = 0, \ldots, n, \ \widehat{Y}_{0|-1} = \mathbf{E}_s\{Y^g_0\}, \widehat{A}_{-1|-1} = \mathbf{E}_s\{A^g_{-1}\}$$

where

$$\overline{\Lambda}_{i,i-1} \triangleq D_{i,i} \Lambda_{i,i-1} + D_{i,i-1}, \quad i = 0, \ldots, n,$$

$$P_{i|i} = \Lambda_{i,i-1} P_{i-1|i-1} \Lambda^T_{i,i-1} + K_{Z_i}$$

$$- \Big(K_{Z_i} D^T_{i,i} + \Lambda_{i,i-1} P_{i-1|i-1} \overline{\Lambda}^T_{i,i-1} \Big)$$

$$\Phi_{i|i-1} \Big(K_{Z_i} D^T_{i,i} + \Lambda_{i,i-1} P_{i-1|i-1} \overline{\Lambda}^T_{i,i-1} \Big)^T,$$

$$\Phi_{i|i-1} \triangleq \Big[D_{i,i} K_{Z_i} D^T_{i,i} + K_{V_i} + \overline{\Lambda}_{i,i-1} P_{i-1|i-1} \overline{\Lambda}^T_{i,i-1} \Big]^{-1},$$

$$\Delta_{i|i-1} \triangleq \Big(K_{Z_i} D^T_{i,i} + \Lambda_{i,i-1} P_{i-1|i-1} \overline{\Lambda}^T_{i,i-1} \Big) \Phi_{i|i-1}$$

The innovations process denoted by $\Big\{\nu^{\overline{e}} : i = 0, \ldots, n\Big\}$ is an orthogonal process, independent of $\{e_i(\cdot) : i = 0, \ldots, n\}$, and satisfies the following identities.

$$\nu^{\overline{e}}_i \triangleq Y^g_i - \widehat{Y}_{i|i-1} = \overline{\Lambda}_{i,i-1} \Big(A^g_{i-1} - \widehat{A}_{i-1|i-1} \Big) + D_{i,i} Z^g_i + V_i$$

$$= \nu^{\overline{e}}_i \Big|_{e=0} \equiv \nu^0_i, \ \nu^0_i \sim N(0, K_{Y_i|Y^{i-1}}), \ i = 0, \ldots, n \tag{30}$$

[2] $\mathbf{E}_s$ means conditional expectations are for fixed $S = s$.

where $\{\nu_i^0 : i = 0, \ldots, n\}$ indicates that the innovations process is independent of the strategy $\{e_i(\cdot) : i = 0, \ldots, n\}$. Then we obtain

$$I(A^{g,n} \rightarrow Y^{g,n}|s) = \frac{1}{2}\sum_{i=0}^{n} \log \frac{|K_{Y_i|Y^{i-1}}|}{|K_{V_i}|}. \qquad (31)$$

Next, we give the decentralized semi-separation principle.

Theorem 3.1: (Decentralized semi-separation of control & information transmission) Consider the G-RM (1)-(6) with $S = (Y^{-1}, A_{-1}) = s$, fixed, and for simplicity assume $C^{i-1}Y^{i-1}$ in (1) is replaced by *unit memory* $C_{i,i-1}Y_{i-1}$. Then the following hold.
(a) *Equivalent Extremum Problem.* The process given by (7)-(12) is optimal, and the following hold.

$$\begin{aligned}
&Y_i^g = C_{i,i-1}Y_{i-1}^g + \overline{\Lambda}_{i,i-1}A_{i-1}^g + D_{i,i}U_i^g + D_{i,i}Z_i^g \\
&\quad +V_i,\ i = 0, \ldots, n, \quad S \triangleq (Y_{-1}, A_{-1}) = s. \qquad (32)\\
&\mathbf{E}_s^{\bar{e}}\Big\{\gamma_i(A_i^g, Y_{i-1}^g)\Big\} \\
&= \mathbf{E}_s^{\bar{e}}\Big\{\langle U_i^g, R_iU_i^g\rangle + 2\langle \Lambda_{i,i-1}\widehat{A}_{i-1|i-1}, R_iU_i^g\rangle \\
&\ +\langle \Lambda_{i,i-1}\widehat{A}_{i-1|i-1}, R_i\Lambda_{i,i-1}\widehat{A}_{i-1|i-1}\rangle + tr\Big(K_{Z_i}R_i\Big) \\
&\ +tr\Big(\Lambda_{i,i-1}^T R_i\Lambda_{i,i-1}P_{i-1|i-1}\Big) + \langle Y_{i-1}^g, Q_iY_{i-1}^g\rangle\Big\}. \qquad (33)
\end{aligned}$$

The FTH information capacity for fixed $S = s$ is given by

$$J_{A^n\rightarrow Y^n|s}(\bar{e}^*, \kappa, s) = \sup_{\overline{\mathcal{P}}^1_{[0,n]}(\kappa)} \frac{1}{2}\sum_{i=0}^{n} \log \frac{|K_{Y_i|Y^{i-1}}|}{|K_{V_i}|} \qquad (34)$$

$$\begin{aligned}
&\overline{\mathcal{P}}_{[0,n]}(\kappa) \triangleq \Big\{\bar{e}_i(\cdot) \triangleq \big(e_i(\cdot), \Lambda_{i,i-1}, K_{Z_i}\big), i = 0, \ldots, n : \\
&\frac{1}{n+1}\sum_{i=0}^{n} \mathbf{E}_s^{\overline{e}}\Big(\gamma_i(A_i^g, Y^{g,i-1})\Big) \leq \kappa\Big\}. \qquad (35)
\end{aligned}$$

(b) *Decentralized Separation of Controller and Encoder Strategies.* The optimal strategy denoted by $\{\bar{e}^*(\cdot) \equiv (e_i^*(\cdot), \Lambda_{i,i-1}^*, K_{Z_i}^*) : i = 0, \ldots, n\}$ is the solution of the dual optimization problem

$$\begin{aligned}
\kappa_{0,n}(C, s) \triangleq &\inf_{\big(e_i(\cdot),\Lambda_{i,i-1},K_{Z_i}\big),i=0,\ldots,n:\frac{1}{2}\sum_{i=0}^{n}\log\frac{|K_{Y_i|Y^{i-1}}|}{|K_{V_i}|}\geq(n+1)C} \\
&\mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\gamma_i(A_i^g, Y_{i-1}^g)\Big\}. \qquad (36)
\end{aligned}$$

Moreover, the following decentralized separation holds.
(i) The optimal strategy $\{e_i^*(\cdot) : i = 0, \ldots, n\}$ is the solution of the stochastic optimal control problem with partial information given by

$$\inf_{e_i(\cdot):i=0,\ldots,n} \mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\gamma_i(A_i^g, Y_{i-1}^g)\Big\} \qquad (37)$$

for a fixed $\{\Lambda_{i,i-1}, K_{Z_i} : i = 0, \ldots, n\}$.
(ii) The optimal strategy $\{\Lambda_{i,i-1}^*, K_{Z_i}^* : i = 0, \ldots, n\}$ is the solution of (36) for $\{e_i(\cdot) = e_i^*(\cdot) : i = 0, \ldots, n\}$.

(c) *Optimal Strategies.* Any candidate of the control strategy $\{e_i(Y^{g,i-1}) : i = 0, \ldots, n\}$ is of the form

$$\begin{aligned}
&e_i(Y^{g,i-1}) \triangleq \Gamma_{i,i-1}^1 Y_{i-1}^g + \Gamma_{i,i-1}^2\widehat{A}_{i-1|i-1}, \qquad (38)\\
&\equiv \overline{\Gamma}_{i,i-1}\overline{Y}_{i-1}^g, \quad \overline{Y}_{i-1}^g \triangleq \begin{bmatrix} Y_{i-1}^g \\ \widehat{A}_{i-1|i-1} \end{bmatrix}, \ i = 0, \ldots, n.
\end{aligned}$$

Define the augmented system

$$\begin{aligned}
&\overline{Y}_i^g = \overline{F}_{i,i-1}\overline{Y}_{i-1}^g + \overline{B}_{i,i-1}U_i^g + \overline{G}_{i,i-1}\nu_i^{\overline{e}}, \qquad (39)\\
&\overline{F}_{i,i-1} \triangleq \begin{bmatrix} C_{i,i-1} & \overline{\Lambda}_{i,i-1} \\ 0 & \Lambda_{i,i-1} \end{bmatrix}, \overline{B}_{i,i-1} \triangleq \begin{bmatrix} D_{i,i} \\ I \end{bmatrix}, \\
&\overline{G}_{i,i-1} \triangleq \begin{bmatrix} I \\ \Delta_{i|i-1} \end{bmatrix}, \quad i = 0, \ldots, n
\end{aligned}$$

and average cost

$$\begin{aligned}
&\mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\gamma_i(A_i^g, Y_{i-1}^g)\Big\} \equiv \mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\overline{\gamma}_i(U_i^g, \overline{Y}_{i-1}^g)\Big\} \\
&\triangleq \mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\Big(\begin{bmatrix} \overline{Y}_{i-1}^g \\ U_i^g \end{bmatrix}^T \begin{bmatrix} \overline{M}_{i,i-1} & \overline{L}_{i,i-1} \\ \overline{L}_{i,i-1}^T & \overline{N}_{i,i-1} \end{bmatrix} \begin{bmatrix} \overline{Y}_{i-1}^g \\ U_i^g \end{bmatrix} \\
&\qquad +tr\big(K_{Z_i}R_i\big) + tr\big(\Lambda_{i,i-1}^T R_i\Lambda_{i,i-1}P_{i-1|i-1}\big)\Big)\Big\}, \\
&\overline{M}_{i,i-1} \triangleq \begin{bmatrix} Q_{i,i-1} & 0 \\ 0 & \Lambda_{i,i-1}^T R_i\Lambda_{i,i-1} \end{bmatrix}, \\
&\overline{L}_{i,i-1} \triangleq \begin{bmatrix} 0 \\ \Lambda_{i,i-1}^T R_i \end{bmatrix}, \quad \overline{N}_{i,i-1} \triangleq R_i.
\end{aligned}$$

Then the following hold.
(1) For a fixed $\{\Lambda_{i,i-1}, K_{Z_i} : i = 0, \ldots, n\}$ the optimal strategy $\{U_i^{g,*} = e_i^*(\overline{Y}^{g,i-1}) : i = 0, \ldots, n\}$ is the solution of the stochastic optimal control problem

$$J_{0,n}(e^*(\cdot), \Lambda, K_Z, \kappa, s) \triangleq \inf_{e_i(\cdot):i=0,\ldots,n} \mathbf{E}_s^{\overline{e}}\Big\{\sum_{i=0}^{n}\overline{\gamma}_i(U_i^g, \overline{Y}_{i-1}^g)\Big\}$$

where $\{\overline{Y}_i^g : i = 0, \ldots, n\}$ satisfy recursion (39). Moreover, the optimal strategy $\{U_i^{g,*} = e_i^*(\overline{Y}^{g,i-1}) : i = 0, \ldots, n\}$ is given by the following equations.

$$\begin{aligned}
&e_i^*(\overline{y}^{i-1}) = \overline{\Gamma}_{i,i-1}\overline{y}_{i-1}, \qquad (40)\\
&\overline{\Gamma}_{i,i-1} = -\Big(\overline{N}_{i,i-1} + \overline{B}_{i,i-1}^T\Sigma(i+1)\overline{B}_{i,i-1}\Big)^{-1} \\
&.\Big(\overline{L}_{i,i-1}^T + \overline{B}_{i,i-1}^T\Sigma(i+1)\overline{F}_{i,i-1}\Big), \quad i = 0, \ldots, n-1 \qquad (41)
\end{aligned}$$

$e_n^*(\overline{y}^{n-1}) = -\overline{N}_{n,n-1}^{-1}\overline{L}_{n,n-1}^T\overline{y}_{n-1}$, where the symmetric positive semidefinite matrix $\{\Sigma(i) : i = 0, \ldots, n\}$ satisfies a matrix difference Riccati equation, for $i = 0, \ldots, n-1$,

$$\begin{aligned}
&\Sigma(i) = \overline{F}_{i,i-1}^T\Sigma(i+1)\overline{F}_{i,i-1} - (\overline{F}_{i,i-1}^T\Sigma(i+1)\overline{B}_{i,i-1} + \overline{L}_{i,i-1}) \\
&.\Big(\overline{N}_{i,i-1} + \overline{B}_{i,i-1}^T\Sigma(i+1)\overline{B}_{i,i-1}\Big)^{-1}\Big(\overline{B}_{i,i-1}^T\Sigma_{i,i-1}\overline{F}_{i,i-1} \\
&+\overline{L}_{i,i-1}^T\Big) + \overline{M}_{i,i-1}, \quad \Sigma(n) = diag\{Q_{n,n-1}, 0\}
\end{aligned}$$

and the optimal pay-off is given by

$$J_{0,n}(e^*(\cdot),\Lambda,K_Z,\kappa,s)=\sum_{j=0}^{n}\Big\{tr\big(K_{Z_j}R_j\big)$$
$$+tr\big(\Lambda^T_{j,j-1}R_j\Lambda_{j,j-1}P_{j-1|j-1}\big)\Big\}+\sum_{j=0}^{n-1}tr\Big(K_{Y_j|Y^{j-1}}\overline{G}^T_{j,j-1}$$
$$.\Sigma(j+1)\overline{G}_{j,j-1}\Big)+\mathbf{E}\langle\overline{Y}_{-1|-1},\Sigma(0)\overline{Y}_{-1|-1}\rangle$$

(2) The optimal strategies $\{(\Lambda^*_{i,i-1},K^*_{Z_i}):i=0,\ldots,n\}$ are the solutions of the optimization problem

$$\kappa_{0,n}(C,s)\overset{\triangle}{=}\inf_{\big(\Lambda_{i,i-1},K_{Z_i}\big),i=0,\ldots,n:\frac{1}{2}\sum_{i=0}^{n}\log\frac{|K_{Y_i|Y^{i-1}}|}{|K_{V_i}|}\geq(n+1)C}\Big\{$$
$$J_{0,n}(e^*(\cdot),\Lambda,K_Z,\kappa)\Big\}.$$

Proof: (a) This follows from (4) and (9). (33) is obtained using the reconditioning property of expectation. (b) (36) follows from the dual relation (22). (i), (ii) follow from the observation that the constraint in (36) depends only on $\{\Lambda,K_Z\}$ and not on $\{e_i(\cdot):i=0,\ldots,n\}$. (c), (i). (38) follows from (27), (30), because $\{Y_i,\widehat{A}_{i|i}:i=0,\ldots,n\}$ is a sufficient statistics for the control process. The rest of the equations follows directly from the solution of partially observable stochastic optimal control problems [19]. ■

Theorem 3.1, (1) and (2) are Person-by-Person Optimality statements of $\{e_i(\cdot):i=0,\ldots,\}$ and $\{\Lambda_{i,i-1},K_{Z_i}:i=0,\ldots,n\}$.

Theorem 3.1, (c) states that the optimal input process consists of 4 strategies, follows.

$$A^g_i=\Gamma^1_{i,i-1}Y^g_{i-1}+\Gamma^2_{i,i-1}\widehat{A}_{i-1|i-1}+\Lambda_{i,i-1}A^g_{i-1}+Z^g_i.\ (42)$$

Remark 3.2: By Theorem 3.1, if $C_{i,i-1}=0,Q_{i,i-1}=0,i=0,\ldots,n$ then $e^*_i(\overline{y}^{i-1})=-\Lambda_{i,i-1}\widehat{A}_{i-1|i-1},i=0,\ldots,n$, and hence

$$A^g_i=\Lambda_{i,i-1}\Big(A^g_{i-1}-\widehat{A}_{i-1|i-1}\Big)+Z^g_i,\ i=0,\ldots,n.\quad(43)$$

That is, $\widehat{A}_{i-1|i-1},i=0,\ldots,n$ is a sufficient statistic for the strategy $e_i(Y^{g,i-1}),i=0,\ldots,n$, as expected.

Next, we discuss item Section I, (d).

Theorem 3.3: (Decentralized coding theorem)
Consider the G-RM of Theorem 3.1.
(a) If $D_{i,i-1}=0,i=0,\ldots,n$, then [2], Theorem IV.1 holds that states, directed information stability holds and separates into (i) a statement related to the ergodic properties of a stochastic optimal control problem with complete information, and (ii) a statement related to an information transmission problem.
(b) For the general G-RMs of Theorem 3.1 with $D_{i,i-1}\neq0,i=0,\ldots,n$, then (a) holds as in [2], Theorem IV.1, with some variations.

Proof: (b) This is done similar to [2], Theorem IV.1. ■

IV. Conclusions

The decentralized features of extremum problems of capacity of models with memory and feedback are illustrated. For Gaussian recursive models with past dependence on inputs and outputs it is illustrated that a semi-separation principle holds, that makes calculations and the derivation of directed information stability simpler.

References

[1] C. D. Charalambous, C. Kourtellaris, and S. Loyka, “Capacity achieving distributions & information lossless randomized strategies for feedback channels with memory: The LQG theory of directed information,” *IEEE Transactions on Information Theory*, submitted in April 2016.

[2] C. Kourtellaris and C. D. Charalambous, “Information structures of capacity achieving distributions for feedback channels with memory and transmission cost: Stochastic optimal control & variational equalities,” *IEEE Transactions on Information Theory*, Accepted in November 2017, submitted in November 2015.

[3] C. D. Charalambous, C. Kourtellaris, I. Tzortzis, and Loyka, “The capacity of unstable dynamical systems-interaction of control and information transmission,” in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 25-30, June 2017, pp. 2663–2667.

[4] C. D. Charalambous and N. U. Ahmed, “Centralized versus decentralized optimization of distributed stochastic differential decision systems with different information structures-Part I: General theory,” *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1194–1209, 2017.

[5] T. Cover and S. Pombra, “Gaussian feedback capacity,” *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.

[6] Y.-H. Kim, “Feedback capacity of stationary Gaussian channels,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 57–85, 2010.

[7] S. Yang, A. Kavcic, and S. Tatikonda, “On feedback capacity of power-constrained Gaussian noise channels with memory,” *Information Theory, IEEE Transactions on*, vol. 53, no. 3, pp. 929–954, March 2007.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[9] H. Marko, “The bidirectional communication theory–A generalization of information theory,” *IEEE Transactions on Communications*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.

[10] J. L. Massey, “Causality, feedback and directed information,” in *International Symposium on Information Theory and its Applications (ISITA ’90)*, Nov. 27-30 1990, pp. 303–305.

[11] C. D. Charalambous and P. A. Stavrou, “Directed information on abstract spaces: Properties and variational equalities,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6019–6052, 2016.

[12] P. A. Stavrou, C. D. Charalambous, and C. Kourtellaris, “Sequential necessary and sufficient conditions for capacity achieving distributions of channels with memory and feedback,” *IEEE Transactions on Information Theory*, accepted in May 2017.

[13] J. Chen and T. Berger, “The capacity of finite-state Markov channels with feedback,” *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 780–798, March 2005.

[14] H. Permuter, T. Weissman, and A. Goldsmith, “Finite state channels with time-invariant deterministic feedback,” *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.

[15] S. Tatikonda and S. Mitter, “The capacity of channels with feedback,” *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[16] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*. Springer-Verlag, 1979.

[17] C. D. Charalambous and C. Kourtellaris, “Information structures of maximizing distributions of feedback capacity for general channel with memory and applications,” *IEEE Transactions on Information Theory*, submitted, July 2016. [Online]. Available: https://arxiv.org/abs/1604.01063

[18] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes: I. General Theory*, 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[19] P. E. Caines, *Linear Stochastic Systems*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1988.

The Operational Capacity of Compound Uniformly-Ergodic Fading Channels

Sergey Loyka, Charalambos D. Charalambous

***Abstract*—The impact of distribution uncertainty on the performance of compound fading channels is studied. To this end, a new class of fading channels, termed ”uniformly-ergodic”, is introduced and several of its equivalent (easy-to-use) characterizations and examples are presented. A single-letter expression for the operational capacity of this class of channels is obtained under the full Rx CSI using the recent general formula for compound channel capacity and the information spectrum approach. The saddle-point property is established, whereby the compound channel capacity is the same as the worst-case capacity so that the full knowledge of the fading distribution at the transmitter does not increase the capacity of this class of channels.**

I. Introduction

The impact of channel uncertainty on its capacity and system design has been extensively studied since late 1950s, see [1] for an extensive literature review up to late 1990s and [2] for a more recent albeit brief review. A widely-accepted approach to the channel uncertainty problem is via the compound channel model, where the channel is assumed to be unknown but is known to belong to a certain class (set) of channels [1]. Since channel estimation is done at the receiver (Rx) and then send back to the transmitter (Tx) via a limited feedback link, many studies concentrate on limited channel state information (CSI) available at the Tx end and assume full CSI at the Rx end.

Fading represents one of the most significant obstacles to reliable wireless communications and respective system design, affecting its performance in a dramatic way [3]. It also makes channel estimation a challenging problem, due to significant channel dynamics, low SNR, limitations of a feedback link etc. In this context, incomplete CSI can also be modelled by assuming that the channel is not known but its distribution is known, the so-called channel distribution information (CDI) [3]. However, complete knowledge of the CDI, which is essential for capacity evaluation and system design, can be questioned on the same grounds as complete CSI: when only a limited sample set is available (always a practicality), the CDI can be obtained with limited accuracy only (especially at the distribution tails); limited feedback link dictates quantization of the estimated CDI before transmission, thus introducing the quantization noise; presence of noise and channel dynamics makes any estimate inaccurate to a certain degree. This motivates us to study the impact of inaccurate CDI on system performance and design.

For quasi-static (and hence non-ergodic) fading channels, the key performance metrics are outage probability/capacity [3]. The impact of CDI uncertainty on these metrics was studied in [5]. In particular, it was shown that the CDI uncertainty induces an error floor effect: increasing the SNR over a certain threshold does not reduce the outage probability and the error floor is determined by the size of the uncertainty set.

For ergodic-fading channels (where the fading process is allowed to have memory provided it is still ergodic), a single-letter capacity expression has been established in [4] under complete CDI at the Tx end and full CSI at the Rx end. However, the standard results on ergodic capacity [3][4] do not apply when only incomplete CDI is available and hence certain performance has to be demonstrated for the whole class of fading distributions, not just for a single one, and, in addition, the Tx does not know the true fading distribution and hence cannot design a codebook using this knowledge (as was done in [3][4]).

The information capacity of ergodic-fading channels under CDI uncertainty, formally defined via the standard max-min expression (of ergodic mutual information), has been studied in [6]. However, its operational meaning as the largest achievable rate subject to the reliability criterion has not been established so it is not clear whether this quantity has practical relevance (while the max-min MI is often the compound channel capacity, it is not always the case [1]). The main difficulty was the lack of general-enough tools for compound channels that would allow one to incorporate CDI uncertainty. Such tools have been recently presented in [7], which are based on the information spectrum approach of Verdu and Han [8][9]. Using these tools, we prove here that the above ”max-min” information capacity has the operational meaning of maximum achievable rate under the CDI uncertainty. This is accomplished by introducing a new concept of ”uniformly-ergodic compound channel” and applying the general formula for compound channel capacity in [7] to such channel, which results in a compact single-letter expression for the capacity of uniformly-ergodic compound channels, subject to the sets of feasible input and fading distributions being convex but otherwise arbitrary. To facilitate applications, we develop several equivalent and easy-to-use criteria for compound channels to be uniformly-ergodic and give some practically-relevant examples. Apart from the single-letter capacity expression, the key contribution of this paper is the recognition of importance of uniform ergodicity for compound fading channels.

II. Channel Model

To isolate and study the impact of CDI uncertainty, we adopt the conditionally-memoryless channel model of [4], where the

S. Loyka is with the School of EECS, University of Ottawa, Canada, sergey.loyka@ieee.org

C.D. Charalambous is with the ECE Department, University of Cyprus, Cyprus, chadcha@ucy.ac.cy

channel is memoryless conditioned on its state sequence s^n:

$$p(y^n|x^n, s^n) = \prod_{i=1}^{n} p(y_i|x_i, s_i) \tag{1}$$

where x, y, s are the input, output and state; n is the block-length, $x^n = \{x_1, .., x_n\}$ and likewise for y^n, s^n; capitals denote random variables while lower-case letters - their realizations or arguments. The random state sequence $S^n = \{S_1, .., S_n\}$ represents the fading process and is assumed to be stationary and ergodic but not necessarily memoryless - it can have memory provided that the ergodicity assumption still holds, so that a correlated fading process is allowed (see Section III for details). Assuming that the receiver has the full CSI, i.e. the state sequence s^n, but the Tx knows only the fading distribution (i.e. the full Tx CDI, see e.g. [3] for a detailed motivation of this assumption), a single-letter ergodic capacity $C[f]$ was obtained in [4] for this ergodic-fading channel:

$$C[f] = \sup_{p(x)} I(X; Y|f) \tag{2}$$

where $I(X; Y|f)$ is the ergodic mutual information under fading distribution $f(s)$ and i.i.d. input:

$$I(X; Y|f) = \sum_{s} f(s) I(X; Y|s) \tag{3}$$

and $I(X; Y|s)$ is the MI under channel state s, and where all alphabets are assumed to be discrete and finite; under some regularity assumptions, this can also be extended to infinite and continuous alphabets. The optimal input is i.i.d. [4]. The maximization over the input distribution $p(x)$ is subject to a suitable constraint, e.g. maximum or average power, and is independent of channel state s (due to no Tx CSI) but may depend on the fading distribution f. We emphasize, for future use, that the ergodic MI $I(X; Y|f)$ as well as the capacity $C[f]$ also depend on the fading distribution. Note that even though the fading process is allowed to have memory (i.e. does not have to be i.i.d.), the ergodic MI as well as the capacity depend only on the marginal fading distribution $f(s)$, not on the joint one (which is ultimately due to the conditionally-memoryless nature of the channel). This makes the analysis much simpler.

Ergodic channel model is suitable in scenarios with significant channel dynamics so that a single codeword spans many different channel realizations and an encoder can take advantage of it [3]. However, in many practical scenarios, complete knowledge of $f(s)$ may be not available at the transmitter, due to e.g.

- inaccuracy in estimating $f(s)$ at the receiver (due to finite sample size or estimation noise);
- limited/quantized feedback link (quantization noise);
- outdated estimate,

so that the true fading distribution f differs from its estimate f_0 available at the transmitter. To model this fading distribution uncertainty (inaccuracy), we consider the scenario where the transmitter has only partial CDI. Namely, it knows that $f \in \mathcal{F}_1$, where $\mathcal{F}_1$ is the uncertainty set known to the Tx, which is further assumed to be convex; the state sequence S^n is not available to the Tx, while the Rx has the full CSI, i.e. the sequence S^n. This forms a compound channel model where the fading distribution f is a (meta) state. Its respective compound channel capacity is defined in the standard way as the maximum achievable rate subject to the reliability criterion, where the error probability converges to zero uniformly over the whole uncertainty set and where the codebooks are independent of the actual channel state s or its fading distribution f (see e.g. [1] for more details and formal definitions).

The following section presents key definitions and properties of ergodic-fading channels in the compound setting, i.e. when the fading distribution is not known exactly.

III. Compound Ergodic-Fading Channels

In order to simplify notations, we use f to refer to marginal $f(s)$ as well as joint distribution $f(s^n)$, which should be clear from the context. If $\{s_1, s_2, ...\}$ is an ergodic process, we call the joint distribution $f(s^n)$ ergodic as well, with understanding that ergodicity reveals itself as $n \to \infty$. $\mathcal{F}$ denotes a set of joint distributions while $\mathcal{F}_1$ – a set of respective marginal distributions. Since the joint fading distribution completely characterises fading channel (in combination with (1)), we will refer to $\mathcal{F}$ as ”channels” as well. $\mathbb{E}\{\cdot\}$ denotes expectation over relevant random variables.

We begin with a standard definition of an ergodic (discrete-time) random process [11]-[14].

Definition 1. *A stationary random process $\{S_1, S_2, ...\}$ is (mean-) ergodic if, for any $g(s)$ such that $\mathbb{E}\{|g(S)|\} < \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} g(S_i) \to \mathbb{E}\{g(S)\} \tag{4}$$

as $n \to \infty$, where the convergence is either in mean-square, or in probability, or with probability 1.

A few modifications to this definition are in order to accommodate the compound channel setting here: (i) we need to consider a class of distributions $\mathcal{F}$ rather than a single distribution f, (ii) there is no need to consider all absolute-integrable/summable functions $g(s)$; instead, we need to consider only the mutual information $I(X; Y|s)$ under channel state s and i.i.d. input as a function of interest; (iii) we will use convergence in the mean-square sense (since it is needed in the proof of coding theorem); this implies convergence in probability but the converse is not true in general; however, when $I(X; Y|s)$ is uniformly bounded (e.g. when either input or output alphabet is of finite cardinality), they are equivalent.

The following definition extends the standard definition of ergodic channels to the compound setting.

Definition 2. *A class of stationary fading channels $\mathcal{F}$ is uniformly (mean-) ergodic if it is ergodic for each $f \in \mathcal{F}$ under i.i.d. input, i.e. as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} I(X; Y|S_i) \to I(X; Y|f) \tag{5}$$

where the convergence is in the mean-square sense, and, in addition, it is uniform over the whole class $\mathcal{F}$, i.e. $\forall\delta > 0$ $\exists n_0(\delta)$ such that $\forall n > n_0(\delta)$

$$\sigma_{nf}^2 \triangleq \mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} I(X;Y|S_i) - I(X;Y|f)\right)^2\right\} < \delta \quad (6)$$

where n_0 depends on δ but not f; δ is also independent of f.

It is straightforward to verify that (6) is equivalent to

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}}\sigma_{nf}^2 = 0 \quad (7)$$

(note that $\lim$ and $\sup$ cannot be swapped). It should be emphasized that the uniform ergodicity property in (5), (6) as well as the ergodicity property in (4) depend on the joint distribution $f(s^n)$, not just marginal $f(s)$, even though the limits depend only on the marginal.

To facilitate applications, we give equivalent criteria of the uniform ergodicity and provide several examples. To simplify notations, let $I_{s_i} = I(X;Y|S_i)$ and $I_f = I(X;Y|f)$ (all under i.i.d. inputs) and let c_{ijf} be the covariance of I_{s_i} and I_{s_j} under fading distribution f,

$$c_{ijf} \triangleq \mathbb{E}\{(I_{s_i} - I_f)(I_{s_j} - I_f)\} \quad (8)$$

Since the channel is stationary, c_{ijf} depends only on $i-j$: $c_{ijf} = c_{(i-j)f}$. We assume below that the variance is uniformly bounded:

$$c_{0f} \le A < \infty \ \forall f \in \mathcal{F} \quad (9)$$

(note that A is independent of f), which is equivalent to $\sup_{f\in\mathcal{F}} c_{0f} < \infty$. This is the case when e.g. the alphabets are discrete (see e.g. [9]) and also holds in many cases for continuous alphabets as well (e.g. Gaussian).

The following proposition is an extension of Slutsky's Theorem (see e.g. [11][13][14]) to the compound setting here.

Proposition 1. *A compound stationary-fading channel is uniformly mean-ergodic iff*

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{l=0}^{n-1}\left(1-\frac{l}{n}\right)c_{lf} = 0 \quad (10)$$

Equivalently,

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{l=0}^{n-1}c_{lf}\right| = 0 \quad (11)$$

Proof. See Appendix. □

The following condition, which follows from (11), is easier to verify in many cases.

Corollary 1.1. *The condition in* (11) *holds if $c_{lf} \to 0$ for each f as $l \to \infty$ and the convergence is uniform over the set $\mathcal{F}$:*

$$\lim_{l\to\infty}\sup_{f\in\mathcal{F}}|c_{lf}| = 0 \quad (12)$$

This condition essentially means that the channel is asymptotically uncorrelated for any possible fading distribution and also uniformly so over the uncertainty set $\mathcal{F}$.

Many special cases can be derived from (12).

1. Assume that the fading process is i.i.d. for each f, in which case $c_{lf} = 0$ for any $l \neq 0$ so that (12) holds if the variance is uniformly bounded: $c_{0f} \le A < \infty\ \forall f \in \mathcal{F}$. The condition of i.i.d. process is trivially extended to a broader condition of uncorrelated fading process.

2. An extension of the previous case is a finite-memory process: $c_{lf} = 0$ for any $|l| > L_f$, where L_f is the memory under fading distribution f, which is uniformly bounded: $L_f \le L < \infty$ for any $f \in \mathcal{F}$.

3. Infinite-memory processes are also allowed provided that the correlation decays to zero asymptotically, e.g. an exponential correlation model: $c_{lf} = c_{0f} r_f^{|l|}$, where r_f is the correlation coefficient under distribution f and $0 \le r_f \le B < 1$ for each $f \in \mathcal{F}$ (i.e. uniformly bounded away from unity), in addition to the standard requirement $c_{0f} \le A < \infty$.

4. Condition (12) is satisfied if the fading process is uniformly, asymptotically independent:

$$\lim_{l\to\infty}\sup_{f\in\mathcal{F}}\sup_{s_1,s_l}\left|\frac{f(s_1,s_l)}{f(s_1)f(s_l)} - 1\right| = 0 \quad (13)$$

i.e. $f(s_1,s_l) \to f(s_1)f(s_l)$ uniformly over $s_1, s_l, f \in \mathcal{F}$, which is equivalent to $f(s_l|s_1) \to f(s_l)$ so that the process forgets its past asymptotically (and uniformly).

5. Cases when c_{lf} does not decay to zero can be included too, e.g. $c_{lf} = (-1)^l$.

6. Any compound fading channel where each $f \in \mathcal{F}$ is ergodic and $\mathcal{F}$ is of finite cardinality is automatically uniformly-ergodic.

One can also construct examples whereby the channel is not uniformly ergodic while being ergodic for each $f \in \mathcal{F}$. Let $1 \le k < \infty$ be an integer index specifying an ergodic distribution f from the uncertainty set $\mathcal{F}$ and consider example 2 with $L_f = k$ and $c_{lf} = c_{0f} > 0$ for any $|l| \le L_f$, or example 3 with $r_f = 1 - 1/k$. In both cases, the uniform convergence condition is broken and the corresponding compound channels are not uniformly ergodic while being ergodic for each $f \in \mathcal{F}$.

IV. The Capacity of Uniformly-Ergodic Channels

Let C be the information capacity of the compound ergodic-fading channel above:

$$C \triangleq \sup_{p(x)}\inf_{f\in\mathcal{F}_1} I(X;Y|f) \quad (14)$$

Note that such $\sup - \inf$ expression appears often in the theory of compound channels and is, in many cases, the operational capacity. However, this is not the case in general [1]. It is the purpose of this section to show that this is indeed the case for the uniformly-ergodic compound channel above. First, we establish the following saddle-point property of the information capacity.

Proposition 2. *Consider the compound ergodic-fading channel in* (1)–(3), *where $f \in \mathcal{F}_1$. Assume that the set of feasible input distributions $p(x)$ is convex (e.g. average or maximum power constraint) and that $\mathcal{F}_1$ is convex. The information capacity C of this compound ergodic-fading channel satisfies the saddle-point property,*

$$C = \sup_{p(x)}\inf_{f\in\mathcal{F}_1} I(X;Y|f) = \inf_{f\in\mathcal{F}_1}\sup_{p(x)} I(X;Y|f) = C_w \quad (15)$$

where C_w is the capacity of worst-case channel in the uncertainty set, i.e. the information capacity equals to the worst-case channel capacity C_w. If the $\inf$ *and* $\sup$ *are achieved, then the following saddle-point inequalities holds for any feasible* $p(x)$ *and* $f(s)$*,*

$$I(X;Y|f^*) \le C = I(X^*;Y|f^*) \le I(X^*;Y|f) \qquad (16)$$

where X^* denotes the input under its optimal distribution $p^*(x)$ and (p^*, f^*) is a saddle point.

Proof. The saddle point property follows from the fact that $I(X;Y|f)$ is concave in $p(x)$ and linear (and thus convex) in f; since the sets of feasible f and $p(x)$ are convex, von Neumann mini-max Theorem [10] guarantees the existence of a saddle point. The saddle-point inequalities in (16) follow from 2nd equality in (15). □

The inequalities in (16) have a well-known game-theoretic interpretation: the Tx chooses $p^*(x)$ and the adversary (nature) chooses f^*; neither player can deviate from this optimal strategy without incurring a penalty.

In the rest of this section, we demonstrate that the information capacity C of the compound ergodic-fading channel has the operational meaning of a maximum achievable rate i.e. the compound channel capacity C_c for the class of uniformly-ergodic fading channels defined above.

Our approach applies to any convex uncertainty set $\mathcal{F}_1$ and also to any convex set of possible input distributions (which may also include a power constraint).

Theorem 1. *Consider a compound uniformly-ergodic fading channel. Let $\mathcal{F}_1$ be a convex set of its marginal fading distributions and $\mathcal{F}$ be a set of its joint fading distributions. Assume that the Rx has the full CSI (i.e. the state sequence s^n) while the Tx has only partial CDI: it knows $\mathcal{F}$ and hence $\mathcal{F}_1$ but neither s^n nor its fading distribution f. Let the set of feasible input distributions $p(x)$ be convex. The operational capacity C_c of this compound channel is*

$$C_c = \sup_{p(x)} \inf_{f \in \mathcal{F}_1} I(X;Y|f) = C = C_w \qquad (17)$$

i.e. the same as the worst-case channel capacity C_w in (15).

Proof. See Appendix. □

Note that, from this Theorem, (i) full knowledge of the fading distribution at the Tx does not increase the capacity, and (ii) a code designed for the worst-case fading distribution also works for the whole class of distributions (and hence much smaller amount of feedback is needed).

We remark that this result cannot be established using Theorem 3.3.5 in [9] and considering fading distribution as a meta-state since there are uncountably many possible distributions in $\mathcal{F}_1$ (since this set is continuous) while Theorem 3.3.5 requires the number of states to be finite - see [7] for details. It should also be noted that while the definition of uniform ergodic channels and the respective uncertainty set $\mathcal{F}$ as well as the error probability depend on the joint fading distribution $f(s^n)$, the capacity depends only on the marginal distribution $f(s)$ and its uncertainty set $\mathcal{F}_1$. This fact, which results in the single-letter capacity expression, cannot be inferred from Theorem 3.3.5 either. In the proof, we make use of the general formula for compound channel capacity in [7], which does not have the restrictions of Theorem 3.3.5, by applying it to the ergodic scenario of the present paper.

It can be further shown, using Fano's inequality, that the first two equalities in (17) do hold even if $\mathcal{F}_1$ is not convex [15][1]. However, the saddle point property and hence the last equality do not need to hold in this case.

V. Appendix

Proof of Proposition 1: It is straightforward to verify (by direct computations) that

$$\sigma_{nf}^2 = \frac{1}{n} \sum_{l=1-n}^{n-1} \left(1 - \frac{|l|}{n}\right) c_{lf} = \frac{2}{n} \sum_{l=0}^{n-1} \left(1 - \frac{l}{n}\right) c_{lf} - \frac{1}{n} c_{0f}$$

(since $c_{(-l)f} = c_{lf}$) and that $\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \sigma_{nf}^2 = 0$ if and only if (10) holds provided that c_{0f} is uniformly bounded: $c_{0f} \le A < \infty \ \forall f \in \mathcal{F}$.

To establish (11), observe that

$$\left| \frac{1}{n} \sum_{l=0}^{n-1} c_{lf} \right| = \left| \mathbb{E} \left\{ \left(\frac{1}{n} \sum_{l=1}^{n} I_{s_l} - I_f \right) (I_{s_1} - I_f) \right\} \right| \le \sigma_{nf} \sqrt{c_{0f}} \qquad (18)$$

where the inequality follows from Cauchy-Schwartz inequality, so that (11) follows provided that c_{0f} is uniformly bounded. This establishes the "only if" part.

To establish the "if" part of (11), let

$$z_{nf} = \frac{1}{n} \sum_{l=0}^{n-1} \left(1 - \frac{l}{n}\right) c_{lf} = \frac{1}{n^2} \sum_{l=0}^{n-1} \sum_{i=1}^{n-l} c_{lf} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{l=0}^{n-i} c_{lf}$$

Observe that (11) implies that for any $\delta > 0$ there exists such $n_0(\delta)$ that for any $n \ge n_0(\delta)$

$$\left| \frac{1}{n} \sum_{l=0}^{n-1} c_{lf} \right| \le \delta \ \forall f \in \mathcal{F} \qquad (19)$$

Set $n \ge n_0^2(\delta)$ and let $L_n = n - \sqrt{n}$ (round off if not integer) so that

$$\begin{aligned} z_{nf} &= \frac{1}{n^2} \sum_{i=1}^{L_n} \sum_{l=0}^{n-i} c_{lf} + \frac{1}{n^2} \sum_{i=L_n+1}^{n} \sum_{l=0}^{n-i} c_{lf} \\ &\le \frac{1}{n} \sum_{i=1}^{L_n} \left| \frac{1}{n-i} \sum_{l=0}^{n-i} c_{lf} \right| + \frac{1}{n^2} \sum_{i=L_n+1}^{n} \sum_{l=0}^{n-i} c_{0f} \qquad (20) \\ &\le \frac{1}{n} \sum_{i=1}^{L_n} \delta + \frac{(n-L_n)^2}{n^2} c_{0f} \le \delta + \frac{c_{0f}}{n} \end{aligned}$$

where 1st inequality is from $c_{lf} \le c_{0f}$ and 2nd one is from $n - i \ge n - L_n \ge n_0(\delta)$. Since $\delta > 0$ is arbitrary, $z_{nf} \ge 0$ and c_{0f} is uniformly bounded, the "if" part follows by taking $\lim_{n\to\infty} \sup_{f\in\mathcal{F}}$:

$$0 \le \lim_{n\to\infty} \sup_{f\in\mathcal{F}} z_{nf} \le \delta \ \forall \delta > 0 \qquad (21)$$

[1] The authors greatly appreciate the very insightful comments by an anonymous reviewer.

□

Proof of Theorem 1: Let $X^n = \{X_1...X_n\}$, $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ and likewise for $\boldsymbol{Y}$. Following Theorem 5 in [7], the capacity of general compound channels (e.g. not necessarily information-stable and where the uncertainty set can be arbitrary) with full Rx CSI but no Tx CSI is given by

$$C_c = \sup_{\boldsymbol{X}} \underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) \tag{22}$$

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,

$$\underline{\underline{I}}(\boldsymbol{X};\boldsymbol{Y}) = \sup_{R} \left\{ R : \lim_{n\to\infty} \sup_{s\in\mathcal{S}} \Pr\{Z_{ns} \le R\} = 0 \right\} \tag{23}$$

where $Z_{ns} = n^{-1} i(X^n;Y^n|s)$ is the normalized information density under channel state s; $\mathcal{S}$ is the (arbitrary) uncertainty set.

To prove (17), first observe that $C_c \le C_w$ holds in full generality[2] and, using (15),

$$C_c \le C_w = C = \sup_{p(x)} \inf_{f\in\mathcal{F}_1} I(X;Y|f) \tag{24}$$

It remains to show that the inequality is actually equality. To this end, apply the general formula in (22) by considering the fading distribution f as a (meta) state s, and restrict the optimization to i.i.d. inputs $\tilde{\boldsymbol{X}}$ to obtain a lower bound

$$C_c \ge \sup_{\tilde{\boldsymbol{X}}} \underline{\underline{I}}(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}}) \tag{25}$$

where $\tilde{\boldsymbol{Y}}$ is the output under i.i.d. input $\tilde{\boldsymbol{X}}$. The following propositions evaluate $\underline{\underline{I}}(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}})$.

Proposition 3. *The compound inf-information rate* $\underline{\underline{I}}(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}})$ *can be upper bounded as follows:*

$$\underline{\underline{I}}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}) \le \inf_{f\in\mathcal{F}_1} I(X;Y|f) \tag{26}$$

where X has the same distribution as the marginals of $\tilde{\boldsymbol{X}}$.

Proof. Let $I_f = I(X;Y|f)$, $i_k = i(X_k;Y_k|S_k)$, $I_{s_k} = \mathbb{E}_{X,Y}\{i_k\}$, $z_n = n^{-1}\sum_{k=1}^{n} i_k$. From Proposition 1 in [7],

$$\underline{\underline{I}}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}) \le \inf_{f\in\mathcal{F}} \underline{I}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}|f) \tag{27}$$

where $\underline{I}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}|f)$ is the inf-information rate under (meta) state f:

$$\underline{I}(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}}|f) = \sup_{R} \left\{ R : \lim_{n\to\infty} \Pr\{z_n \le R\} = 0 \right\} \tag{28}$$

Note that $\mathbb{E}\{i_k\} = I_f$ and

$$\mathbb{E}\left\{|z_n - I_f|^2\right\} = \frac{1}{n^2}\sum_{k,l} \mathbb{E}\{(I_{s_k} - I_f)(I_{s_l} - I_f)\} = \sigma_{nf}^2$$

where 1st equality follows from the fact that (X_i, Y_i) and (X_j, Y_j) are independent of each other ($i \ne j$) given the state sequence $\{s_1, s_2, ...\}$, so that, from Chebychev inequality,

$$\Pr\{|z_n - I_f| \ge \delta\} \le \sigma_{nf}^2/\delta^2 \to 0 \ \forall\ f \in \mathcal{F} \tag{29}$$

for any $\delta > 0$ as $n \to \infty$ since, due to (7), $\lim_{n\to\infty}\sigma_{nf} = 0$. Therefore, $z_n = n^{-1}\sum_{k=1}^{n} i_k \to I_f$ in probability, from which it follows that

$$\underline{I}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}|f) = I(X;Y|f) \tag{30}$$

Combining this with (27), one obtains (26). □

Proposition 4. *The compound inf-information rate* $\underline{\underline{I}}(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}})$ *can be lower bounded as follows:*

$$\underline{\underline{I}}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}) \ge \inf_{f\in\mathcal{F}_1} I(X;Y|f) \tag{31}$$

Proof. Observe that, for each $\delta > 0$,

$$\Pr\left\{ z_n \le \inf_{f\in\mathcal{F}_1} I_f - \delta \right\} \le \Pr\{|z_n - I_f| \ge \delta\} \le \frac{\sigma_{nf}^2}{\delta^2}$$

Applying $\lim - \sup$ and using (7), one obtains

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \Pr\left\{ z_n \le \inf_{f\in\mathcal{F}_1} I_f - \delta \right\} = 0 \tag{32}$$

i.e. $\underline{\underline{I}}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}) \ge \inf_{f\in\mathcal{F}_1} I_f - \delta$. Since this holds for any $\delta > 0$, (31) follows. □

Combining Propositions 3 and 4,

$$\underline{\underline{I}}(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}) = \inf_{f\in\mathcal{F}_1} I(X;Y|f) \tag{33}$$

for any i.i.d. input. Applying $\sup_{\tilde{\boldsymbol{X}}}$ to this equality in combination with (24) and (25), one obtains the desired result.

References

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

[2] S. Loyka, C.D. Charalambous, Novel Matrix Singular Value Inequalities and Their Applications to Uncertain MIMO Channels, IEEE Trans. Info. Theory, v.61, n.12, pp. 6623 - 6634, Dec. 2015.

[3] 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.

[4] G. Caire, S. Shamai, On The Capacity of Some Channels With Channel State Information, IEEE Transactions on Information Theory, vol. 45, no. 6, pp. 2007–2019, Sep. 1999.

[5] I. Ioannou, C.D. Charalambous, S. Loyka, ”Outage Probability Under Channel Distribution Uncertainty”, *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6825-6838, Nov. 2012.

[6] S. Loyka, C. D. Charalambous, Ergodic Capacity Under Channel Distribution Uncertainty, 52nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, Oct. 1-3 2014.

[7] S. Loyka, C.D. Charalambous, A General Formula for Compound Channel Capacity, IEEE Transactions on Information Theory, v. 62, no.7, pp.3971-2991, July 2016.

[8] 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.

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

[10] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[11] A.M. Yaglom, An Introduction to the Theory of Stationary Random Functions, Uspehi Matematicheskih Nauk, v. 7, no. 5(51), pp. 3–168, 1952.

[12] D.R. Brillinger, Time Series: Data Analysis and Theory, Holt, Rinehart and Winston Inc., New York, 1975.

[13] S.M. Rytov, An Introduction to Statistical Radio Physics, v.1: Stochastic Processes, Nauka, Moscow, 1976.

[14] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw Hill, Boston, 2002.

[15] A Private Communication by An Anonymous Reviewer, 2017.

[2] the compound capacity never exceeds the worst-case one since a code that works for the whole uncertainty set has also to work on the worst-case channel in the set [1].

Strong Converse Bounds for High-Dimensional Estimation

Ramji Venkataramanan
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ, United Kingdom
Email: ramji.v@eng.cam.ac.uk

Oliver Johnson
School of Mathematics
University of Bristol
Bristol, BS8 1TW, United Kingdom
Email: O.Johnson@bristol.ac.uk

Abstract

In statistical inference problems, we wish to obtain lower bounds on the minimax risk, that is to bound the performance of any possible estimator. A standard technique to obtain risk lower bounds involves the use of Fano's inequality. In an information-theoretic setting, it is known that Fano's inequality typically does not give a sharp converse result (error lower bound) for channel coding problems. Moreover, recent work has shown that an argument based on binary hypothesis testing gives tighter results. We adapt this technique to the statistical setting, and argue that Fano's inequality can always be replaced by this approach to obtain tighter lower bounds that can be easily computed and are asymptotically sharp. We illustrate our technique in three applications: density estimation, active learning of a binary classifier, and compressed sensing, obtaining tighter risk lower bounds in each case.

Horizont Independent MDL

Peter Harremoës
Niels Brock Copenhagen Business College
Copenhagen, Denmark
Email: harremoes@ieee.org

Abstract

In the minimum description length (MDL) approach to prediction, one may use the conditional normalized maximum-likelihood predictor to predict the future given the past. This strategy is, however, computationally involved and in general, depending on how many future symbols one wants to predict. For special exponential family models, the conditional normalized maximum-likelihood predictor does not depend on the number of symbols that one wants to predict. In this case, the prediction strategy equals a Bayesian strategy based on Jeffreys' prior. These special exponential families can be characterized as those for which the conjugated exponential family has a saddle-point approximation that is exact after renormalization. For 1-dimensional exponential families, the only families with exact renormalized saddle-point approximations are the Gaussian location family, the Gamma family, and the inverse Gaussian family. They are conjugated families of the Gaussian location family, the Gamma family and a less familiar family that we will call the Poisson-exponential family. This approach can be also used to construct exponential families with horizont independent MDL in higher dimensions.

Universal Batch Learning — Information Theoretical View

Meir Feder
School of Electrical Engineering
Tel-Aviv University
Tel-Aviv, Israel
Email: meir@eng.tau.ac.il

Abstract

Universal learning with log-loss discussed in this talk follows information theoretical concepts of universal prediction and universal compression. However, the fact that in learning problems data features are given and the goal is to predict the outcome, requires an extension of the theory. In previous work we analyzed on-learning, so in this talk we focus on universal batch learning. In the stochastic setting we propose a minimax universal learning solution that minimizes the worst case log-loss regret. The resulting universal learning solution is a mixture over the models in the considered class. Utilizing the minimax theorem and information-theoretical tools, we also come up with a redundancy capacity theorem and an upper bound on the performance of the optimal solution. This performance bound on the generalization error decays as $O(\log N/N)$, where N is the sample size, instead of $O\left(\sqrt{\log N/N}\right)$ that I attained in statistical learning theory. Finally, we propose a setting for universal batch learning in the individual setting, based on the leaving-one-out (LOO) principle, and show its performance in some batch learning examples.

Relations among the Minimum Error Probability, Guessing Moments, and Arimoto-Rényi Conditional Entropy

Igal Sason
Andrew and Erna Viterbi Faculty
of Electrical Engineering
Technion — Israel Institute of Technology
Technion City, Haifa 3200003, Israel
Email: sason@ee.technion.ac.il

Sergio Verdú
Department of Electrical Engineering
Princeton University
Pinceton, New Jersey 08544, USA
Email: verdu@princeton.edu

Abstract

This talk presents upper and lower bounds on the minimum error probability of Bayesian M-ary hypothesis testing in terms of the Arimoto-Rényi conditional entropy of an arbitrary order α. The improved tightness of these bounds over their specialized versions with the Shannon conditional entropy ($\alpha = 1$) is explained. In particular, in the case where M is finite, we generalize Fano's inequality under both the conventional and list-decision settings. As a counterpart to the generalized Fano's inequality, allowing M to be infinite, a lower bound on the Arimoto-Rényi conditional entropy is derived as a function of the minimum error probability. We further provide upper and lower bounds on the optimal guessing moments of a random variable taking values on a finite set when side information may be available. These moments quantify the number of guesses required for correctly identifying the unknown object and, similarly to Arıkan's bounds, they are expressed in terms of the Arimoto-Rényi conditional entropy. Although Arıkan's bounds are asymptotically tight, the improvement of the bounds in this paper is significant in the non-asymptotic regime. Relationships between moments of the optimal guessing function and the MAP error probability are also presented, characterizing the exact locus of their attainable values.

Rates of linear codes based on bipartite graphs with low decoding error probability

Ghurumuruhan Ganesan

New York University, Abu Dhabi

***Abstract*—Consider binary linear codes obtained from bipartite graphs as follows. There are $k \geq 1$ left nodes each representing a message bit and there are $m = m(k)$ right nodes each representing a parity bit, generated from the corresponding set of message node neighbours. Both the message and the parity bits are sent through a memoryless binary input channel that either retains, flips or erases each transmitted bit, independently. Based on the received set of symbols, the decoder at the receiver obtains an estimate of the original message sent. If the decoding error probability $P_k \longrightarrow 0$ and the average degree per parity node remains bounded as $k \to \infty$, then the rate of the code $\frac{k}{k+m} \longrightarrow 0$ as $k \to \infty$.**

Key words: Linear codes, low decoding error probability, asymptotic rates.

AMS 2000 Subject Classification: Primary: 60J10, 60K35; Secondary: 60C05, 62E10, 90B15, 91D30.

I. Introduction

Parity check codes are used extensively in today's communication systems particularly in the form of Low Density Parity Check (LDPC) Codes (see [3] for an introduction). One of the main challenges here is to achieve low decoding error probability. There exists extensive literature on LDPC codes and previous literature mainly focus on developing decoding schemes that achieve low error probability (see for example [2], [7] and references therein). The emphasis there is to design schemes that achieve low error probability but possibly at the cost of increased overhead.

Performance tradeoffs with regards to density of LDPC codes for a fixed rate have been studied in [4] and [6]. In this paper, we study the rate versus decoding error probability tradeoff and show that low decoding error probability necessarily requires a low rate or equivalently a large number of parity bits to be appended to the message. In other words, if the decoder is such that the asymptotic decoding error probability converges to zero as the number of message bits $k \to \infty$, then the asymptotic encoded rate also converges to zero as $k \to \infty$.

Model description

We are interested in sending a random message through a communication channel reliably. We describe the underlying communication system below.

Messages: Messages are $k-$bit vectors satisfying the following condition:
$(A1)$ A random message $X = (X_1, \ldots, X_k)$ has independent and identically distributed (i.i.d.) bits $X_i \in \{0,1\}$ with

$$\mathbb{P}(X_i = 0) = p_x = 1 - \mathbb{P}(X_i = 1). \tag{1}$$

In particular, this implies that the *raw rate* defined as

$$R_{raw} := \frac{H(X)}{k} = H(p_x) > 0, \tag{2}$$

where $H(p_x) = -p_x \log p_x - (1-p_x)\log(1-p_x)$ and

$$H(X) := -\sum_{w} p(w) \log p(w) \tag{3}$$

is the entropy of the vector X (see Chapter 1, Section 1.1 of [5]). In (3), $p(.)$ is the probability mass function of X and the summation is over all possible $k-$bit vectors. All logarithms are to the base 2 and for simplicity we assume throughout that $p_x = \frac{1}{2}$ so that $H(p_x) = 1$.

Encoder: We consider binary linear codes obtained from bipartite graphs as follows. There are $k \geq 1$ left nodes called *message nodes* and there are $m = m(k)$ right nodes called *parity nodes*. For parity node $1 \leq j \leq m$, let $R_k(j)$ be the message nodes adjacent to j. The j^{th} parity bit Z_j is obtained as

$$Z_j = \oplus_{w \in R_k(j)} X_w, \tag{4}$$

where $\oplus$ is XOR operation, i.e., addition modulo 2. The vector

$$(X, Z_1, \ldots, Z_m) = (X_1, \ldots, X_k, Z_1, \ldots, Z_m)$$

is the *codeword* associated with the message X and the *encoded rate* is defined as

$$R_{enc} := \frac{H(X)}{k+m} = \frac{k}{k+m}, \tag{5}$$

by (2). We make the following assumption regarding the encoder:
$(A2)$ For $1 \leq j \leq m$, let $\#R_k(j)$ be the degree of the parity node j and suppose that the average degree per parity node remains bounded as $k \to \infty$; i.e.,

$$\limsup_k \frac{1}{m} \sum_{j=1}^{m} \#R_k(j) < \infty. \tag{6}$$

Channel: The codeword $(X, Z_1, \ldots, Z_m)$ is sent through a binary input channel which introduces noise that either retains, flips or erases the transmitted bit. Formally, we assume that the noise alphabet is $\{\alpha_0, \alpha_1, \alpha_{er}\}$ and the i^{th} received message symbol is

$$\begin{aligned}\tilde{X}_i &= 1\!\!1(N_x(i) = \alpha_{er})\alpha_{er} + 1\!\!1(N_x(i) = \alpha_0)X_i \\ &+ 1\!\!1(N_x(i) = \alpha_1)(1 - X_i) \qquad (7)\end{aligned}$$

Here α_{er} is the erasure symbol and $\tilde{N}_x(i)$ is the noise symbol. Similarly, the j^{th} received parity symbol is

$$\begin{aligned}\tilde{Z}_j &= \alpha_{er}1\!\!1(N_z(j) = \alpha_{er}) + 1\!\!1(N_z(j) = \alpha_0)Z_j \\ &+ 1\!\!1(N_z(j) = \alpha_1)(1 - Z_j). \qquad (8)\end{aligned}$$

The overall received codeword is

$$Y = (\tilde{X}_1, \ldots, \tilde{X}_k, \tilde{Z}_1, \ldots, \tilde{Z}_m). \qquad (9)$$

$(A3)$ We assume that the noise random variables $\{N_x(i)\}$ and $\{N_z(j)\}$ are independent and identically distributed (i.i.d.) with

$$\begin{aligned}&\mathbb{P}(N_x(i) = \alpha_{er}) = p_{er}, \mathbb{P}(N_x(i) = \alpha_1) = p_1 \\ &\text{and } \mathbb{P}(N_x(i) = \alpha_0) = 1 - p_1 - p_{er}. \qquad (10)\end{aligned}$$

The term $0 < p_{er} + p_1 < 1$ is the probability that a channel error occurs; i.e., the noise in the channel corrupts (either erases or flips) a transmitted bit. We also assume that the noise is independent of the transmitted bits $\{X_i\}$ and $\{Z_j\}$.

The above channel model is general and with particular choices of α_0, α_1 and α_{er}, we realize various channels. For example if $\alpha_0 = 0, \alpha_1 = 1$ and $p_0 = p = 1 - p_1$, then we obtain the binary symmetric channel (BSC). Similarly if $p_0 = p = 1 - p_{er}$, we obtain the binary erasure channel (BEC).

Decoder: At the receiver, a pre installed decoder uses the received word Y to obtain an estimate $\hat{X}$ of the message sent and let

$$P_k = \mathbb{P}(X \neq \hat{X}) \qquad (11)$$

be the decoding error probability. The following is the main result of this paper.

Theorem 1. *Suppose assumptions $(A1) - (A3)$ hold. If the decoding error probability $P_k \longrightarrow 0$ as $k \to \infty$, then the encoded rate $R_{enc} = \frac{k}{k+m} \longrightarrow 0$ as $k \to \infty$.*

In other words, search for codes with positive rate and low decoding error probability must be outside the set of linear bounded degree codes as described above. Equivalently, linear bounded degree codes having low decoding error probability must necessarily contain a lot of parity bits. One example of such a code is the $r-$repetition code, where each message bit is simply repeated r times. Recall that for a fixed r, an $r-$repetition code has an encoded rate of $\frac{1}{r+1}$ and using majority decision rule, it is possible to correct up to $\frac{r-1}{2}$ channel errors, irrespective of the number of bits k in the message (for more on repetition codes see [1]). If however, we allow $r = r(k)$ to depend on k, we can correct all errors in the message with high probability.

Proposition 1. *Suppose $2p_1 + p_{er} < 1$ and $r = r(k) = M \log k$. There are constants $M_0 = M_0(p_1, p_{er}) \geq 1$ and $K_0 = K_0(p_1, p_{er}) \geq 1$ so that the following holds for all $M \geq M_0$ and $k \geq K_0$: For an $r-$repetition code, the decoding error probability with the majority decision rule is bounded above by $P_k \leq \frac{1}{k}$.*

The paper is organized as follows. In Section I, we prove Theorem 1 and Proposition 1.

PROOF OF THEOREM 1 AND PROPOSITION 1

Recall that X is the message and Y as defined in (9) is the received codeword. Define

$$H(X|Y) := -\sum p(x,y) \log p(x|y) \qquad (12)$$

to be the uncertainty in X given the random vector Y, where $p(x,y)$ and $p(x|y)$ respectively, refer to probability mass functions of the joint distribution of (X,Y) and the conditional distribution of X given Y (see Chapter 1, [5]). Since the total number of messages is 2^k, we have from Fano's inequality (Theorem 2.10.1, [5]) that

$$H(X|Y) \leq H(X|\hat{X}) \leq H(P_k) + P_k \log\left(2^k - 1\right) \leq 1 + kP_k$$

and so

$$\frac{1}{k}H(X|Y) \leq \frac{1}{k} + P_k \longrightarrow 0 \qquad (13)$$

as $k \to \infty$.

To evaluate $H(X|Y)$, let $X_0 = 0$ and write

$$\begin{aligned}H(X|Y) &= \sum_{i=1}^{k} H(X_i|Y, X_1, \ldots, X_{i-1}) \\ &\geq \sum_{i=1}^{k} H(X_i|\tilde{X}_i, \{\tilde{Z}_j\}, \{X_w\}_{w \neq i}). \qquad (14)\end{aligned}$$

The first equality in (14) follows by chain rule for entropy (Theorem 2.5.1, [5]) and the inequality in (14) follows from the data processing inequality (Theorem 2.8.1, [5]).

We evaluate each term in the summation in (14) separately. First, we use the received parity symbols $\tilde{Z}_1, \ldots, \tilde{Z}_m$ to obtain estimates for the i^{th} transmitted bit X_i. Formally, for $1 \leq i \leq k$ let $T_k(i)$ denote the set of parity nodes adjacent to the message node i. Recall that for $u \in T_k(i)$, the term $R_k(u)$ denotes the set of message nodes adjacent to the parity node u and by definition $i \in R_k(u)$. For $1 \leq i \leq k$, define

$$\begin{aligned}\hat{X}_i(u) &:= \alpha_{er}1\!\!1(\tilde{Z}_u = \alpha_{er}) \\ &+ 1\!\!1(\tilde{Z}_u \neq \alpha_{er})\tilde{Z}_u \oplus_{w \in R_k(u)\setminus\{i\}} X_w \\ &= \alpha_{er}1\!\!1(N_z(u) = \alpha_{er}) + 1\!\!1(N_z(u) = \alpha_0)X_i \\ &+ 1\!\!1(N_z(u) = \alpha_1)(1 - X_i). \qquad (15)\end{aligned}$$

Equation (15) follows from the expression for $\tilde{Z}_u$ in (8) and the fact that if $Z_u = X_i \oplus_{w \in R_k(u)\setminus\{i\}} X_w$, then

$$1 - Z_u = (1 - X_i) \oplus_{w \in R_k(u)\setminus\{i\}} X_w.$$

The map

$$\left(\tilde{X}_i, \{\tilde{Z}_j\}, \{X_w\}_{w\neq i}\right) :\longrightarrow$$
$$\left(\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)}, \{\tilde{Z}_j\}_{j\notin T_k(i)}, \{X_w\}_{w\neq i}\right)$$

is one to one and invertible and so the i^{th} term in the final summation in (14) is

$$\begin{aligned} &H(X_i|\tilde{X}_i, \{\tilde{Z}_j\}, \{X_w\}_{w\neq i}) = \\ &H\left(X_i|\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)}, \{\tilde{Z}_j\}_{j\notin T_k(i)}, \{X_w\}_{w\neq i}\right). \end{aligned} \tag{16}$$

The set of random variables $(\{\tilde{Z}_j\}_{j\notin T_k(i)}, \{X_w\}_{w\neq i})$ are independent of the rest of random variables $(\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)})$ and are also independent of X_i. Thus the final term in (16) is

$$\begin{aligned} &H\left(X_i|\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)}, \{\tilde{Z}_j\}_{j\notin T_k(i)}, \{X_w\}_{w\neq i}\right) \\ &= H\left(X_i|\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)}\right) \end{aligned} \tag{17}$$

and substituting this into (14) gives

$$H(X|Y) \geq \sum_{i=1}^{k} G(d_k(i)) \tag{18}$$

where $d_k(i) := \#T_k(i)$ is the degree of the message node i and

$$G(d_k(i)) := H\left(X_i|\tilde{X}_i, \{\hat{X}_i(u)\}_{u\in T_k(i)}\right) > 0$$

is the uncertainty in the bit X_i given $d_k(i)+1$ independently noise corrupted copies.

We have the following properties regarding $G(.)$.
$(g1)$ Using the fact that conditioning reduces entropy, we obtain that $G(d)$ is a decreasing function of d.
$(g2)$ Using (18) and (13) we get that

$$\frac{1}{k}\sum_{i=1}^{k} G(d_k(i)) \longrightarrow 0 \tag{19}$$

as $k \to \infty$.

We use properties $(g1)-(g2)$ to get the following properties.
$(g3)$ The average degree per message node

$$\frac{1}{k}\sum_{i=1}^{k} d_k(i) \longrightarrow \infty \tag{20}$$

as $k \to \infty$.
$(g4)$ The encoded rate $\frac{k}{k+m} \longrightarrow 0$ as $k \to \infty$.
This proves Theorem 1.

Proof of $(g3)-(g4)$: We prove $(g3)$ first. For integer $q \geq 1$, let

$$S_k(q) = \{i : d_k(i) \leq q\} \tag{21}$$

be the set of message nodes whose degree is at most q. For a fixed q, it is true that

$$\frac{\#S_k(q)}{k} \longrightarrow 0 \tag{22}$$

as $k \to \infty$. If (22) is not true, then there exists $\epsilon_0 > 0$ and a subsequence $\{k_r\}$ such that $\frac{\#S_{k_r}(q)}{k_r} \geq \epsilon_0$ for all large r. Using property $(g1)$ that $G(.)$ is decreasing, we get that

$$\begin{aligned} \frac{1}{k_r}\sum_{i=1}^{k_r} G(d_{k_r}(i)) &\geq \frac{1}{k_r}\sum_{i\in S_{k_r}(q)} G(d_{k_r}(i)) \\ &\geq G(q)\frac{\#S_{k_r}(q)}{k_r} \\ &\geq \epsilon_0 G(q) \end{aligned} \tag{23}$$

for all large r. The final term in (23) is positive, contradicting (19) in property $(g2)$.

From the above paragraph, we obtain that (22) is true and so for any integer $q \geq 1$, we get that

$$\frac{1}{k}\sum_{i=1}^{k} d_k(i) \geq \frac{1}{k}\sum_{i\notin S_k(q)} d_k(i) \geq q\left(\frac{k - S_k(q)}{k}\right) \geq \frac{q}{2}$$

for all large k. Since $q \geq 1$ is arbitrary, we get (20).

To prove $(g4)$, we use the fact that the number of edges in the graph is

$$\sum_{i=1}^{k} d_k(i) = \sum_{j=1}^{m} f_k(j)$$

where $f_k(j) = \#R_k(j)$ is the degree of the parity node j. Using $(g3)$, we therefore get

$$\frac{1}{k}\sum_{j=1}^{m} f_k(j) = \frac{m}{k}\frac{1}{m}\sum_{j=1}^{m} f_k(j) \longrightarrow \infty \tag{24}$$

as $k \to \infty$. Since by assumption, the average degree per parity node is bounded (see (6)) we get from (24) that $\frac{m}{k} \longrightarrow \infty$ and so $\frac{k}{k+m} \longrightarrow 0$ as $k \to \infty$. ■

Proof of Proposition 1: Let $X = (X_1, \ldots, X_k)$ be the message bits. For $1 \leq i \leq k$ and $1 \leq j \leq r$, define $Z_i(j) = X_i$ be the parity bits for the message bit X_i. Thus each message bit is repeated r times and for convenience define $Z_i(0) = X_i$ to be the message bit to be transmitted. Let $\{\tilde{Z}_j(i)\}$ be corresponding received symbols as defined in (8).

The decoding is majority based as follows. For each $1 \leq i \leq k$ and $l \in \{0,1\}$, let $W_l(i) \subseteq \{0,1,2,\ldots,r\}$ be the random set of all indices for which the received symbol is l; i.e.,

$$\tilde{Z}_i(j) = 0 \text{ for all } j \in W_0(i) \text{ and } \tilde{Z}_i(j) = 1 \text{ for all } j \in W_1(i).$$

If $\#W_1(i) \geq \#W_0(i)$, set $\hat{X}_i = 1$; else set $\hat{X}_i = 0$. The estimated message is $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_k)$.

A decoding error occurs if $\hat{X}_i \neq X_i$ for some $1 \leq i \leq r$. For a fixed $1 \leq i \leq k$ and $0 \leq j \leq r$, let $N_z(i,j) \in \{0,1,\alpha\}$ be the noise random variable affecting the bit $Z_i(j)$ as in (8). If message bit i is decoded wrongly then necessarily

$$\sum_{j=0}^{r} \mathbb{1}(N_z(i,j) = 1) \geq \sum_{j=0}^{r} \mathbb{1}(N_z(i,j) = 0).$$

Defining

$$L(i,j) = \mathbb{1}(N_z(i,j)=1) - \mathbb{1}(N_z(i,j)=0) \in \{-1,1\}$$

we have that

$$\mathbb{E}L(i,j) = p_1 - (1-p_1-p_{er}) = 2p_1 + p_{er} - 1 < 0,$$

by the assumption in the statement of the Proposition.

For a fixed $1 \leq i \leq k$, the random variables $\{L(i,j)\}_{0\leq j\leq r}$ are i.i.d and so using the Chernoff bound, we have for $s > 0$ and $c \geq 0$ that

$$\begin{aligned} \mathbb{P}\left(\sum_{j=0}^{r} L(i,j) \geq c\right) &\leq e^{-sc}\prod_{j=0}^{r}\mathbb{E}e^{sL(i,j)} \\ &= e^{-sc}\left(e^{s}p_1 + e^{-s}(1-p_1-p_\alpha)\right)^{r+1}. \end{aligned} \tag{25}$$

Writing $e^s = 1 + s + R_1(s)$ and $e^{-s} = 1 - s + R_2(s)$, we have

$$p_1e^s + e^{-s}(1-p_1-p_{er}) = 1 - (1-2p_1-p_{er})s + T(s),$$

where $T(s) = R_1(s)p_1 + R_2(s)(1-p_1-p_{er})$. Choosing $s > 0$ small, we have $|T(s)| \leq s^2$ and $1-(1-2p_1-p_{er})s+T(s) \leq \delta$ for some constant $\delta < 1$. Substituting into (25) and setting $c = 0$ gives

$$\mathbb{P}\left(\sum_{j} L(i,j) \geq 0\right) \leq \delta^{r+1} \leq \frac{1}{k^2}$$

if $r = \frac{2}{\delta}\log k$. But $\sum_j L(i,j) \geq 0$ only if the bit X_i is decoded wrongly i.e., $\hat{X}_i \neq X_i$ and so $\mathbb{P}(\hat{X}_i \neq X_i) \leq \frac{1}{k^2}$ and so the overall decoding error probability is at most $\frac{1}{k} \longrightarrow 0$ as $k \to \infty$. ■

Acknowledgements

I thank Professors Rahul Roy, Federico Camia, Wolfgang Gerstacker and the referee for crucial comments that led to an improvement of the paper. I also thank Professors Rahul Roy and Federico Camia for my fellowships.

REFERENCES

[1] Martin Bossert, *Channel coding for Telecommunications*. Wiley, 1999.

[2] M. Luby, M. Mitzenmacher, A. Shokrollahi and D. Spielman, Efficient erasure correcting codes. *IEEE Transactions on Information Theory*, **47**, pp. 569–584, 2001.

[3] A. Shokrollahi, LDPC codes: An introduction. *Link: https://www.ics.uci.edu/ welling/teaching/ICS279/LPCD.pdf*, 2003.

[4] I. Sason and R. Urbanke, Parity-check density versus performance of binary block codes over memoryless symmetric channels. *IEEE Transactions on Information Theory*, **7**, pp. 1611–1635, 2003.

[5] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Interscience, 2006.

[6] G. Wiechman and I. Sason, Parity-check density versus performance of binary linear block codes: New bounds and applications. *IEEE Transactions on Information Theory*, **59**, pp. 1505–1516, 2007.

[7] G. Liva, E. Paolini, Balasz Matuz and M. Chiani, A decoding algorithm for LDPC codes over erasure channels with sporadic errors. 48^{th} *Annual Allerton Conference on Communication, Control and Computing*, November 2010.

A Lower Bound on the Error Exponent of Random Gilbert-Varshamov Codes

Anelia Somekh-Baruch
Bar-Ilan University
somekha@biu.ac.il

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

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

We consider transmission over a discrete memoryless channel (DMC) $W(y|x)$ with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. We consider an (n, M_n)-codebook $\mathcal{M}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M_n}\}$ with rate $R_n = \frac{1}{n}\log M_n$. The type-dependent maximum-metric decoder estimates the transmitted message as

$$\hat{m} = \operatorname*{arg\,max}_{\boldsymbol{x}_i \in \mathcal{M}_n} q(\hat{P}_{\boldsymbol{x}_i,\boldsymbol{y}}), \tag{1}$$

where $\hat{P}_{\boldsymbol{x},\boldsymbol{y}}$ is the joint empirical distribution [1, Ch. 2] of the pair $(\boldsymbol{x}, \boldsymbol{y})$ and the metric $q : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is continuous. Maximum-likelihood (ML) decoding is a special case of (1), but the decoder may in general be *mismatched* [2], [3].

We construct the code $\mathcal{M}_n$ such that any two distinct codewords $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{M}_n$ satisfy $d(\boldsymbol{x}, \boldsymbol{x}') > \Delta$ for a given distance function $d(\cdot,\cdot)$ and $\Delta \in \mathbb{R}$. This guarantees that the minimum distance of the codebook exceeds Δ. Similar constructions are used to prove the Gilbert-Varshamov bound in Hamming spaces [4], [5]. Our construction depends on an input distribution $P \in \mathcal{P}(\mathcal{X})$, and we let P_n denote an arbitrary type [1, Ch. 2] whose entries are $\frac{1}{n}$-close to P. The set of sequences with type P_n is denoted by $\mathcal{T}(P_n)$.

Fixing n, M_n, an input distribution $P \in \mathcal{P}(\mathcal{X})$, a distance function $d(\cdot,\cdot)$, and constants $\delta > 0, \Delta \in \mathbb{R}$, the construction is described by the following steps:

1) The first codeword, $\boldsymbol{x}_1$, is drawn uniformly over $\mathcal{T}_1(P_n)$, given by $\mathcal{T}_1(P_n) = \mathcal{T}(P_n)$;
2) The second codeword $\boldsymbol{x}_2$ is uniformly drawn from
$$\mathcal{T}_2(P_n, \boldsymbol{x}_1) = \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) : d(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) > \Delta\} \tag{2}$$
the set of sequences of composition P_n whose distance to $\boldsymbol{x}_1$ exceeds Δ;
3) The i-th codeword $\boldsymbol{x}_i$ is drawn uniformly from
$$\begin{aligned}&\mathcal{T}_i(P_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1})\\ &= \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) : d(\bar{\boldsymbol{x}}, \boldsymbol{x}_j) > \Delta,\, j = 1 \ldots, i-1\}\end{aligned} \tag{3}$$

In order to ensure that the above procedure generates the desired number of codewords $M_n = \mathrm{e}^{nR_n}$ (i.e., the sets $\mathcal{T}_i$ are non-empty for $i = 1, \ldots, M_n$), set Δ and δ such that

$$\mathrm{e}^{n(R_n+\delta)}\mathrm{vol}_{\boldsymbol{x}}(\Delta) \le |\mathcal{T}(P_n)| \tag{4}$$

where $\mathrm{vol}_{\boldsymbol{x}}(\Delta) = |\{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) : d(\bar{\boldsymbol{x}}, \boldsymbol{x}) \le \Delta\}|$ is the volume of a ball of radius Δ according to distance $d(\cdot,\cdot)$ centered at $\boldsymbol{x} \in \mathcal{T}(P_n)$. If the distance d is symmetric and type-dependent, $\mathrm{vol}_{\boldsymbol{x}}(\Delta)$ does not depend on $\boldsymbol{x} \in \mathcal{T}(P_n)$.

Our main result is as follows, namely, a single-letter lower bound for the error exponent of the RGV construction.

Theorem 1. *For any $P \in \mathcal{P}(\mathcal{X})$, $\delta > 0$, $\Delta \in \mathbb{R}$, type-dependent distance function d, and $R > 0$ satisfying*

$$R \le \min_{P_{X\widetilde{X}} : d(P_{X\widetilde{X}}) \le \Delta,\, P_X = P_{\widetilde{X}} = P} I(X; \widetilde{X}) - 3\delta, \tag{5}$$

the RGV construction with parameters $(n, R, P, d, \Delta, \delta)$ and decoding metric $q(\cdot)$ over the DMC W achieves the following error exponent

$$\begin{aligned}&E_{\mathrm{RGV}}(R, P, W, q, d, \Delta) =\\ &\qquad \min_{V \in \mathcal{T}_{d,q,P}(\Delta)} D(V_{Y|X} \| W|P) + \big|I(\widetilde{X}; Y, X) - R\big|_+,\end{aligned} \tag{6}$$

and

$$\begin{aligned}\mathcal{T}_{d,q,P}(\Delta) \triangleq \Big\{&V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) : V_X = V_{\widetilde{X}} = P,\\ &q(V_{\widetilde{X}Y}) \ge q(V_{XY}),\, d(P_{X\widetilde{X}}) \ge \Delta\Big\}.\end{aligned} \tag{7}$$

The following corollary shows that when the distance function $d(\cdot,\cdot)$ is optimized, and Δ is chosen appropriately, the exponent in Theorem 1 recovers the exponent of [6], denoted by $E_q(R, P, W)$, known to be at least as large as the maximum of the random-coding and expurgated exponents.

Corollary 1. *Setting $d(P_{X\widetilde{X}}) = -I(X; \widetilde{X})$, $\Delta = -(R + 3\delta)$ gives that for sufficiently small $\delta > 0$ and $\epsilon > 0$*

$$E_{\mathrm{RGV}}(R, P, W, q, d, \Delta) \ge E_q(R, P, W) - \epsilon. \tag{8}$$

References

[1] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[2] 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.

[3] 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.

[4] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Labs Tech. J.*, vol. 31, no. 3, pp. 504–522, 1952.

[5] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," in *Dokl. Akad. Nauk SSSR*, vol. 117, no. 5, 1957, pp. 739–741.

[6] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, 1981.

This work was supported in part by the European Research Council under Grant 725411, and the Spanish Ministry of Economy and Competitiveness under Grant TEC2016-78434-C3-1-R.

Bounds on Binary Locally Repairable Codes Tolerating Multiple Erasures

Matthias Grezet*, Ragnar Freij-Hollanti†, Thomas Westerbäck*, Oktay Olmez‡ and Camilla Hollanti*
*Department of Mathematics and Systems Analysis, Aalto University, Finland, Email: firstname.lastname@aalto.fi
†Department of Electrical and Computer Engineering, Technical University of Munich, Germany, Email: ragnar.freij@tum.de
‡Department of Mathematics, Ankara University, Turkey, Email: oolmez@ankara.edu.tr

***Abstract*—Recently, locally repairable codes have gained significant interest for their potential applications in distributed storage systems. However, most constructions in existence are over fields with size that grows with the number of servers, which makes the systems computationally expensive and difficult to maintain. Here, we study linear locally repairable codes over the binary field, tolerating multiple local erasures. We derive bounds on the minimum distance on such codes, and give examples of LRCs achieving these bounds. Our main technical tools come from matroid theory, and as a byproduct of our proofs, we show that the lattice of cyclic flats of a simple binary matroid is atomic.**

I. Introduction

In modern distributed storage systems (DSSs) failures happen frequently, whence decreasing the number of connections required for node repair is crucial. Removing even one connection locally can easily imply huge gains in the overall system functionality, thanks to shortened queues and improved data availability. Consequently, locally repairable codes (LRCs) have gained a lot of interest in the past few years [1]–[3]. Namely, LRCs allow to repair a small number of failures locally, *i.e.*, by only contacting few close-by nodes and hence avoiding congesting the system. Related Singleton-type bounds have been derived for various cases, see [4], [5]. The first bound on the minimum distance for fixed field size was obtained by Cadambe and Mazumdar in [6]. Recently, this bound was improved and generalized, via the observation that any log-convex bound on the "local rank" of a code can be blown up to obtain a bound on the global rank [7]. Interestingly, the bounds in [7] do not depend on the linearity of the code. However, all the bounds in [6], [7] are implicit, except for special classes of codes. For more details and tentative comparison, see the last section of this paper.

In this paper, we consider binary codes motivated by the fact that the computational complexity when retrieving a file or repairing a node grows with the field size. We derive new, improved Singleton-type bounds for this special case alongside with sporadic examples, in particular when the local repair sets can tolerate multiple failures. In contrast to the bounds in [6], [7], our bounds are explicit, and do not depend on any prior bounds on binary codes of shorter length.

As our main contribution, in Theorem 5, we obtain a closed-form bound on the minimum distance d of a binary (n, k)-code of length n and dimension k and with all-symbol (r, δ)-locality, where the local distance $\delta > 2$. Such bounds were previously only known when $\delta = 2$. The bound is in terms of the rank ℓ of the repair sets, but can easily be transformed to bounds in terms of the size $r + \delta - 1$. Interestingly, while the two parameters r and ℓ can be assumed to agree when $\delta = 2$, as well as when the field size is unbounded, this is no longer the case over the binary field with $\delta > 2$. While both parameters are of independent interest in applications, we have chosen to focus on the number of nodes ℓ that need to be contacted for local repair, rather than on the size of the local clouds

In addition, in Section III we prove that every element of a non-degenerate binary locally repairable code without replication is contained in an atomic cyclic flat, and hence that the lattice of cyclic flats is atomic. From a practical point of view, this implies a hierarchy of failure tolerance, as explained in the end of Section III. In particular, whenever a symbol e fails, we can start by downloading nodes in an atomic cyclic flat in order to repair e. If it turns out that some other nodes in this local set have failed as well, we can repair them while still keeping the part that we already downloaded, and simply contact some more nodes in the corresponding repair set to repair all the failed symbols. Thus, we do not have to restart from the beginning if we find out during the repair process that a small amount of other nodes have failed as well.

Several constructions are known for optimal LRCs over the binary field, for specified ranges of parameters, and almost exclusively in the case $\delta = 2$. The first such construction, for codes with exponentially low rate and locality $r = 2, 3$, was obtained by deleting carefully chosen columns from the simplex code [8]. These constructions are also optimal when taking the availability t, *i.e.*, the number of disjoint sets that can recover a given symbol, into consideration. A slightly more flexible family of codes, allowing for higher rate, was given in [9], [10], where also a slight improvement over the Cadambe–Mazumdar bound was given for linear codes. In the realm of multiple erasures, *i.e.*, when $\delta > 2$, rate-optimal codes were studied in [11]. There, rate-optimal codes for short length codes were characterized when $\delta = 3$, and analogous constructions without optimality proof were given for $\delta > 3$. However, to the best of our knowledge, no previous work has studied bounds on the global minimum distance in the regime $\delta > 2$.

In the interest of space, we have relegated proofs to an extended version of this paper available on *arXiv* [12].

II. Preliminaries of LRCs and Matroids

As is common practice, we say that C is an (n, k, d)-code if it has length n, dimension k, and minimum Hamming distance d. A linear (n, k, d)-code C over a finite field $\mathbb{F}$ is a *non-degenerate storage code* if $d \geq 2$ and there is no zero column in a generator matrix of C. For a fixed code C, we denote by d_Y the minimum Hamming distance of the punctured code $C|Y$, where $Y \subseteq [n]$ is a set of coordinates of the code C.

Definition 1. *Let C be an (n, k, d)-code over $\mathbb{F}^n$. A symbol $x \in [n]$ has* locality (r, δ) *if there exists a subset R of $[n]$, called* repair set *of x, such that $x \in R$, $|R| \leq r + \delta - 1$, and $d_R \geq \delta$.*

Definition 2. *A* linear (n, k, d, r, δ)-LRC *over a finite field $\mathbb{F}^n$ is a non-degenerate linear (n, k, d)-code C over $\mathbb{F}^n$ such that every coordinate $x \in [n]$ has locality (r, δ). In the literature, this is specifically called* all-symbol locality.

The parameters (n, k, d, r, δ) can immediately be defined and studied for matroids in general, as in [2], [5], [13].

a) Matroid fundamentals: Matroids have many equivalent definitions in the literature. Here, we choose to present matroids via their rank functions.

Definition 3. *A* (finite) matroid *$M = (E, \rho)$ is a finite set E together with a* rank function *$\rho : 2^E \to \mathbb{Z}$ such that for all subsets $X, Y \subseteq E$:*

$$\begin{array}{ll} (R.1) & 0 \leq \rho(X) \leq |X|, \\ (R.2) & X \subseteq Y \quad \Rightarrow \quad \rho(X) \leq \rho(Y), \\ (R.3) & \rho(X) + \rho(Y) \geq \rho(X \cup Y) + \rho(X \cap Y). \end{array}$$

A subset $X \subseteq E$ is called *independent* if $\rho(X) = |X|$. If X is independent and $\rho(X) = \rho(E)$, then X is called a *basis*. A subset that is not independent is called *dependent*. A circuit is a minimal dependent subset of E, that is, a dependent set whose proper subsets are all independent. Strongly related to the rank function is the *nullity function* $\eta : 2^E \to \mathbb{Z}$, defined by $\eta(X) = |X| - \rho(X)$ for $X \subseteq E$.

There is a straightforward connection between linear codes and matroids. Indeed, any linear code C over a field $\mathbb{F}$ generates a matroid $M_C = (E, \rho_C)$, where E is the set of coordinates of C, and $\rho_C(X)$ is the dimension of the punctured code $C|X$. For a given set $X \subseteq E$, we define the *restriction* of M to X to be the matroid $M|X = (X, \rho_{|X})$ by $\rho_{|X}(Y) = \rho(Y)$ for all subsets $Y \subseteq X$.

Two matroids $M_1 = (E_1, \rho_1)$ and $M_2 = (E_2, \rho_2)$ are *isomorphic* if there exists a bijection $\psi : E_1 \to E_2$ such that $\rho_2(\psi(X)) = \rho_1(X)$ for all subsets $X \subseteq E_1$.

Definition 4. *A matroid that is isomorphic to M_C for some code C over $\mathbb{F}$ is said to be* representable *over $\mathbb{F}$. We also say that such a matroid is $\mathbb{F}$-representable. A* binary *matroid is a matroid that is $\mathbb{F}_2$-representable.*

Definition 5. *A matroid is called* simple *if if has no circuits consisting of 1 or 2 elements. A element $e \in E$ is called a* co-loop *if $\rho(E - e) < \rho(E)$.*

b) Fundamentals on cyclic flats: The main tool from matroid theory in this paper are the cyclic flats. We will define them using the closure and cyclic operators.

Let $M = (E, \rho)$ be a matroid. The *closure* operator $\mathrm{cl} : 2^E \to 2^E$ and *cyclic* operator $\mathrm{cyc} : 2^E \to 2^E$ are defined by

$$\begin{array}{ll} (i) & \mathrm{cl}(X) = X \cup \{e \in E - X : \rho(X \cup e) = \rho(X)\}, \\ (ii) & \mathrm{cyc}(X) = \{e \in X : \rho(X - e) = \rho(X)\}. \end{array}$$

A subset $X \subseteq E$ is a *flat* if $\mathrm{cl}(X) = X$ and a *cyclic set* if $\mathrm{cyc}(X) = X$. Therefore, X is a *cyclic flat* if

$$\rho(X \cup y) > \rho(X) \quad \text{and} \quad \rho(X - x) = \rho(X)$$

for all $y \in E - X$ and $x \in X$. The collection of flats, cyclic sets, and cyclic flats of M are denoted by $\mathcal{F}(M)$, $\mathcal{U}(M)$, and $\mathcal{Z}(M)$, respectively. Some more fundamental properties of flats, cyclic sets, and cyclic flats are given in [14].

If $C \subseteq \mathbb{F}^E$ is a linear code, then the cyclic flats of M_C can be described as sets $X \subseteq E$ such that

$$|C|(X \cup y)| > |C|X| \quad \text{and} \quad |C|(X - x)| = |C|X|$$

for all $y \in E - X$ and $x \in X$.

Before going deeper in the study of $\mathcal{Z}(M)$, we need a minimum background on poset and lattice theory. We will use the standard notation of $\wedge$ and $\vee$ for the meet and join operator, we will denote by $0_{\mathcal{L}}$ and $1_{\mathcal{L}}$ the bottom and top element of a lattice $\mathcal{L}$, and we will denote by $\lessdot$ the covering relation, *i.e.*, for $X, Y \in (\mathcal{L}, \leq)$, we say that $X \lessdot Y$ if $X < Y$ and there is no $Z \in \mathcal{L}$ with $X < Z < Y$.

The atoms and coatoms of a lattice $(\mathcal{L}, \subseteq)$ are defined as

$$A_{\mathcal{L}} = \{X \in \mathcal{L} : 0_{\mathcal{L}} \lessdot X\} \text{ and } coA_{\mathcal{L}} = \{X \in \mathcal{L} : X \lessdot 1_{\mathcal{L}}\},$$

respectively. A lattice $\mathcal{L}$ is said to be *atomic* if every element of $\mathcal{L}$ is the join of atoms.

We can now give a crucial property of the set of cyclic flats.

Proposition 1 (See [14]). *Let $M = (E, \rho)$ be a matroid. Then*

1) *$(\mathcal{Z}(M), \subseteq)$ is a lattice with $X \vee Y = \mathrm{cl}(X \cup Y)$ and $X \wedge Y = \mathrm{cyc}(X \cap Y)$.*
2) *$1_{\mathcal{Z}} = \mathrm{cyc}(E)$ and $0_{\mathcal{Z}} = \mathrm{cl}(\emptyset)$.*

c) Relation between LRCs and the lattice of cyclic flats: Recently, some work has been done to emphasise the relation between cyclic flats and linear codes over finite fields. In [5], the authors proved that the minimum distance can be expressed in terms of the nullity of certain cyclic flats:

Proposition 2. *Let C be a non-degenerate (n, k, d)-code and $M = (E, \rho)$ the matroid associated to C. Then,*

$$d = \eta(E) + 1 - \max\{\eta(Z) : Z \in \mathcal{Z}(M) - \{E\}\}.$$

Moreover, [15] gives us necessary conditions on the structure of the lattice of cyclic flats when the code and hence the matroid are binary. The key results from [15] are the following proposition and theorem that constrain the edges of the associated Hasse diagram.

Proposition 3 ([15]). *Let $M = (E, \rho)$ be a binary matroid. Then, every $X, Y \in \mathcal{Z}(M)$ with $X \lessdot Y$ satisfy exactly one of the following:*

- *$\rho(Y) - \rho(X) = l > 1$ and $\eta(Y) - \eta(X) = 1$. We call such an edge in the Hasse diagram of $\mathcal{Z}(M)$ a rank edge and label it $\rho = l$.*
- *$\rho(Y) - \rho(X) = 1$ and $\eta(Y) - \eta(X) = l > 1$. We call such an edge a nullity edge and label it $\eta = l$.*
- *$\rho(Y) - \rho(X) = 1$ and $\eta(Y) - \eta(X) = 1$. We call such an edge a elementary edge.*

Theorem 1 (Announced in [15]). *Let C be a non-degenerate, binary linear (n, k, d, r, δ)-LRC with $d > 2$ and without replication. Let $M = (E, \rho)$ be the associated matroid. Then $\mathcal{Z} = \mathcal{Z}(M)$ satisfies the following:*

1) *$\emptyset$ and E are cyclic flats.*
2) *Every covering relation $Z \lessdot E$ is a nullity edge labeled with a number $\geq d - 1$.*
3) *If $\delta = 2$, then for every $i \in E$, there is $X \in \mathcal{Z}$ with $i \in X$ such that $\rho(X) \leq r$.*
4) *If $\delta > 2$, then for every $i \in E$, there is $X \in \mathcal{Z}$ with $i \in X$ such that*
    a) *Every covering relation $Y \lessdot X$ is a nullity edge labeled with a number $\geq \delta - 1$.*
    b) *Every cyclic flat Y with $Y \lessdot X$ has rank $\leq r - 2$.*

III. Lattice structure and repair properties

The first part of this section is devoted to understanding how restricting to binary linear codes affects the structure of the lattice of cyclic flats. The main result of this section consists of proving that the lattice of cyclic flats has the property of being atomic.

In the second part, we will discuss the meaning of these results for binary linear codes and LRCs. In particular, we will see that every non-degenerate binary linear (n, k, d)-code without replication is already a binary linear $(n, k, d, r', 2)$-LRC for a certain r'. Furthermore, for LRCs with $\delta > 2$, we will see that these codes have a hierarchy in failure tolerance.

a) Structural properties of the lattice of cyclic flats: We will first begin by the relation between binary linear codes and the associated matroid. The following proposition is a reformulation of Proposition 8 in [5] together with the easy observation that, in a binary linear code, two symbols are dependent if and only if they are equal.

Proposition 4. *Let C be a binary linear (n, k, d)-code. Then C is non-degenerate with no replication if and only if the associated matroid $M = (E, \rho)$ is simple and contains no co-loops.*

Now that we have established the type of matroids that is revelant to our case, we can study the implications of Proposition 3 for the lattice of cyclic flats. The following lemma, in addition to being a crucial step towards proving Theorem 2, has even stronger implications, as it shows that any element in $\mathcal{Z}(M)$ is equal not only to the join, but also to the union of all the atoms it contains.

Lemma 1. *Let $M = (E, \rho)$ be a simple binary matroid that contains no co-loops, e an element of E, and C a circuit of M. We have the following results:*

1) *Z is an atom of $\mathcal{Z}(M)$ if and only if $\eta(Z) = 1$.*
2) *$\mathrm{cl}(C)$ is an atom of $\mathcal{Z}(M)$ if and only if $\mathrm{cl}(C) = C$.*
3) *If C is a circuit containing e of minimal length, then $\mathrm{cl}(C) = C$.*
4) *Every element $e \in E$ is contained in an atom.*

Theorem 2. *Let $M = (E, \rho)$ be a simple binary matroid that contains no co-loops. Then the lattice of cyclic flats $\mathcal{Z}(M)$ is atomic.*

b) Hierarchy of failure tolerance: By Proposition 4, we can reinterpret Lemma 1 and Theorem 2 as statements about non-degenerate binary storage codes. Indeed, combining Lemma 1.4 with Proposition 2, we see that every symbol is directly contained in a small repair set with $\delta = 2$, *i.e.*, in a repair set that can correct exactly one erasure. Hence we obtain the following theorem.

Theorem 3. *For every non-degenerate binary linear (n, k, d)-code C with no replication, C is also an $(n, k, d, r', 2)$-LRC for some $r' \in \{2, \ldots, k\}$.*

Now, if we want to be able to correct more than one erasure, then the repair sets cannot be atoms of $\mathcal{Z}(M)$ as these have $d_Z = 2$. They have to be at least one level above some atoms. However, the previous theorem still holds, meaning that for every symbol, there is also an atom containing it. Thus, we get a natural hierarchy in failure tolerance. If one node fails, then we can contact the close-by nodes in the atom to repair it. If more nodes fail, but no more than $\delta - 1$, we can contact other repair sets to fix them. And if more than $\delta - 1$ nodes fail, then we need to use the global properties of the code.

Moreover, by the remark following Theorem 2, it follows that repair sets are unions of all the atoms below them. Since the collection of repair sets contains every symbol, we can choose the collection of atoms that will give us the $(r', 2)$ locality to be inside repair sets. The following corollary summaries the previous observations in one statement.

Corollary 1. *Let C be a non-degenerate binary linear (n, k, d, r, δ)-LRC with no replication and with $\delta > 2$. Let $\{R_i\}_{i \in I}$ be the list of repair sets. Then, there exists a collection of sets $\{X_j\}_{j \in J}$ such that for all X_j, there exists R_i with $X_j \subsetneq R_i$ such that C is also an $(n, k, d, r', 2)$-LRC.*

From a practical point of view, this reinforces the usefulness of the failure tolerance hierarchy. For example, suppose that the symbol $e \in E$ fails. We can start by downloading nodes in the atom Z_{at}^{e} in order to repair e. Now, if we realize that some other nodes in Z_{at}^{e} have failed as well, we can keep the part that we already downloaded from Z_{at}^{e} and contact more nodes in the corresponding repair set to repair all the failed symbols in Z_{at}^{e}. Thus, we do not have to restart from the beginning if we find out during the repair process that a small amount of other nodes have failed as well.

IV. Improving the Singleton-type bound for $\delta > 2$

The goal of this section is to improve the existing bound for non-degenerate linear (n, k, d, r, δ)-codes C when the codes are binary, contain no replication and $\delta > 2$. It has been proven in [16] that, for a linear (n, k, d, r, δ)-code over $\mathbb{F}_q$, we have

$$d \leq n - k + 1 - \left(\left\lceil \frac{k}{r} \right\rceil - 1 \right) (\delta - 1). \tag{1}$$

We start by defining a new parameter that represents the maximum rank of a repair set.

Definition 6. *Let C be an (n, k, d, r, δ)-LRC. Let $\{R_i\}_{i \in I}$ be the list of repair sets and $M = (E, \rho)$ the associated matroid to C. We define ℓ to be $\ell := \max_{i \in I} \rho(R_i)$.*

As mentioned in the introduction, even if the maximum rank and r can be assumed to coincide over large field size, this is no longer the case for binary codes with $\delta > 2$. Indeed, as binary MDS codes do not exist when the minimum distance is greater than 2, we must have $\ell < r$ for such codes. Moreover, any bound on the rank of a binary code in terms of its size and minimum distance, gives a bound on ℓ in terms of r, and as a consequence a new translation of our bounds to a bound in terms of r. As our main interest is in small values of δ, Proposition 5 is strong enough for the purposes of this paper.

Proposition 5. *Let C be a non-degenerate, binary linear(n, k, d, r, δ)-LRC with $\delta > 2$ and without replication. Let $\{R_i\}_{i \in I}$ be the collection of repair sets. Then,*

$$\ell \leq r - 1 \qquad \textit{and} \qquad \eta(R_i) \geq \delta \quad \textit{for all } i \in I.$$

Since we will focus on the rank of the repair sets rather than on the size, we assume from now on that repair sets are maximal after fixing its rank or, in matroid terms, that repair sets are cyclic flats. We will denote these repair sets by Z_i instead of R_i to avoid confusion.

Remember that Proposition 2 links the minimum distance of a code to the coatoms among the cyclic flats. We would like to construct a cyclic flat that is close to the coatoms level, to give an accurate lower bound on $\max\{\eta(Z) : Z \in \mathcal{Z}(M) - \{E\}\}$. We will do this by creating a chain in $\mathcal{Z}(M)$ of joins of repair sets and we will call these type of chains *repair-sets*-chain or, for short, *rps*-chain.

Definition 7. *Let C be a non-degenerate (n, k, d, r, δ)-code and $\{Z_i\}_{i \in I}$ the collection of repair sets. Let $M = (E, \rho)$ the associated matroid. An rps*-chain

$$\emptyset = Y_0 \subsetneq Y_1 \subsetneq \ldots \subsetneq Y_m = E$$

is a chain in $\mathcal{Z}(M)$ defined inductively by

1) *Let $Y_0 = \emptyset$.*
2) *Given $Y_{i-1} \subsetneq E$, we choose $x_i \in E \setminus Y_{i-1}$ and Z_i with $x_i \in Z_i$ arbitrarily, and assign $Y_i = Y_{i-1} \vee Z_i$.*
3) *If $Y_i = E$, we set $m = i$.*

Notice that this chain is not uniquely defined since we can choose symbols and corresponding repair sets freely.

Now, we are interested in how the rank and the nullity can increase at each step. To bound the rank and nullity difference at each step of the *rps*-chain $(Y_i)_{i=2}^m$, one can use the rank axioms to obtain $\rho(Y_i) - \rho(Y_{i-1}) \leq \ell$ and $\eta(Y_i) - \eta(Y_{i-1}) \geq \delta - 1$. However, this does not take binarity, and in particular Proposition 3, into account. Indeed, if $\rho(Y_i) - \rho(Y_{i-1}) = \ell$, then we must have $Y_{i-1} \cap Z_i = \emptyset$ and $\rho(Z_i) = \ell$. Hence, there is no code nor an *rps*-chain that can simultaneously achieve both bounds.

Next, we introduce an indicator function that will capture when the intersection is a coatom of a repair set. To be more concise, we will denote by $\mathcal{A}_i$ the event that $Y_{i-1} \cap Z_i \in coA_{\mathcal{Z}(M|Z_i)}$. First, this is a necessary condition to have $\eta(Y_i) - \eta(Y_{i-1}) = \delta - 1$. Secondly, this will also imply that $\rho(Y_i) - \rho(Y_{i-1}) = 1$ since every covering relation of a repair set has a nullity edge by Proposition 3. The following lemma summarizes these observations.

Lemma 2. *Let C be a non-degenerate, binary linear (n, k, d, r, δ)-LRC with $\delta > 2$ and without replication, and let $M = (E, \rho)$ be the associated matroid. Let $\{Z_i\}_{i \in I}$ be the collection of repair sets and $(Y_i)_{i=0}^m$ an associated rps-chain. Then $(Y_i)_{i=0}^m$ has the following properties:*

1) *$\rho(Y_i) - \rho(Y_{i-1}) \leq \ell - (\ell - 1)\mathbb{1}_{\mathcal{A}_i}$ for all $i = 2, \ldots, m$.*
2) *$\eta(Y_i) - \eta(Y_{i-1}) \geq \delta - \mathbb{1}_{\mathcal{A}_i}$ for all $i = 2, \ldots, m$.*

In order to use Lemma 2 to get a Singleton-type bound with these assumptions, we need to define a new parameter that will count the number of times the intersection $Y_{i-1} \cap Z_i$ is a coatom of Z_i.

Definition 8. *Let $(Y_i)_{i=0}^m$ be an rps-chain. Define $0 \leq \alpha \leq 1$ by $\alpha m = \#\{i : Y_{i-1} \cap Z_i \in coA_{\mathcal{Z}(M|Z_i)}\}$. We say that $(Y_i)_{i=0}^m$ is an rps$_\alpha$*-chain.

We can now derive a new Singleton-type bound with the extra parameter α.

Theorem 4. *Let C be a non-degenerate, binary linear (n, k, d, r, δ)-LRC with $\delta > 2$ and without replication. Let $\alpha \in [0, 1]$ be such that C has an rps$_\alpha$-chain. Then,*

$$d \leq n - k + 1 + \delta - \left\lceil \left\lceil \frac{k}{\ell - (\ell - 1)\alpha} \right\rceil (\delta - \alpha) \right\rceil .$$

Since this bound is valid for all α, we can optimize α to get a bound for all types of *rps*-chain, *i.e.*, a Singleton-type bound that only depends on the parameters n, k, d, ℓ and δ.

Theorem 5. *Let C be a non-degenerate, binary linear (n, k, d, r, δ)-LRC with $\delta > 2$ and without replication. Then,*

$$d \leq n - k + 1 - \left(\left\lceil \frac{k}{\ell} \right\rceil - 1 \right) \delta + \mathbb{1}_{\ell | (k-1) \text{ and } l \neq k-1}. \tag{2}$$

In order to make the comparison to the previously known bound (1) easier and to emphasize the improvement provided by the new bound, we state the following corollary of Theorem 5 using Proposition 5. This gives us a bound that only depends on n, k, d, r and δ.

Corollary 2. *Let C be a non-degenerate, binary linear (n, k, d, r, δ)-LRC with $\delta > 2$ and without replication. Then,*

$$d \leq n - k + 1 - \left(\left\lceil \frac{k}{r - 1} \right\rceil - 1 \right) \delta + \mathbb{1}_{(r-1)|(k-1) \text{ and } r \neq k}. \tag{3}$$

We provide one small example that achieves the bound from Corollary 2.

Example 1. *Let C be the binary linear $(10,4,4)$-code given by the following generator matrix,*

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

We can define three repair sets by their corresponding columns in G, $Z_1 = \{1,2,3,5,6,8\}$, $Z_2 = \{2,3,6,7,9,10\}$, and $Z_3 = \{1,4,6,7,8,10\}$.

For all $i \in \{1,2,3\}$, we have $|Z_i| = 6, \rho(Z_i) = 3$, and $d_{Z_i} = 3$, and hence we obtain a binary linear $(10,4,4,4,3)$-LRC that achieves the bound from Corollary 2.

The following graph is a comparison of the previous known Singleton-type bound (1) and the new one from Corollary 2 for two different values of r. We can see that the new bound is always better than (or equivalently, smaller) or equal to the previous bound.

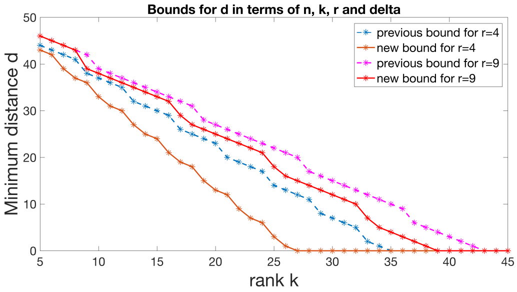


Fig. 1: Comparison of the previous Singleton-type bound (1) and the new bound (3) for $n = 50$ and $\delta = 3$.

The most commonly used field-dependent distance bound for LRCs is the Cadambe–Mazumdar bound [6], which is only stated without reference to the local minimum distance (or equivalently, for $\delta = 2$). It can straightforwardly be generalized to codes tolerating more than one local erasure, as also noted in [17]:

Proposition 6. *(Cf. [17, Remark 3]) Let C be a non-degenerate linear (n,k,d,r,δ)-LRC over $\mathbb{F}_q$ with maximal local rank ℓ. Then,*

$$k \leq \min_{t \in \mathbb{Z}_+} \left[t\ell + k_{opt}^{(q)}(n - t(r+\delta-1), d) \right], \qquad (4)$$

where $k_{opt}^{(q)}(n,d)$ is the maximum rank of a linear code of length n and minimum distance d over $\mathbb{F}_q$.

However, the determination of $k_{opt}^{(q)}$ is a classical open problem in coding theory. Moreover, even given a formula for $k_{opt}^{(q)}$, evaluating (4) may be a tedious task. In that sense, the bound (2) is more explicit than this one.

Comparing the bounds (2) and (4) represents a challenge since (2) is a bound on d with, on the right-hand side, a ceiling function on k, while (4) is a bound on k with a minimum over another unknown bound that includes d. One method is to transform (2) to have every term in k on the left, so it will be bounded by the remaining terms on the right hand side and also by the left hand side after replacing k by the maximum dimension given by (4). The best bound will be the one that gives the minimum of the two alternative right hand sides. This allows for a partial but computable comparison of (2) and (4).

We estimated $k_{opt}^{(q)}$ in (4) via the Plotkin bound and took the approach described above and it turned out, for the values we tried, the extension of the Cadambe–Mazumdar bound (4) is better or equal than (2), *i.e.*, gives the afore-mentioned minimum. However this is only a glimpse of the relation between the two bounds and the complete comparison is left for future work. The bound (2) improves the known Singleton-type bound for binary LRCs and highlights explicitly constraints on the minimum distance d. In conclusion, this work takes the first step toward an explicit bound for binary LRCs via matroidal techniques, and further improvements can be obtained by extending the techniques developed in this paper, via a more detailed (and technical) study of the cyclic flats. This is left to an extended version of this article.

Acknowledgment. This work is supported in part by the Academy of Finland (grants #276031, #282938 and #303819 to C. Hollanti), and by the TU Munich – Institute for Advanced Study, funded by the German Excellence Initiative and the EU 7th Framework Programme (grant #291763).

References

[1] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, “On the locality of codeword symbols,” *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6925–6934, 2012.

[2] I. Tamo, D. Papailiopoulos, and A. Dimakis, “Optimal locally repairable codes and connections to matroid theory,” *IEEE Trans. Inf. Theory*, vol. 62, pp. 6661–6671, 2016.

[3] I. Tamo and A. Barg, “A family of optimal locally recoverable codes,” *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4661–4676, 2014.

[4] I. Tamo, A. Barg, and A. Frolov, “Bounds on the parameters of locally recoverable codes,” *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3070–3083, 2016.

[5] T. Westerbäck, R. Freij-Hollanti, T. Ernvall, and C. Hollanti, “On the combinatorics of locally repairable codes via matroid theory,” *IEEE Trans. Inf. Theory*, vol. 62, pp. 5296–5315, 2016.

[6] V. Cadambe and A. Mazumdar, “An upper bound on the size of locally recoverable codes,” in *IEEE NetCod*, 2013, pp. 1–5.

[7] A. Agarwal, A. Barg, S. Hu, A. Mazumdar, and I. Tamo, “Combinatorial alphabet-dependent bounds for locally repairable codes,” 2017, arXiv: 1702.02685.

[8] N. Silberstein and A. Zeh, “Optimal binary locally repairable codes via anticodes,” in *IEEE ISIT*, 2015, pp. 1247–1251.

[9] P. Huang, E. Yaakobi, H. Uchikawa, and P. Siegel, “Cyclic linear binary locally repairable codes,” in *IEEE ITW*, 2015.

[10] ——, “Binary linear locally repairable codes,” *IEEE Trans. Inf. Theory*, vol. 62, pp. 5296–5315, 2016.

[11] S. Balaji, K. Prashant, and P. V. Kumar, “Binary codes with locality for multiple erasures having short block length,” in *IEEE ISIT*, 2016, pp. 655–659.

[12] M. Grezet, R. Freij-Hollanti, T. Westerbäck, O. Olmez, and C. Hollanti, “Bounds on binary locally repairable codes tolerating multiple erasures,” 2017, arXiv: 1709.05801.

[13] R. Freij-Hollanti, T. Westerbäck, and C. Hollanti, “Locally repairable codes with availability and hierarchy: matroid theory via examples,” in *International Zürich Seminar on Communications*. IEEE/ETH, 2016, pp. 45–49, invited paper.

[14] J. E. Bonin and A. de Mier, “The lattice of cyclic flats of a matroid,” *Ann. Comb.*, vol. 12, pp. 155–170, 2008.

[15] M. Grezet, R. Freij-Hollanti, T. Westerbäck, and C. Hollanti, “On binary matroid minors and applications to data storage over small fields,” in *ICMCTA*, 2017, pp. 139–153.

[16] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar, “Codes with local regeneration and erasure correction,” *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4637–4660, 2014.

[17] A. S. Rawat, A. Mazumdar, and S. Vishwanath, “Cooperative local repair in distributed storage,” *EURASIP J. Adv. Signal Proc.*, pp. 1–17, 2015.

Absorbing Sets of Generalized LDPC Codes

Marco Ferrari*, Alessandro Tomasoni*, Luca Barletta† and Sandro Bellini†
*CNR-IEIIT, Politecnico di Milano, Milan, Italy, Email: marco.ferrari@ieiit.cnr.it, alessandro.tomasoni@ieiit.cnr.it
†Politecnico di Milano, Milan, Italy, Email: luca.barletta@polimi.it, sandro.bellini@polimi.it

***Abstract*—In this paper we propose a definition of Absorbing Sets for binary Generalized LDPC (GLDPC) codes. We show that under practical Max-Log iterative decoding, our AS definition enables a local description of the message evolution with the iterations, with a simplified model very similar to the one used for the analysis of Min-Sum LDPC decoding. Accordingly, these ASs exhibit a threshold behavior also in GLDPC codes.**

***Index Terms*—Generalized Low-Density Parity-Check codes, Error floor, Absorbing sets, Max-Log decoding, Tanner graph.**

I. Introduction

After the introduction of Turbo-Codes and the rediscovery of Low-Density Parity-Check (LDPC) codes, the idea of Generalized LDPC (GLDPC) codes was also recovered from [1]. GLDPC codes raised new attention as a compromise between the two aforementioned classes of iterative concatenated codes, especially because they appeared not to suffer from the *error floor* phenomenon. In [2] and [3] it is proven that in the ensembles of GLDPC codes with Hamming component codes there exist codes with minimum Hamming distance growing linearly with the block-size, even with low Variable Node (VN) degree $d_v = 2$ (GLDPC codes are usually considered with VN-degree 2 since decoding complexity is minimized and the code rate is maximized). In other terms, GLDPC codes show *good spectral shape behavior* [4].

The good spectral shape behavior, although necessary for floorless codes, is not sufficient under suboptimal decoding, such as message passing on graphs with loops. For GLDPC codes it has been observed that iterative decoders can fail, ending without valid decisions, both over the Binary Erasure Channel (BEC) and over the Binary Symmetric Channel (BSC). In [5] the definition of *Stopping Set* (SS) from [6] is generalized to *Stopping Set* of order m, as a subset $\mathcal{S}$ of VNs whose neighboring Check Nodes (CNs) are connected to $\mathcal{S}$ at least m times. In [5] SSs are identified as the main cause of error floors both on the BEC and on the BSC under iterative Hard Bounded Distance decoding. In [7], the asymptotic exponent of the SS size distribution in GLDPC codes is investigated in conjunction with the Hamming weight distribution of the code.

Absorbing sets (ASs), defined in [8], are combinatorial substructures of the Tanner graph in LDPC codes that describe the dominant decoding failures of various soft message passing decoders over AWGN channels [8], [9]. Recently Non-Binary (NB) LDPC codes have gained new interest and the definition of ASs has been extended to NB-LDPC codes in [10]. These ASs are named Generalized AS (GAS) in [11].

Elementary ASs (i.e. with CNs connected no more than twice to the VNs of the AS) enable a linear state-space model for the local analysis of the iterative decoder (see [12] and references therein). In [13] and [14] we studied through a linear state-space model with saturation, the behavior of practical iterative decoders in binary LDPC Tanner graphs with ASs and we defined an AS parameter, the *threshold*, that discriminates the existence/non-existence of misleading equilibria for the iterative decoder.

In this paper we propose a definition of ASs for GLDPC binary codes that captures decoding failures of practical, Max-Log [15] iterative decoders, over AWGN channels. We focus on degree-2 VNs, for which the GAS definition cannot be trivially extended to GLDPC codes. We show that our definition of ASs for GLDPC codes, under Max-Log decoding, enables a linear model similar to that used in [13], [14] for binary LDPC codes. Therefore also GLDPC decoders exhibit a threshold behavior in presence of ASs. We show a couple of examples of GLDPC codes with ASs of size provably smaller than the minimum Hamming distance of the code, that can indeed entrap the iterative decoder. Thereby these ASs are responsible for an error floor whose probability also depends on the multiplicity of these structures inside the graph. Finally, we discuss the problem of the search of these ASs in GLDPC codes with extended Hamming component codes and we check their multiplicity against a probabilistic computation.

II. Generalized LDPC Codes and Notation

A binary regular GLDPC code, with N_v VNs of degree $d_v = 2$ and N_c CNs, is defined by the biadjacency matrix Γ and by the code constraints imposed by the CNs. The CNs could be a mixture of various component codes. In this paper, to keep notation simple, we assume one type of component code only, $\mathcal{C}(N, K)$. The matrix Γ has $d_v = 2$ ones per column, N ones per row, and size $N_c \times N_v$, where $N_c = 2N_v/N$. Each row of Γ has ones in the columns corresponding to the N VNs that are constrained to form a codeword of $\mathcal{C}$. Replacing each 1 in Γ with a column of the parity check matrix H_c of $\mathcal{C}$ we obtain the parity check matrix H of the GLDPC code, of size $(N-K)N_c \times N_v$. The design code rate is $R = 1 - (N-K)N_c/N_v = 2R_c - 1$ with $R_c = K/N$.

In Fig. 1 we draw the bipartite graph of a GLDPC code with the above constraints. We order the VNs according to the first set of component codewords, and we let a permutation matrix π assign the VNs to the CNs according to the matrix Γ, and their position inside each codeword of $\mathcal{C}$. Iterative decoding is

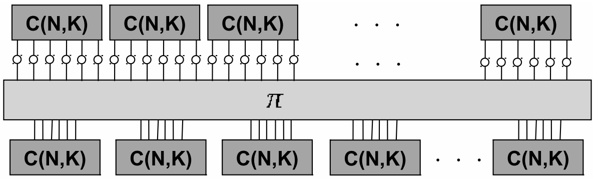


Fig. 1. GLDPC Tanner graph with $d_v = 2$ and component code $\mathcal{C}(N, K)$.

run by activating the CNs on one side of π, then the VNs, then the CNs on the other side of π and finally the VNs again, and then iterating this procedure. An optimal MAP decoder for $\mathcal{C}$, when activated with input LLRs $L_k, (k = 1 \dots N)$, computes extrinsic messages $E_j, (j = 1 \dots N)$ for the VNs, by

$$E_j = \log \frac{\sum_{\mathbf{c} \in \mathcal{C}: c_j = +1} \prod_{k=1}^{N} \exp(c_k L_k/2)}{\sum_{\mathbf{c} \in \mathcal{C}: c_j = -1} \prod_{k=1}^{N} \exp(c_k L_k/2)} - L_j. \quad (1)$$

When activated, each VN $v_i, (i = 1 \dots N_v)$ computes the a posteriori LLR adding the two extrinsic LLRs E'_i and E''_i received by the neighboring CNs, with the channel LLR λ_i

$$O_i = \lambda_i + E'_i + E''_i \quad (2)$$

and the input messages for the two CNs by

$$L'_i = \lambda_i + E''_i, \quad L''_i = \lambda_i + E'_i. \quad (3)$$

III. Absorbing Sets of GLDPC Codes

An absorbing set [8] in LDPC codes is a subset of VNs that, although not forming a codeword, locally satisfy a majority of neighboring CNs of each VN. These subgraphs can lock the iterative decoders to wrong decisions, despite some CNs left unsatisfied, because iterative decoding processes messages only at a local level. Assume to transmit the all-zero codeword, corresponding to symbols $c_i = +1, \forall i$: messages greater than zero correspond to correct decisions, whereas negative messages correspond to errors. Suppose that, at a certain iteration, the decoder has negative decisions for all the VNs of an AS. Satisfied CNs propagate negative messages that reinforce the wrong decisions. Unsatisfied CNs try to correct these values forwarding positive messages, but they are a minority and thus can fail to correct the decisions.

In GLDPC codes, CNs compute messages based on the component code $\mathcal{C}$ as in (1). In practical implementations, MAP decoders (1) are generally replaced by their *Max-Log* versions, and messages are quantized and saturated to a maximal value. Assuming Max-Log decoding and a function *sat* that clips extrinsic messages to their maximum value,[1] (1) is replaced by

$$E_j = sat \left[\max_{\mathbf{c} \in \mathcal{C}: c_j = +1} \sum_{k \neq j}^{N} \frac{c_k L_k}{2} - \max_{\mathbf{c} \in \mathcal{C}: c_j = -1} \sum_{k \neq j}^{N} \frac{c_k L_k}{2} \right] \quad (4)$$

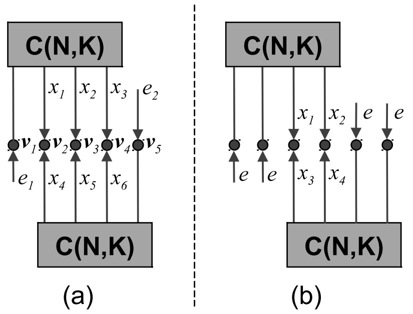


Fig. 2. Examples of Absorbing Sets for GLDPC codes with component codes of minimum Hamming distance $d_H = 4$.

By (4) it is apparent that a CN propagates a negative message E_i whenever the most likely codeword (neglecting L_i) has $c_i = -1$. And this can happen, even with one single negative input message L_k, provided its reliability is higher than the sum of the positive ones. In other words, a simple classification of neighboring CNs *satisfied/unsatisfied* does not capture the harmful subgraphs that entrap the decoder.

In GLDPC codes, a set of VNs with values -1 must match a codeword of $\mathcal{C}$ to locally satisfy the CN. This set must have at least size d_H, where d_H is the minimum Hamming distance of $\mathcal{C}$. Suppose that the two $\mathtt{max}$ operators in (4) select $\mathbf{c} = +\mathbf{1}$ and a codeword of weight d_H, e.g., without loss of generality, with $c_k = -1$ in positions $k = 1, 2 \dots d_H$. The output (4) of the CN does not depend on the rest of the LLRs in the component codeword, and it reads

$$E_j = sat \left[L_1 + L_2 + \dots L_{d_H} - L_j, \right] \quad j = 1 \dots d_H. \quad (5)$$

Thereby, under Max-Log decoding, we can write a quasi-linear relation between input and output messages inside the set of VNs that correspond to low weight codewords of the component codes.

A. Saturated Linear Model for CN and VN Decoders

For instance, assume that $\mathcal{C}$ is an extended Hamming (eH) code with $d_H = 4$, and that the permutation matrix π allows subgraphs like that drawn in Fig. 2 (a). The subset of VNs $\mathcal{D} = \{v_1, v_2, v_3, v_4, v_5\}$, of size $a = 5$, is the union of the VNs that form two minimum Hamming weight codewords of $\mathcal{C}$, constrained by the two CNs. Let $\mathbf{x} = [x_1, \dots x_6]^T$ be the extrinsic messages sent by the two CNs to the VNs in $\mathcal{D}_2 \subset \mathcal{D}$ which are connected to both of them, i.e., $\mathcal{D}_2 = \{v_2, v_3, v_4\}$. Using (5), we can write a quasi-linear system (apart from saturation) that describes the iterative activation of VNs and CNs and involves the messages $\mathbf{x}^{(k)}$ generated by the CNs at the kth iteration, the channel LLRs $\boldsymbol{\lambda}$, and the messages $\mathbf{e} = [e_1, e_2]^T$ received by v_1 and v_5, respectively, from CNs outside the subgraph. In matrix form, the system reads

$$\mathbf{x}^{(k)} = sat \left(\mathbf{A} \mathbf{x}^{(k-1)} + \mathbf{R} \mathbf{e} + \mathbf{C} \boldsymbol{\lambda} \right) \quad (6)$$

where the 6×6 *routing matrix* $\mathbf{A}$ and the 6×2 *external LLRs matrix* $\mathbf{R}$ forward the internal extrinsic messages $\mathbf{x}^{(k-1)}$,

[1] Apart from saturation, in Eq. (4) E_j is linear in the inputs L_k. The saturation level can be set arbitrarily as long as all L_k and E_j are scaled accordingly. In this paper, as in [13], we assume a function $sat(x)$ that clips x to ± 1 and input LLRs L_k scaled by the maximal extrinsic value E_{max}.

and the external extrinsic messages $\mathbf{e}$, respectively. The 6×5 *channel LLRs matrix* $\mathbf{C}$ combines the channel LLRs λ_i of each VN $v_i, (i = 1...5)$ inside each message x_j. I.e.,

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix},$$

$$\mathbf{R}^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Note that, unlike in the state-space linear model of [13], here the CNs generate the linear combinations of internal messages, whereas the degree-2 VNs swap the messages of the two incoming edges. The system (6), that locally describes iterative decoding, allows *misleading equilibria*, i.e. pairs of vectors $(\mathbf{x}, \boldsymbol{\lambda})$ such that $\mathbf{x} = sat\left(\mathbf{Ax} + \mathbf{Re} + \mathbf{C}\boldsymbol{\lambda}\right)$ is stable along the iterations and produces wrong decisions. For instance, the pair $(\mathbf{x}, \boldsymbol{\lambda}) = (-\mathbf{1}, -\mathbf{1})$ is an equilibrium for any vector $\mathbf{e}$ since $\mathbf{e} \leq \mathbf{1}$ because of saturation,[2] and it corresponds to wrong decisions about all VNs in $\mathcal{D}$ since $O_i < 0, i = 1..5$ according to (2). These decisions cannot be changed by further iterations, independently of the messages incoming from the external graph, despite the CNs are not satisfied. In other words, the subgraph of Fig. 2 (a) is an *absorbing set*.

We can look for sufficient conditions for system (6) to converge to an equilibrium corresponding to correct decisions. As in [13], we assume that in the rest of the graph messages converge towards correct decisions and we start the analysis of the iterations when the messages received by $\mathcal{D}$ from external CNs are already saturated to their maximal value, $\mathbf{e} = +\mathbf{1}$. This point of view is chosen to decouple the dynamical behavior of the decoder inside and outside the AS. By this choice we should use an initial vector $\mathbf{x}^{(0)}$ which is the result of the message evolution up to that iteration, which is unknown. Since we are looking for sufficient conditions, we can consider any starting configuration $\mathbf{x}^{(0)}$. If no $\mathbf{x}^{(0)}$ results in a convergence failure, the AS cannot trap the decoder independently of this message evolution.

Note that the ith row weight of $\mathbf{A}$ is equal to the number of internal messages x_j that are added to the external messages e_k to compute the extrinsic LLR x_i as by (5). Since we are assuming that $\mathbf{e} = +\mathbf{1}$, we have $\mathbf{A1} + \mathbf{Re} = (d_H - 1)\mathbf{1}$, and (6) can be rewritten as

$$\mathbf{x}^{(k)} = sat\left(\mathbf{A}\left(\mathbf{x}^{(k-1)} - \mathbf{1}\right) + (d_H - 1)\,\mathbf{1} + \mathbf{C}\boldsymbol{\lambda}\right). \quad (7)$$

If we let $\mathbf{C}\boldsymbol{\lambda} = \boldsymbol{\mu} = [\mu_1, \mu_2...\mu_6]^T$, the system (7) is formally identical to the system assumed in [13] and [14]. The only difference is that each entry μ_i is the sum of the $d_H - 1$ independent channel LLRs of the VNs that complete a codeword of weight d_H with the recipient of x_i.

The formal equivalence of the dynamical system (7) with [14, Eq.(8)], reveals a threshold behavior similar to ASs for binary LDPC codes. Given the equilibrium of the system (7), i.e. pairs $(\mathbf{x}, \boldsymbol{\mu})$ such that $\mathbf{x} = f\left(\mathbf{x}, \boldsymbol{\mu}\right)$ with

$$f\left(\mathbf{x}, \boldsymbol{\mu}\right) = sat\left(\mathbf{A}\left(\mathbf{x} - \mathbf{1}\right) + (d_H - 1)\,\mathbf{1} + \boldsymbol{\mu}\right), \quad (8)$$

we can compute τ_μ defined as

$$\tau_\mu = \max_{(\boldsymbol{\mu}, \mathbf{x})} \quad \min(\boldsymbol{\mu}) \quad (9)$$
$$s.t. \quad -\mathbf{1} \leq \mathbf{x} \leq \mathbf{1},\ \exists j : x_j < 1,\ \mathbf{x} = f\left(\mathbf{x}, \boldsymbol{\mu}\right)$$

In [13] it is proven that if $\mu_i > \tau_\mu, \forall i$, no misleading equilibrium, nor periodic or aperiodic sequence $\mathbf{x}^{(k)}$ is generated by (7) for any initial state $\mathbf{x}^{(0)}$. Thus (7) converges to $\mathbf{x} = +\mathbf{1}$ that is the only equilibrium allowed. This equilibrium leads the VNs to correct decisions, since the a posteriori LLR is equal to $O_i = 2 + \lambda_i$.[3] The corresponding threshold for each channel LLR λ_i can be taken as $\tau = \tau_\mu/(d_H - 1)$. If $\lambda_i > \tau, \forall i$ then $\mu_i > (d_H - 1)\tau = \tau_\mu, \forall i$. Since in this condition there exist no equilibrium with $\mathbf{x} \neq +\mathbf{1}$, the decoder cannot be trapped by this AS.

If no channel LLR can take values below the AS threshold τ, the AS cannot trap the decoder, and we say it is *deactivated*. As shown in [13], a practical way to deactivate an AS with threshold $\tau < 0$, is by setting different saturation levels λ_{max} and E_{max} for the channel and extrinsic LLRs, representing them with a different number of bits, say q_I and q, respectively. If $\tau < -\lambda_{max}/E_{max}$, the AS is deactivated.

The AS of Fig. 2 (a) has threshold $\tau = 0$ since $\tau_\mu = 0$, and it cannot be deactivated. We verified by simulation over a real GLDPC graph, namely $\mathcal{C}_1$, with eH (128,120) component codes and blocksize $N_v = 16384$, that these ASs can indeed trap the iterative decoder. Using $\lambda_{max} = 7$ for the channel LLRs and increasing values E_{max} by using $q = 4, 5$ and 6 bits to represent the extrinsic LLRs, did lower the Word Error Rate (WER) contribution of each one of these ASs (from $3 \cdot 10^{-8}$, to $2.5 \cdot 10^{-9}$ and $2 \cdot 10^{-11}$ respectively, at SNR $E_s/N_0 = 3.6$ dB, 20 iterations) but they could trap the decoder anyway.

A different type of AS, also found in the GLDPC graph, has the subgraph drawn in Fig. 2 (b) and can be analyzed with the same method. Since its threshold $\tau = -2/3$, it can trap the decoder with $q = 4$, but it is deactivated with $q = 5$ since $-\lambda_{max}/E_{max} = -7/15 > \tau$. Importance Sampling (IS) simulation with received vectors biased in the direction of these ASs, did not deliver any error event with $q \geq 5$.

An intuitive picture of the decoders behavior with these ASs is shown in Fig. 3 where the two plots (a) and (b) refer to the ASs of Fig. 2 (a) and (b), respectively. We plot the received vectors that generated a decoding failure, separating the two components r_1 and r_2: r_1 is the component (normalized by $1/a$) along the direction joining the a-length transmitted vector $+\mathbf{1}$ and the a-length AS vector $-\mathbf{1}$; r_2 is the orthogonal component, in the a dimensional subspace. In Fig. 3 (a) we see that error events are registered for any value of q. The

[2] In our notation, $\mathbf{1}$ is the all-ones column vector and the inequalities when applied to vectors are to be meant component-wise.

[3] Here, the dynamic range of λ_i is assumed not higher than E'_i and E''_i.

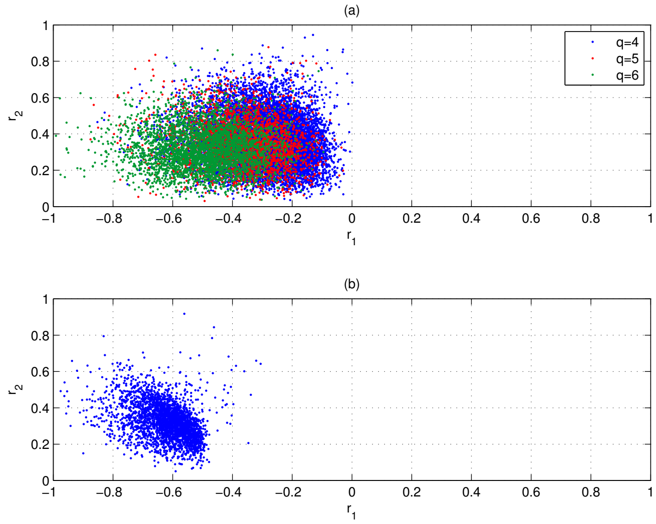


Fig. 3. Error regions for the iterative GLDPC decoders with the small Absorbing Sets of Fig. 2 (a) and (b).

error region becomes slightly smaller increasing q, but does not disappear. On the contrary in Fig. 3 (b) we see error events for the $q = 4$ decoder only, and the error region is faraway from $+\mathbf{1}$. Even with $q = 4$, the WER contribution of each of the AS of Fig. 2 (b) is much smaller (approximately $7 \cdot 10^{-12}$) and the error-floor is dominated by the ASs of Fig. 2 (a).

B. GLDPC Absorbing Sets Definition

The most important difference between the two ASs of Fig. 2 is that their CNs receive a different number of positive messages e from the external graph. We define the *degree of dissatisfaction* (o) of a CN by negative decisions about the VNs in $\mathcal{D}$, as the number of messages received from outside the AS. In the AS we need to consider all CNs with a degree of dissatisfaction $o \leq w - 2$, where w is the Hamming weight of the codeword of $\mathcal{C}$ considered. In fact, a CN with $o = w - 1$ behaves as the external graph. In other words, we include in the AS all the CNs that exchange at least two messages with the VNs of degree two in the AS.

Definition III.1. *In the Tanner graph of a binary GLDPC code, with VN-degree $d_v = 2$, an Absorbing Set is a subgraph with a subset $\mathcal{E}$ of the CNs, and a subset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ of the VNs, where $\mathcal{D}_i$ are the VNs in $\mathcal{D}$ with i neighboring CNs in $\mathcal{E}$, if*

1) *$\mathcal{D}$ is the union of low Hamming weight codewords for each CN $c \in \mathcal{E}$.*
2) *$\mathcal{E}$ is the subset of the neighboring CNs of $\mathcal{D}$, that are connected at least twice to $\mathcal{D}_2$.*

If we are interested in the smallest ASs, we need to consider the minimum weight (d_H) codewords for each CN in $\mathcal{E}$. Each AS can be classified by a triplet (a, b, o) where $a = |\mathcal{D}|$, $b = |\mathcal{E}|$ and o is the degree of dissatisfaction of the CNs in $\mathcal{E}$. For instance, the AS in Fig. 2 (a) is a (5,2,1) AS, whereas in Fig. 2 (b) we have a (6,2,2) AS.[4] We can imagine other ASs but these two have the minimum size a. However, constraints on the adjacency matrix can be easily imposed to exclude these small ASs. In the next subsection we discuss this topic.

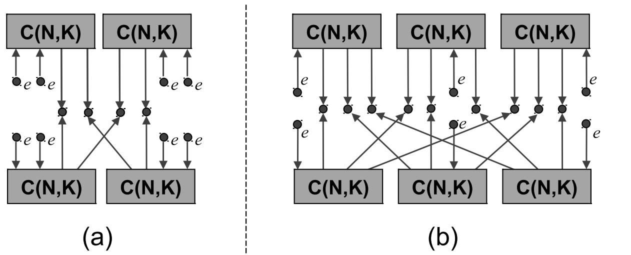


Fig. 4. Smallest ASs compatible with girth-8 constrained adjacency matrix and $d_H = 4$ component codes: AS (12,4,2) (a) and AS (15,6,1) (b).

C. Absorbing Sets of Girth-Constrained GLDPC Codes

Subgraphs like those drawn in Fig. 2, can be found in GLDPC codes with random interleaving π, but they cannot occur in graphs with adjacency matrix Γ of *girth* 8, i.e., with the property that any two CNs share no more than one VN (see, for instance, [16] or [4]). Consider again $d_H = 4$ component codes, but with a girth-8 adjacency matrix Γ. The smallest possible AS that can exist in this GLDPC graph is a (12,4,2) AS, that involves four CNs with $o = 2$ and 12 VNs, and it is represented in Fig. 4 (a). The corresponding system of equations (6) returns a threshold $\tau = -2/3$. These ASs can therefore be easily deactivated.

The smallest possible AS with $o = 1$ for all CNs is the (15,6,1) AS shown in Fig. 4 (b). This is a more dangerous AS, and cannot be deactivated since its threshold is $\tau = 0$. We checked the behavior of these ASs by IS simulation over a GLDPC graph $\mathcal{C}_2$ with girth-8 adjacency matrix Γ built by circulant blocks [16], extended Hamming (64,57) component codes and blocksize $N_v = 32768$ ($R = 25/32$). We verified that the ASs shown in Fig. 4 (b) can trap iterative decoders with $q = 4$, 5 or 6 bits for the representation of the extrinsic LLRs. In particular at $E_s/N_0 = 2.5$ dB we found that each AS contribution to the total WER is $3 \cdot 10^{-19}$, $6 \cdot 10^{-22}$ and $5 \cdot 10^{-26}$ with $q = 4$, 5 or 6, respectively.

IV. Search and Enumeration of GLDPC AS

The error probability due to the ASs with a certain topology also depends on their multiplicity. Their search and enumeration in a specific graph requires the inspection of both the adjacency matrix Γ and of the component codebook $\mathcal{C}$. This search is quite complex in general. Hamming codes exhibit the simplifying property that any pair of ones can be completed with a third one in a specific position to get a codeword. This property is inherited by eH codes: any triplet of ones can be turned into a codeword with a single fourth one in a specific position. Our inspection can thus focus on Γ, enumerating all

[4] In general, different CNs inside an AS could have different degrees o, but in our examples, this does not occur, so we take it as a scalar value.

triplets of VNs shared by two CNs. The two codewords of weight 4 including those VNs can be identified later.

We want to enumerate in $\mathcal{C}_1$ the ASs (5,2,1) with subgraph shown in Fig. 2 (a). The GLDPC code $\mathcal{C}_1$ has blocksize $N_v =$ 16384 bits, constrained by $N_c = 2N = 256$ eH(128,120) CNs, $R = 2R_c - 1 = 7/8$, and a purely random permutation matrix π. We stress the fact that the Hamming weight of these ASs ($a = 5$) is smaller than the minimum Hamming distance d_{min} of $\mathcal{C}_1$, since we checked that there is no Hamming weight 4 or 6 codeword allowed by π, hence $d_{min} \geq 8$ in $\mathcal{C}_1$. We have one AS (5,2,1) for every triplet of bits shared by two CNs, which can be enumerated. The number of bits shared by a pair of CNs under a random permutation π, as a first approximation, has a binomial probability distribution of parameter $2/N_c = 1/N$. The expected number of AS (5,2,1) in a code like $\mathcal{C}_1$ is

$$A_5 = N^2 \sum_{k=1}^{N} \binom{k}{3} \binom{N}{k} \frac{1}{N}^k \left(1 - \frac{1}{N}\right)^{N-k} \approx 2667. \quad (10)$$

The exhaustive inspection of $\mathcal{C}_1$ enumerated 2705 ASs (5,2,1). These are responsible for an error-floor at WER $8 \cdot 10^{-5}, 7 \cdot 10^{-6}$ and $5 \cdot 10^{-8}$ for Max-Log decoders with $q =$ 4, 5 and 6, respectively (at $E_s/N_0 = 3.6$ dB, 20 iterations).

Later, we chose eH(64,57) component codes for a Quasi-Cyclic GLDPC code $\mathcal{C}_2$ with blocksize $N_v = 32768$, $R = 2R_c - 1 = 25/32$ and a girth-8 adjacency matrix Γ built as in [16]. The matrix Γ has $d_v = 2$ row-blocks of $N = 64$ circulant matrices of size $S \times S$, with $S = 512$. The shifts of the first row-block were set to zero. The shifts of the second row-block $s_1, s_2 ... s_N$ have been chosen randomly, but all distinct to guarantee girth $g = 8$. With $g = 8$ the minimum Hamming distance of the code is $d_{min} \geq 16$ [4] and thus larger than the most critical AS analyzed, of size $a = 15$.

To enumerate the ASs (15,6,1), we need to look in the graph of $\mathcal{C}_2$, for triplets of cycles of length 8 that share 9 VNs and 6 CNs. For each triplet we have exactly one AS (15,6,1). The exhaustive inspection of the graph enumerated about $4400 \times S \approx 2.3 \cdot 10^6$ ASs (15,6,1). We can check this number against a probabilistic argument. Select three VNs, in columns c_1, c_2, c_3 from three different column-blocks of Γ, with shifts s_1, s_2, s_3 in the second row-block. We have $S^3 \binom{N}{3}$ different choices. Pick any two of these three VNs. The probability that there exists a cycle of length 8 across these two VNs is the probability that there exist in Γ two circulant blocks of shifts $s_1 \pm (c_1 - c_2) \mod S$. This probability is $(N/S)^2$ by random choice of the shifts, hence the probability that all the three pairs belong to cycles of length 8 is $(N/S)^6$. Finally, if a triplet of cycles like this exists, it is counted 6 times by this combinatorial argument. As a first approximation, the expected number of these ASs (15,6,1) is

$$A_{15} = \frac{1}{6} S^3 \binom{N}{3} \left(\frac{N}{S}\right)^6 \approx 3 \cdot 10^6. \quad (11)$$

Taking the multiplicity into account, the total estimated WER contribution of these ASs is $7 \cdot 10^{-13}, 10^{-15}$ and $2 \cdot 10^{-19}$ for Max-Log decoders with $q =$ 4, 5 and 6, respectively (at $E_s/N_0 = 2.5$ dB, 20 iterations).

V. Conclusions

In this paper we have proposed a definition for combinatorial substructures of the Tanner graph of binary VN-degree 2, GLDPC codes, that can trap practical Max-Log decoders over AWGN channels, i.e., *Absorbing Sets* of GLDPC codes. For these structures we can derive a quasi-linear model that reveals a threshold behavior similar to ASs in binary LDPC codes. The model predictions have been checked via IS simulation over two examples. Design constraints on the adjacency matrix of the code can avoid the smallest structures, but larger ASs able to trap the iterative decoders do exist. In case of extended Hamming component codes we enumerated by exhaustive search the most critical ASs and we checked our results against combinatorial arguments.

References

[1] R. Tanner, “A recursive approach to low complexity codes,” *IEEE Trans. Inf. Theory*, vol. 27, no. 5, pp. 533–547, May 1981.

[2] M. Lentmaier and K. Zigangirov, “On generalized low-density parity-check codes based on Hamming component codes,” *IEEE Commun. Letters*, vol. 3, no. 8, pp. 248–250, Aug. 1999.

[3] J. Bouthros, O. Pothier, and G. Zemor, “Generalized low density (Tanner) codes,” in *Proc. of IEEE Int. Conf. on Commun. (ICC’99)*, Vancouver, Canada, June 6-10 1999, pp. 441–445.

[4] M. Lentmaier, G. Liva, E. Paolini, and G. Fettweis, “From product codes to structured generalized LDPC codes,” in *Proc. of 5th Int. ICST Conf. on Commun. and Networking*, Beijing, China, Aug. 25-27 2010, pp. 1–8.

[5] N. Miladinovic and M. Fossorier, “Generalized LDPC codes and generalized stopping sets,” *IEEE Trans. Commun.*, vol. 56, no. 2, pp. 201–212, Feb. 2008.

[6] C. Di, D. Proietti, E. Telatar, T. J. Richardson, and R. Urbanke, “Finite-Length Analysis of Low-Density Parity-Check Codes on the Binary Erasure Channel,” *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1570 – 1579, June 2002.

[7] E. Paolini, M. Flanagan, M. Chiani, and M. Fossorier, “Spectral shape of check-hybrid GLDPC codes,” in *Proc. IEEE Int. Conf. Communications*, Capetown, SA, May 23-27 2010, pp. 1–6.

[8] L. Dolecek, Z. Zhang, V. Anantharam, M. Wainwright, and B. Nikolic, “Analysis of absorbing sets and fully absorbing sets of array-based LDPC codes,” *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 181–201, Jan. 2010.

[9] Z. Zhang, L. Dolecek, B. Nikolic, V. Anantharam, and M. Wainwright, “Design of LDPC decoders for improved low error rate performance: quantization and algorithm choices,” *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3258–3268, Nov. 2009.

[10] B. Amiri, J. Kliewer, and L. Dolecek, “Analysis and enumeration of absorbing sets for non-binary graph-based codes,” *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 398–409, Feb. 2014.

[11] A. Hareedy, C. Lanka, and L. Dolecek, “A general non-binary LDPC code optimization framework suitable for dense flash memory and magnetic storage,” *IEEE J. Select. Areas Commun.*, vol. 34, no. 9, pp. 2402–2415, Sep. 2016.

[12] B. K. Butler and P. H. Siegel, “Error floor approximation for LDPC codes in the AWGN channel,” *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7416–7441, Dec. 2014.

[13] A. Tomasoni, S. Bellini, and M. Ferrari, “Thresholds of absorbing sets in low-density-parity-check codes,” *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3238–3249, Aug. 2017.

[14] M. Ferrari, A. Tomasoni, and S. Bellini, “Analysis of practical LDPC decoders in Tanner graphs with absorbing sets,” in *Proc. of IEEE Info. Theory Workshop*, Kaohsiung, Taiwan, Nov. 6-10 2017, pp. 141–145.

[15] J. Hagenauer, E. Offer, and L. Papke, “Iterative decoding of binary block and convolutional codes,” *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 429–445, Feb. 1996.

[16] G. Liva, W. Ryan, and M. Chiani, “Quasi-cyclic generalized LDPC codes with low error floors,” *IEEE Trans. Commun.*, vol. 56, no. 1, pp. 49–57, Jan. 2008.

Lengthening and Extending Binary Private Information Retrieval Codes

Hsuan-Yin Lin and Eirik Rosnes
Simula@UiB, N–5020 Bergen, Norway
(Emails: hsuan-yin.lin@ieee.org and eirikrosnes@simula.no)

***Abstract*—It was recently shown by Fazeli *et al.* that the storage overhead of a traditional t-server private information retrieval (PIR) protocol can be significantly reduced using the concept of a *t-server PIR code*. In this work, we show that a family of t-server PIR codes (with increasing dimensions and blocklengths) can be constructed from an existing t-server PIR code through lengthening by a single information symbol and code extension by at most $\lceil t/2 \rceil$ code symbols. Furthermore, by extending a code construction notion from Steiner systems by Fazeli *et al.*, we obtain a specific family of t-server PIR codes. Based on a code construction technique that lengthens and extends a t-server PIR code simultaneously, a basic algorithm to find good (i.e., small blocklength) t-server PIR codes is proposed. For the special case of $t = 5$, we find provably optimal PIR codes for code dimensions $k \leq 6$, while for all $7 \leq k \leq 32$ we find codes of smaller blocklength than the best known codes from the literature. Furthermore, in the case of $t = 8$, we also find better codes for $k = 5, 6, 11, 12$. Numerical results show that most of the best found 5-server PIR codes can be constructed from the proposed family of codes connected to Steiner systems.**

I. Introduction

Private information retrieval (PIR) has attracted significant attention for well over a decade since its introduction by Chor *et al.* in [1]. A formal PIR protocol allows to privately retrieve a single file among the servers storing it without revealing any information about the requested file to each individual server. Traditional PIR protocols operate on a database of n bits, which is replicated among several servers to achieve PIR. Thus, the storage overhead of traditional PIR protocols is at least 2, and the overall goal is to reduce the total upload and download cost of the protocol.

PIR for distributed storage systems was first addressed in [2]. For distributed storage systems the size of the requested file is typically much larger than the number of files, and thus the upload cost is much lower than the download cost. Hence, only the download cost is considered, as opposed to traditional PIR protocols. Recent work on PIR protocols for distributed storage systems typically assumes that the storage code is given, and then the PIR protocol is designed as a second layer to the system [3], [4]. This is in contrast to the work by Fazeli *et al.* in [5], where, in order to reduce the storage overhead of traditional PIR protocols, the concept of a *t-server PIR code* was proposed. A t-server PIR code is an $[n, k]$ linear code satisfying the so-called t-PIR property, i.e., for every information symbol, there exist t mutually disjoint subsets of $\{1, 2, \ldots, n\}$ such that it can be recovered from the code symbols indexed by any of these t subsets. By employing an $[n, k]$ t-server PIR code, they have shown that all known t-server information-theoretic PIR protocols can be emulated by a coded PIR protocol with storage overhead equal to n/k.

Finding good codes that operate efficiently with a small storage overhead, i.e., designing a t-server PIR code with a small blocklength for a given dimension, is an important research challenge. In [5], an insightful series of t-server PIR code constructions based on existing code construction techniques were presented. In the recent work of [6], the authors found that the so-called *shortened projective Reed Muller (SPRM)* codes are good t-server PIR codes for $t = 2^\ell - 1$ and 2^ℓ where ℓ is a positive integer. For $t = 3, 4$, it was shown in [6] that SPRM codes are indeed optimal in the sense of achieving a lower bound on the blocklength of a t-server PIR code.

In this work, we will show that a t-server PIR code with small blocklength can be constructed by lengthening and extending an existing PIR code. Furthermore, we prove that a certain family of codes associated with Steiner systems possesses the t-PIR property. Since optimal codes for $t \leq 4$ are known (see [5], [6]), we mainly focus on the special case of $t = 5$ (or, equivalently, $t = 6$) for which we show that provably optimal PIR codes can be constructed from lengthening and extending an existing PIR code for code dimensions $k \leq 6$, while for all $7 \leq k \leq 32$ we find codes of smaller blocklength than the best known codes from the literature. Moreover, we also show that for certain values of k, SPRM codes are not optimal for $t = 8$.

II. Definitions and Preliminaries

Throughout this paper, we will focus on binary codes only. Component-wise addition of vectors from a vector space will be written as normal addition, and as is customary in coding theory, we denote row vectors by boldface italic Roman letters, e.g., $\boldsymbol{x}$. However, sometimes we will slightly abuse this notational convention by using $\boldsymbol{c}$ to refer to a column vector. Moreover, whether an all-zero vector $\mathbf{0}$ (or an all-one vector $\mathbf{1}$) is a row vector or a column vector will become clear from the context. The Hamming weight of a binary vector $\boldsymbol{x}$ is denoted by $w_{\mathrm{H}}(\boldsymbol{x})$ throughout the paper.

A. t-Server PIR Codes

Definition 1: Consider an $[n, k]$ linear code $\mathscr{C}$ and its corresponding generator matrix $\mathsf{G} \triangleq [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n]$. This $[n, k]$

This work was partially funded by the Research Council of Norway (grant 240985/F20).

code is said to be an $[n,k;t]$ PIR code if for every $i \in \mathbb{N}_k \triangleq \{1,2,\ldots,k\}$, there exist t mutually disjoint sets $\mathcal{R}_h^{(i)}$, $h \in \mathbb{N}_t$, such that

$$\boldsymbol{e}_i \triangleq (\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0)^{\mathsf{T}} = \sum_{j\in\mathcal{R}_h^{(i)}} \boldsymbol{c}_j, \quad \forall\, h \in \mathbb{N}_t,$$

where superscript "$^{\mathsf{T}}$" denotes vector transposition. We also say that such a code $\mathscr{C}$ (or G) has the t-PIR property. Moreover, given a message symbol u_i, $i \in \mathbb{N}_k$, those mutually disjoint sets $\mathcal{R}_h^{(i)}$, $h \in \mathbb{N}_t$, are called the recovering sets for u_i.

For given values of k and t, the minimum value of n for which an $[n,k;t]$ PIR code exists is of great interest. This motivates us to look at a related parameter in conventional coding theory: the length of the shortest binary linear code with dimension k and minimum Hamming distance d. The smallest blocklength of a linear code for fixed values of (k,d) has been discussed extensively in the existing literature. Note that our notation of an $[n,k;t]$ PIR code should not be confused with the usual three parameters notation of an $[n,k,d]$ linear code, where the third parameter d denotes the minimum Hamming distance of the $[n,k]$ code. We make the following definitions.

Definition 2:

$$N_{\mathrm{P}}(k,t) \triangleq \min\{n\colon \text{an } [n,k;t] \text{ binary PIR code exists}\}.$$
$$N(k,d) \triangleq \min\{n\colon \text{an } [n,k,d] \text{ binary linear code exists}\}.$$

B. Bounds for t-Server PIR Codes

It is well-known that the minimum Hamming distance d of a t-server PIR code must be at least t [7].

Proposition 1: If an $[n,k;t]$ PIR code exists, then its minimum Hamming distance d must satisfy $d \geq t$.

Corollary 1: For given values of k and t, $N_{\mathrm{P}}(k,t)$ is lower-bounded by the smallest blocklength n such that an $[n,k,t]$ code exists, i.e., $N_{\mathrm{P}}(k,t) \geq N(k,t)$.

Proof: See the extended version [8]. ■

In [6], a lower bound on the minimum blocklength $N_{\mathrm{P}}(k,t)$ for any *systematic* $[n,k;t]$ PIR code was presented. As shown in [9], the bound from [6] also holds for any binary $[n,k;t]$ PIR code. The lower bound from [6], denoted by $L_{\mathrm{P}}(k,t)$, is

$$L_{\mathrm{P}}(k,t) \triangleq k + \left\lceil \sqrt{2k+\frac{1}{4}} + \frac{1}{2} \right\rceil + t - 3, \quad t \geq 3.$$

It can easily be verified that in general $N(k,t) \geq L_{\mathrm{P}}(k,t)$ for small values of $t > 4$. In fact, we will show in Section V that $N(k,t)$ is a tighter lower bound on $N_{\mathrm{P}}(k,t)$ than $L_{\mathrm{P}}(k,t)$ for $t = 6$.

Some useful upper and lower bounds on $N_{\mathrm{P}}(k,t)$ were provided by Fazeli *et al.* in [5]. Together with the constructions introduced therein, the authors provided an upper bound table on $N_{\mathrm{P}}(k,t)$ for all values of $k \leq 32$ and $t \leq 16$. We briefly summarize their results below.

Lemma 1 (Lemmas 13 and 14 in [5]):

(a) $N_{\mathrm{P}}(k,t+t') \leq N_{\mathrm{P}}(k,t) + N_{\mathrm{P}}(k,t')$,
(b) $N_{\mathrm{P}}(k+k',t) \leq N_{\mathrm{P}}(k,t) + N_{\mathrm{P}}(k',t)$,
(c) $N_{\mathrm{P}}(k,t) \leq N_{\mathrm{P}}(k+1,t) - 1$,
(d) $N_{\mathrm{P}}(k,t) \leq N_{\mathrm{P}}(k,t+1) - 1$, and
(e) if t is odd, then $N_{\mathrm{P}}(k,t+1) = N_{\mathrm{P}}(k,t) + 1$.

III. Code Constructions

In this section, we first present a code construction by lengthening and extending a given PIR code, and then present an extension of a code construction inspired by Steiner systems proposed by Fazeli *et al.* in [5]. An earlier work constructing PIR codes (and even stronger batch codes) for $t = k$ based on Steiner systems (and more general block designs) was presented in [10].

A. Lengthening and Extending PIR Codes

In the following theorem, we will investigate an important property of a PIR code with an arbitrary positive integer t.

Theorem 1: For any given $t \in \mathbb{N} \triangleq \{1,2,\ldots\}$, we have

$$N_{\mathrm{P}}(k+1,t) \leq N_{\mathrm{P}}(k,t) + \left\lceil \frac{t}{2} \right\rceil.$$

Proof: See the extended version [8]. ■

Theorem 1 is an improved version of part (b) of Lemma 1 for $k' = 1$, while for $k' > 1$, it is an improved version only if $k'\lceil \frac{t}{2} \rceil < N_{\mathrm{P}}(k',t)$. This theorem suggests that for a given even value of t, a new t-server PIR code can always be generated by adding one information symbol and appending at most $t/2$ code symbols to the original t-server PIR code.

Next, we will discuss a special family of systematic codes that will help in the numerical search for good PIR codes with small blocklength, especially when k is large.

B. Construction of PIR Codes Based on Steiner Systems

In [5], a systematic code construction based on Steiner systems was proposed, in which the authors introduce a representation method of systematic codes, and give a sufficient (but not necessary) condition for constructing PIR codes.

Definition 3: Let $\mathscr{P}_k = \{\mathcal{P}_j\}_{j=1}^{r}$ be a collection of subsets of $\mathbb{N}_k$. A systematic $[n = k+r, k]$ code $\mathscr{C}$ can be represented by defining the codewords of $\mathscr{C}$ as $\boldsymbol{x} \triangleq (u_1,\ldots,u_k,x_{k+1},\ldots,x_{k+r})$, where $u_1,\ldots,u_k$ are the information bits of the code and each redundancy bit x_{k+j} is defined as $x_{k+j} \triangleq \sum_{i\in\mathcal{P}_j} u_i$, $j \in \mathbb{N}_r$.

We denote the constructed code by $\mathscr{C}(\mathscr{P}_k)$. Furthermore, for the sake of notational convenience, we define $\mathcal{J}^{(i)} \triangleq \bigl\{j \in \mathbb{N}_r : i \in \mathcal{P}_j\bigr\}$ to be the set of indices $j \in \mathbb{N}_r$ such that $i \in \mathcal{P}_j$.

The systematic generator matrix G of this code can be written as $\mathsf{G} = \bigl[\mathsf{I}_k | \mathsf{P}_{k\times r}\bigr]$, where I_k is the $k \times k$ identity matrix and the $k \times r$ redundancy matrix $\mathsf{P}_{k\times r} = \{p_{ij}\}_{1\leq i\leq k,\,1\leq j\leq r}$ is defined by

$$p_{ij} \triangleq \begin{cases} 1, & \text{if } i \in \mathcal{P}_j, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 2 (Lemma 7 in [5]): Suppose that a collection $\mathscr{P}_k = \{\mathcal{P}_j\}_{j=1}^{r}$ satisfies the following properties.

1) For all $i \in \mathbb{N}_k$, $\bigl|\mathcal{J}^{(i)}\bigr| \geq t-1$, and
2) for all $j \neq j' \in \mathbb{N}_r$, $\bigl|\mathcal{P}_j \cap \mathcal{P}_{j'}\bigr| \leq 1$.

Then, the corresponding systematic code $\mathscr{C}(\mathscr{P}_k)$ is a t-server PIR code.

The above lemma only leads to an absorbing upper bound on the redundancy $N_{\mathrm{P}}(k,t)-k$ for fixed t and sufficiently large k, which shows that it is equal to $O(\sqrt{k})$. However, for smaller values of the parameter k, whether or not this upper bound is tight is still unknown. Moreover, in [5] a similar PIR code construction based on constant-weight codes was provided, where all rows of $\mathsf{P}_{k\times r}$ have constant weight and a given minimum Hamming distance.

It is known that the minimum Hamming distance d of a PIR code must be larger than or equal to the desired parameter t (see Proposition 1), and so are the row Hamming weights of any generator matrix G for the code. Hence, it is reasonable to change the sufficient condition of $\bigl|\mathcal{J}^{(i)}\bigr| \geq t-1$ in Lemma 2 to $\bigl|\mathcal{J}^{(i)}\bigr| = t-1$, $\forall\, i \in \mathbb{N}_k$.

Motivated by Steiner systems, we define a more elaborate systematic code family as follows.

Definition 4: For any integer $t \in \mathbb{N}$ and a given collection $\mathscr{P}_k = \{\mathcal{P}_j\}_{j=1}^{r}$ of subsets of $\mathbb{N}_k$, we say that a systematic code $\mathscr{C}(\mathscr{P}_k)$ (or its corresponding generator matrix) has property S_t if all of the following conditions are satisfied.

1) $\mathcal{P}_r = \mathbb{N}_k$,
2) $\bigl|\mathcal{J}^{(i)}\bigr| = t-1$ for all $i \in \mathbb{N}_k$,
3) $\bigl|\mathcal{P}_j \cap \mathcal{P}_{j'}\bigr| \leq 1$ for all $j \neq j' \in \mathbb{N}_{r-1}$, and
4) for any given $m \in \mathbb{N}_k$, there exists a subset $\mathcal{I}(m) \subseteq \mathbb{N}_k$ with $\mathcal{I}(m) \cap \bigl(\bigcup_{j\in\mathcal{J}^{(m)}\setminus\{r\}} \mathcal{P}_j\bigr) = \emptyset$ and a subset $\mathcal{V}^{(m)} \subseteq \mathbb{N}_{r-1}$ with $\mathcal{V}^{(m)} \cap \mathcal{J}^{(m)} = \emptyset$ such that
$$u_m + \sum_{i\in\mathcal{I}(m)} u_i + \sum_{j\in\mathcal{V}^{(m)}} \sum_{i\in\mathcal{P}_j} u_i = \sum_{i=1}^{k} u_i.$$

Similarly to Lemma 2, a systematic code with property S_t turns out to be an $[n,k;t]$ PIR code.

Lemma 3: If a systematic code $\mathscr{C}(\mathscr{P}_k)$ has property S_t, then it is an $[n=k+r,k;t]$ PIR code.

Proof: See the full version [8]. ■

The following example illustrates the code design of Lemma 3.

Example 1: For an $[n,k] = [17,8]$ systematic code, we describe it in terms of $\mathscr{P}_8$ as follows:
$$\begin{aligned}\mathscr{P}_8 \triangleq \bigl\{&\mathcal{P}_1 \triangleq \{1,2,3\}, \mathcal{P}_2 \triangleq \{1,4,6\}, \mathcal{P}_3 \triangleq \{1,5,7\},\\ &\mathcal{P}_4 \triangleq \{2,4,8\}, \mathcal{P}_5 \triangleq \{2,5,6\}, \mathcal{P}_6 \triangleq \{3,4,7\},\\ &\mathcal{P}_7 \triangleq \{3,5,8\}, \mathcal{P}_8 \triangleq \{6,7,8\}, \mathcal{P}_9 \triangleq \mathbb{N}_8\bigr\}.\end{aligned}$$

One can see that $r=9$ and that the systematic code $\mathscr{C}(\mathscr{P}_8)$ has property S_5. Here, condition 4) can be verified by the following observations (e.g., take $m=1,8$):
$$\begin{aligned}&\mathcal{J}^{(1)} = \{1,2,3,9\},\ \mathcal{J}^{(8)} = \{4,7,8,9\},\\ &\mathcal{I}(1) = \{8\},\ \mathcal{I}(8) = \{1\},\ \mathcal{V}^{(1)} = \mathcal{V}^{(8)} = \{5,6\},\\ &\mathbb{N}_8 = \{1\} \cup \mathcal{P}_5 \cup \mathcal{P}_6 \cup \{8\}.\end{aligned}$$

Then, we can conclude that this code is a 5-server $[17,8]$ PIR code. For example, the recovering sets for the first information bit are determined by $\mathcal{R}_1^{(1)} = \{1\}$,
$$\begin{aligned}\mathcal{R}_2^{(1)} &= \{m \in \mathcal{P}_1 \colon m \neq 1\} \cup \{k+1\} = \{2,3,9\},\\ \mathcal{R}_3^{(1)} &= \{m \in \mathcal{P}_2 \colon m \neq 1\} \cup \{k+2\} = \{4,6,10\},\\ \mathcal{R}_4^{(1)} &= \{m \in \mathcal{P}_3 \colon m \neq 1\} \cup \{k+3\} = \{5,7,11\},\\ \mathcal{R}_5^{(1)} &= \{8, k+5, k+6, k+r\} = \{8,13,14,17\}.\end{aligned}$$

In fact, the idea behind Lemma 3 is to try to combine the properties of Steiner systems and part (e) of Lemma 1, in such a way that we can construct an $[n+1,k;t+1]$ PIR code from an $[n,k;t]$ PIR code when t is even.

We also remark that a systematic $[n,k;t]$ PIR code with property S_t usually has different cardinalities of its recovering sets (the so-called *non-uniform information-symbol locality* property). For instance, for the code of Example 1, each information symbol has 1 recovering set of cardinality 1, 3 recovering sets of cardinality 3, and 1 recovering set of cardinality 4. This is also in alignment with [6], where the presented PIR codes in general have recovering sets of different cardinalities. In Section V, we will show that codes having property S_5 are good 5-server PIR codes with small blocklength.

IV. Searching for Optimal PIR Codes

In this section, we present an algorithm to search for good (i.e., small blocklength) PIR codes. Since optimal codes for $t \leq 4$ are already known for all code dimensions k, we concentrate on $t=5$. Because Theorem 1 implies that we can construct a t-server PIR code by lengthening and extension, hence, combined with the idea of lexicographic code construction [11], Algorithm 1 is proposed to find a sequence of good systematic PIR codes for $t=5$.[1]

Initially, we choose the best known $[n,k;5]$ code with a systematic generator matrix in which all rows have weight 5. Note that for small values of n and k, such a code is not too difficult to find. As an example, the generator matrix G of a systematic $[8,2;5]$ code in which all rows have weight 5 is
$$\mathsf{G} = \left[\begin{array}{cc|ccccc|c} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1\\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1\end{array}\right]. \tag{1}$$

The outer while loop of Algorithm 1 increases a counter (denoted by i) from 1 to $\binom{r+w}{4}$ (the counter runs over all possible length-$(r+w)$ binary vectors of weight 4). The function $\mathtt{LengtheningExtending}(\mathsf{G}_{\text{best}}, \boldsymbol{z})$ in Line 6 of Algorithm 1 is defined by
$$\tilde{\mathsf{G}} \triangleq \left[\begin{array}{c|c|c} \mathsf{I}_{k_{\text{best}}} & \mathbf{0} & \mathsf{P}_{k_{\text{best}}\times(r+w)}\\ \hline \mathbf{0} & 1 & \boldsymbol{z}\end{array}\right], \qquad \underbrace{\hphantom{\mathsf{I}_{k_{\text{best}}}\ \mathbf{0}\ 1}}_{k_{\text{best}}+1}\underbrace{\hphantom{\mathsf{P}_{k_{\text{best}}\times(r+w)}}}_{r+w}$$

where $\mathsf{G}_{\text{best}} = \bigl[\mathsf{I}_{k_{\text{best}}} \big| \mathsf{P}_{k_{\text{best}}\times(r+w)}\bigr]$ and $w_{\mathrm{H}}(\boldsymbol{z}) = 4$.[2] Note that if $w=2$, it follows from the proof of Theorem 1 in [8] that $k_{\text{best}} \geq k+1$; explaining why we choose $1 \leq w \leq 2$ from the beginning. Furthermore, notice that for $w=1$, sometimes

[1] In general, this algorithm can be applied for any t. The main reason why we focus on small values of t is that when t is increasing, the complexity to determine whether a code has the t-PIR property is also increasing.

[2] Note that the definition of $\tilde{\mathsf{G}}$ guarantees that G_{best} is always in systematic form in each iteration.

```
Algorithm 1: Searching for optimal 5-server PIR codes
  Input  : A systematic constant row-weight-5 generator
           matrix $\mathsf{G} = [\mathsf{I}_k|\mathsf{P}_{k\times r}]$ for an $[n,k;5]$ code, and a
           given $w \in \mathbb{N}_2$.
  Output: A systematic constant row-weight-5 generator
           matrix $\mathsf{G}_{\text{best}}$ for an $[n_{\text{best}}, k_{\text{best}}; 5]$ code, where
           $k_{\text{best}} \geq k$ is the largest possible code dimension
           found and $n_{\text{best}} = k_{\text{best}} + r + w$.
 1 $\mathsf{G}_{\text{best}} \leftarrow [\mathsf{I}_k|\mathsf{P}_{k\times r}|\mathsf{O}_{k\times w}]$, $k_{\text{best}} \leftarrow k$
 2 /* $\mathsf{O}_{k\times w}$ is a $k \times w$ all-zero matrix                    */
 3 $i \leftarrow 1$
 4 $\boldsymbol{z} \leftarrow$ the row vector $(1,1,1,1,0,\ldots,0)$ of length $r + w$
 5 while $i \leq \binom{r+w}{4}$ do
 6     $\tilde{\mathsf{G}} \leftarrow$ LengtheningExtending($\mathsf{G}_{\text{best}}, \boldsymbol{z}$)
 7     $\tilde{d} \leftarrow$ minimum Hamming distance of $\tilde{\mathsf{G}}$
 8     /* we simply say a code $\tilde{\mathsf{G}}$ is the set of all rows of $\tilde{\mathsf{G}}$     */
 9     if $\tilde{d} \geq 6$ then
10         if $\tilde{\mathsf{G}}$ has the 5-PIR property then
11             $\mathsf{G}_{\text{best}} \leftarrow \tilde{\mathsf{G}}$, $k_{\text{best}} \leftarrow k_{\text{best}} + 1$
12         else
13             return $(\mathsf{G}_{\text{best}}, k_{\text{best}})$
14         end
15     end
16     $i \leftarrow i + 1, \boldsymbol{z} \leftarrow$ Lexical($\boldsymbol{z}$)
17 end
18 if $k_{\text{best}} = k$ then
19     $\mathsf{G}_{\text{best}} \leftarrow \mathsf{G}$
20 end
21 return $(\mathsf{G}_{\text{best}}, k_{\text{best}})$
```

the algorithm only results in the original input code. We also verify whether $\tilde{d} \geq 6$ or not in Line 9 of Algorithm 1. This is to ensure that the resulting code generated by $\tilde{\mathsf{G}}$ can potentially satisfy Proposition 1.[3] Finally, given a vector $\boldsymbol{z}$, Lexical($\boldsymbol{z}$) generates the next lexicographical constant-weight $\boldsymbol{z}$ of length $r + w$, e.g., Lexical$(\boldsymbol{z}) = (1,1,1,0,1,0,\ldots,0)$ for $\boldsymbol{z} = (1,1,1,1,0,\ldots,0)$.

We also remark that the resulting k_{best} from Algorithm 1 strongly depends on the selected $\mathsf{G} = [\mathsf{I}_k|\mathsf{P}_{k\times r}]$ and the given w in the input. It is difficult to predict whether the corresponding blocklength n_{best} is good or not. For example, given the systematic $[n = k + r, k; t] = [8, 2; 5]$ code defined in (1) and $w = 1$, the output from Algorithm 1 is an $[n_{\text{best}}, k_{\text{best}}; 5] = [11, 4; 5]$ code without property S_5, while for $w = 2$, Algorithm 1 results in an $[n_{\text{best}}, k_{\text{best}}; 5] = [13, 5; 5]$ code with property S_5 (see Section V that follows). Now, for code dimension $k = 4$, the $[11, 4; 5]$ code is better than the $[12, 4; 5]$ code obtained by shortening the optimal $[13, 5; 5]$ code. Hence, for a fixed code dimension k, to find a good 5-server PIR code with small blocklength, we have to compare all the resulting $[n, k; 5]$ codes found by Algorithm 1.

In general, the complexity of exhaustively examining the t-PIR property for a given code becomes infeasible for large n and k, even for $t = 5$. However, according to our numerical results, for small code dimensions k, an optimal 5-server PIR code often has property S_5. Therefore, we investigate a sequence of good PIR codes with respect to property S_5. In fact, a sequence of good codes with small blocklength can always be generated by lengthening by one information symbol and extending at most 2 coordinates from a smaller-sized code with property S_5, as shown in the theorem below.

Theorem 2: For any given values of n and k, if a systematic $[n, k]$ code has property S_5, then there must exist a systematic $[n + 3, k + 1]$ code that also has property S_5.

Proof: See the details in the extended version [8]. ■

Based on Theorem 2, we can slightly modify Algorithm 1 to investigate 5-server PIR codes with property S_5. First, we replace the input generator matrix by a generator matrix $\mathsf{G} = [\mathsf{I}_k|\mathsf{P}_{k\times(r-1)}|\mathbf{1}]$ with property S_5, and modify the starting G_{best} to $[\mathsf{I}_k|\mathsf{P}_{k\times(r-1)}|\mathsf{O}_{k\times w}|\mathbf{1}]$ in Line 1 of Algorithm 1. The function LengtheningExtending($\mathsf{G}_{\text{best}}, \boldsymbol{z}$) for $\mathsf{G}_{\text{best}} = \bigl[\mathsf{I}_{k_{\text{best}}}|\mathsf{P}_{k_{\text{best}}\times(r+w-1)}|\mathbf{1}\bigr]$ in Line 6 of Algorithm 1 is accordingly re-defined as

$$\tilde{\mathsf{G}} \triangleq \left[\begin{array}{c|c|c|c} \mathsf{I}_{k_{\text{best}}} & \mathbf{0} & \mathsf{P}_{k_{\text{best}}\times(r+w-1)} & \mathbf{1} \\ \hline \mathbf{0} & 1 & \boldsymbol{z} & 1 \end{array}\right], \qquad \underbrace{\mathsf{I}_{k_{\text{best}}}\ \ \mathbf{0}}_{k_{\text{best}}+1} \quad \underbrace{\mathsf{P}_{k_{\text{best}}\times(r+w-1)}}_{r+w-1}$$

where $w_{\mathrm{H}}(\boldsymbol{z}) = 5 - 2 = 3$. Notice that the outer while loop counter now should increase from 1 to $\binom{r+w-1}{3}$, and the initial $\boldsymbol{z}$ in Line 4 should be replaced by the length-$(r + w - 1)$ vector $\boldsymbol{z} = (1,1,1,0,\ldots,0)$. In fact, there is no need to modify Line 9 of Algorithm 1, since the resulting $\tilde{\mathsf{G}}$ will again satisfy conditions 1)–3) of Definition 4.[4] As a result, after the modifications to Algorithm 1 outlined above, and if Line 10 of Algorithm 1 is replaced by the verification of property S_5 for $\tilde{\mathsf{G}}$, we are able to find good 5-server PIR codes with property S_5 for large code dimensions $k \geq 16$ (see Section V below). From Theorem 2 it follows that if $w = 2$, $k_{\text{best}} \geq k + 1$.

V. Numerical Results

In this section, upper bounds on $N_{\mathrm{P}}(k, t)$ for $1 \leq k \leq 32$ and $t = 4, 6, 8$ are summarized in Table I. In particular, for $t = 6$, we also present the numerical results obtained using the search algorithm from Section IV. Entries for which strictly better codes are found than in the current literature are marked in bold. In comparison with the obtained improved upper bound, a lower bound on $N_{\mathrm{P}}(k, 6)$ is also given. For $t = 4$, the SPRM codes provided in [6] are optimal. More specifically, the blocklength is equal to the lower bound $L_{\mathrm{P}}(k, 4)$.

In order to show how good our constructed 6-server PIR codes are, we also list the best (smallest) known blocklength

[3] Since the construction guarantees that all rows have equal Hamming weights, the Hamming distance between any pair of rows is even, i.e., the necessary condition $\tilde{d} \geq 5$ is equivalent to $\tilde{d} \geq 6$.

[4] Note that the construction of $\tilde{\mathsf{G}}$ will make all the row-weights of $\tilde{\mathsf{G}}$ equal to 5 and the last column equal to the all-one vector (i.e., conditions 1) and 2) of Definition 4 are satisfied). In order to satisfy condition 3) of Definition 4, the minimum Hamming distance of $\tilde{\mathsf{G}}$ must be larger than or equal to $2\cdot(5-2) = 6$, since any two row vectors in $\tilde{\mathsf{G}}$ must have a common 1 in at most two coordinates.

TABLE I
BEST KNOWN BOUNDS ON $N_{\mathrm{P}}(k,t)$ FOR SMALL VALUES OF k AND EVEN $t = 4, 6, 8$. IN THE CASE OF $t = 6$, n_{B} DENOTES THE BEST FOUND BLOCKLENGTH BASED ON OUR PROPOSED SEARCH ALGORITHM, AND n_{U} IS DEFINED IN (2). STARRED VALUES (OR COLUMNS) CAN BE PROVED TO BE OPTIMAL, WHILE BOLD ENTRIES ARE NEW RESULTS.

$k\backslash t$	4^* [6]	6			8 [6]
		$N(k,t)$ [12]	n_{B}	n_{U}	
1	4	–	6^*	–	8^*
2	6	–	9^*	–	12^*
3	7	–	11^*	–	14^*
4	9	–	$12^{*\diamond}$	–	15^*
5	10	14	14^*	$13^!$	19
6	11	15	$15^{*\diamond}$	$14^!$	21
7	13	16	**17**	$15^!$	22
8	14	17	**18**	20	24
9	15	18	**20**	23	25
10	16	20	**21**	24	26
11	18	21	**22**	25	30
12	19	22	**23**	26	32
13	20	23	$\mathbf{25}^{\diamond}$	27	33
14	21	24	$\mathbf{27}^{\diamond}$	29	35
15	22	26	$\mathbf{28}^{\diamond}$	34	36
16	24	27	**31**	35	37
17	25	28	**32**	37	39
18	26	29	**33**	38	40
19	27	30	**35**	39	41
20	28	31	**36**	40	42
21	29	32	**37**	42	46
22	31	33	**39**	46	48
23	32	34	**40**	47	49
24	33	36	**41**	49	51
25	34	37	**42**	50	52
26	35	38	**43**	51	53
27	36	39	**44**	53	55
28	37	40	**46**	54	56
29	39	41	**47**	55	57
30	40	42	**48**	56	58
31	41	43	**50**	58	60
32	42	44	**52**	59	61

for $t = 8$ (the smallest blocklength of the SPRM codes from [6]). They will result in an improved upper bound for $t = 6$, since by part (d) of Lemma 1, $N_{\mathrm{P}}(k,6) \leq N_{\mathrm{P}}(k,8) - 2$. Hence,

$$n_{\mathrm{U}} \triangleq \min\{n_1, n_2 - 2\} \qquad (2)$$

is the best known upper bound for $t = 6$, where n_1 denotes the best known blocklength provided in [5], and n_2 is the smallest blocklength of SPRM codes for $t = 8$ provided in [6].

Note again that, according to part (e) of Lemma 1 and in order to compare our findings with [5, Table III] and [6, Table II], only even values of t are interesting. Here, for $t = 6$ the blocklengths n_{B} of Table I are obtained by adding one to the blocklengths of our best found 5-server PIR codes. We make the following remarks to Table I.

1) The superscript "$*$" indicates that the corresponding blocklength can be shown to be optimal. We use the lower bound $N(k,t)$, whose value can be obtained from [12], since $L_{\mathrm{P}}(k,6) = L_{\mathrm{P}}(k,4) + 2 \leq N(k,6)$ and no tighter lower bound for $t = 6$ is known.
2) The superscript "$\diamond$" indicates that the best found systematic $[n,k;5]$ code has a constant-weight generator matrix of row-weight 5 and without property S_5.
3) The superscript "!" indicates that the corresponding blocklength is impossible, since it is smaller than $N(k,t)$ (a contradiction to Corollary 1). We believe that the value of $n_{\mathrm{U}} = 15$ for $(k,t) = (7,6)$ in [5, Table III] was obtained from [5, Thm. 9] and should have corresponded to $(k,t) = (6,6)$ due to a misprint in [13, p. 289] in the redundancy of *type*-1 *doubly transitive invariant codes*. We believe this explains the contradictions.
4) The superscript "[$\cdot$]" indicates the reference number.

We also remark that for $t = 8$, using our algorithm we are able to find better PIR codes for certain values of k: we have obtained $n_{\mathrm{B}} = 18, 20, 29, 31$ for $k = 5, 6, 11, 12$, respectively. This indicates that the SPRM codes are not optimal for $t = 8$.

VI. CONCLUSION

In this paper, we presented a construction of a t-server PIR code by lengthening and extension of an existing PIR code. We also presented an extension of a code construction inspired by Steiner systems proposed by Fazeli *et al.*, which was used in the proposed algorithm to search for good (i.e., small blocklength) 5-server PIR codes. For code dimensions $k \leq 6$, provably optimal PIR codes were found, while for all $7 \leq k \leq 32$, codes of smaller blocklength than the best known codes from the literature were found and presented. Moreover, better 8-server PIR codes were also found for $k = 5, 6, 11, 12$.

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th IEEE Symp. Found. Comp. Sci.*, Milwaukee, WI, USA, Oct. 1995, pp. 41–50.

[2] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 856–860.

[3] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, Jul. 2016, pp. 1411–1415.

[4] S. Kumar, E. Rosnes, and A. Graell i Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jun. 2017, pp. 1421–1425.

[5] A. Fazeli, A. Vardy, and E. Yaakobi, "PIR with low storage overhead: Coding instead of replication," May 2015, arXiv:1505.06241v1 [cs.IT]. [Online]. Available: http://arxiv.org/abs/1505.06241

[6] M. Vajha, V. Ramkumar, and P. V. Kumar, "Binary, shortened projective Reed Muller codes for coded private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jun. 2017, pp. 2648–2652.

[7] V. Skachek, "Batch and PIR codes and their connections to locally repairable codes," Jun. 2017, arXiv:1611.09914v3 [cs.IT]. [Online]. Available: https://arxiv.org/abs/1611.09914

[8] H.-Y. Lin and E. Rosnes, "Lengthening and extending binary private information retrieval codes," Jan. 2018, arXiv:1707.03495v3 [cs.IT]. [Online]. Available: https://arxiv.org/abs/1707.03495

[9] S. Rao and A. Vardy, "Lower bound on the redundancy of PIR codes," Feb. 2017, arXiv:1605.01869v2 [cs.IT]. [Online]. Available: https://arxiv.org/abs/1605.01869

[10] Z. Wang, O. Shaked, Y. Cassuto, and J. Bruck, "Codes for network switches," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013, pp. 1057–1061.

[11] V. I. Levenstein, "A class of systematic codes," *Sov. Math.-Dokl.*, vol. 1, no. 1, pp. 368–371, 1960.

[12] M. Grassl, "Bounds on the minimum distance of linear codes and quantum codes," accessed on 2017-03-31. [Online]. Available: http://www.codetables.de

[13] S. Lin and D. J. Costello, Jr., *Error Control Coding*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2004.

Construction $C^\star$: an inter-level coded version of Construction C

Maiara F. Bollauf [*], Ram Zamir [†] and Sueli I. R. Costa [‡]

[*] [‡] Institute of Mathematics, Statistic and Computer Science
University of Campinas, Sao Paulo, Brazil
Email: maiarabollauf@ime.unicamp.br, sueli@ime.unicamp.br

[†] Deptartment Electrical Engineering-Systems
Tel Aviv University, Tel Aviv, Israel
Email: zamir@eng.tau.ac.il

***Abstract*—Besides all the attention given to lattice constructions, it is common to find some very interesting nonlattice constellations, as Construction C, for example, which also has relevant applications in communication problems (multi-level coding, multi-stage decoding, good quantization efficieny). In this work we present a constellation which is a subset of Construction C, based on inter-level coding, which we call Construction $C^\star$. This construction may have better immunity to noise and it also provides a simple way of describing the Leech lattice Λ_{24}. A condition under which Construction $C^\star$ is a lattice constellation is given.**

***Index terms*—Lattice construction, Bit-interleaved coded modulation (BICM), Construction $C^\star$, Construction by Code-Formula, Leech lattice.**

I. Introduction

Communication problems involve, in general, transmitting digital information over a channel with minimum losses. One way to approach it is by using coded modulation [8], where not only coding, but also mapping the code bits to constellation symbols is significant. In the latest years, a prevalent coded modulation scheme is the bit-interleaved coded modulation (BICM), which is the motivation to our study.

BICM, first introduced by Zehavi [11], [6], asks mainly to have: an $nL-$dimensional binary code $\mathcal{C}$, an interleaver (permutation) α and a one-to-one binary labeling map $\mu : \{0,1\}^L \to \mathcal{X}$, where $\mathcal{X}$ is a signal set $\mathcal{X} = \{0, 1, \ldots, 2^L - 1\}$ in order to construct a constellation Γ_{BICM} in $\mathcal{X}^n \subseteq \mathbb{R}^n$. The code and interleaveled bit sequence c is partitioned into L subsequences c_i of length n :

$$c = (c_1, \ldots, c_L), \quad \text{with} \quad c_i = (c_{i1}, c_{i2}, \ldots, c_{in}). \tag{1}$$

The bits $c_{1j}, \ldots, c_{Lj}$ are mapped at a time index j to a symbol x_i chosen from the 2^L-ary signal constellation $\mathcal{X}$ according to the binary labeling map μ. Hence, for a $nL-$binary code $\mathcal{C}$ to encode all bits, then we have the scheme below:

$$\boxed{\text{codeword } (c)} \rightarrow \boxed{\text{interleaver } \alpha} \rightarrow$$
$$\boxed{\text{partitioning into L subsequences of length n}} \rightarrow \boxed{\text{mapping } \mu} \rightarrow$$
$$\boxed{x_j = \mu(c_{1j}, \ldots, c_{Lj}),\ j = 1, \ldots, n}$$

Under the natural labeling $\mu(c_1, c_2, , \ldots, c_L) = c_1 + 2c_2 + \cdots + 2^{L-1}c_L$ and assuming identity interleaver $\alpha(\mathcal{C}) = \mathcal{C}$, it is possible to define an extended BICM constellation in a way very similar to the well known multilevel Construction C, that we call Construction $C^\star$.

The constellation produced via Construction $C^\star$ is always a subset of the constellation produced via Construction C for the same projection codes (as defined below) and it also does not usually produce a lattice. The objective of our paper is to explore this new construction, aiming to find a condition that makes it a lattice and also to describe the Leech lattice Λ_{24} with Construction $C^\star$.

The paper is organized as follows: Section II shows some preliminary definitions; in Section III we introduce Construction $C^\star$, illustrate it with examples and also show how to describe the Leech lattice using this construction; in Section IV we exhibit a condition for $\Gamma_{C^\star}$ to be a lattice and Section V is devoted to conclusions.

II. Mathematical background

In this section, we will introduce the basic concepts, notation and results to be used in the sequel. We will denote by $+$ the real addition and by $\oplus$ the sum in $\mathbb{F}_2$, i.e., $x \oplus y = (x + y) mod\ 2$.

Definition 1. *(Lattice) A lattice $\Lambda \subset \mathbb{R}^N$ is a set of integer linear combinations of independent vectors $v_1, v_2, \ldots, v_n \in \mathbb{R}^N$, with $n \leq N$.*

It is possible to derive lattice constellations from linear codes using the known Constructions A and D [7].

Definition 2. *(Construction A) Let $\mathcal{C}$ be a linear $(n, k, d)-$binary code. We define the binary Construction A as*

$$\Lambda_A = \mathcal{C} + 2\mathbb{Z}^n. \tag{2}$$

Definition 3. *(Construction D) Let $\mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_L \subseteq \mathbb{F}_2^n$ be a family of nested linear binary codes. Let $k_i = \dim(\mathcal{C}_i)$ and* *let $b_1, b_2, \ldots, b_n$ be a basis of $\mathbb{F}_2^n$ such that $b_1, \ldots, b_{k_i}$ span $\mathcal{C}_i$. The lattice Λ_D consists of all vectors of the form*

$$\sum_{i=1}^{L} 2^{i-1} \sum_{j=1}^{k_i} \alpha_{ij} b_j + 2^L z \tag{3}$$

where $\alpha_{ij} \in \{0,1\}$ and $z \in \mathbb{Z}^n$.

Another remarkable and well studied multi-level construction, that in general does not produce a lattice constellation, even when the underlying codes are linear, is Construction C, defined below using the terminology in [9] (more details and applications also in [1] [3]).

Definition 4. *(Construction C) Consider L binary codes $\mathcal{C}_1, \ldots, \mathcal{C}_L \subseteq \mathbb{F}_2^n$, not necessarily nested or linear. Then we define an infinite constellation Γ_C in $\mathbb{R}^n$ that is called Construction C as:*

$$\Gamma_C := \mathcal{C}_1 + 2\mathcal{C}_2 + \cdots + 2^{L-1}\mathcal{C}_L + 2^L\mathbb{Z}^n, \tag{4}$$

or equivalently

$$\begin{aligned} \Gamma_C \quad := \quad & \{c_1 + 2c_2 + \cdots + 2^{L-1}c_L + 2^L z : c_i \in \mathcal{C}_i, \\ & i = 1, \ldots, L,\ z \in \mathbb{Z}^n\}. \end{aligned} \tag{5}$$

Note that if $L = 1$ and we consider a single level with a linear code, then both Constructions C and D reduce to lattice Construction A. There exists also a relation between Constructions C and D that will be presented in what follows.

Definition 5. *(Schur product) For $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n) \in \mathbb{F}_2^n$, we define $x * y = (x_1y_1, \ldots, x_ny_n)$.*

It is easy to verify, using the Schur product that for $x, y \in \mathbb{F}_2^n$

$$x + y = x \oplus y + 2(x * y). \tag{6}$$

Denote by Λ_C the smallest lattice that contains Γ_C. Kositwattanarerk and Oggier [10] give a condition that if satisfied guarantees that Construction C will provide a lattice which coincides with Construction D.

Theorem 1. *[10] (Lattice condition for Constructions C and D) Given a family of nested linear binary codes $\mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_L \subseteq \mathbb{F}_2^n$, then the following statements are equivalent:*

1. *Γ_C is a lattice.*
2. *$\Gamma_C = \Lambda_C$.*
3. *$\mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_L \subseteq \mathbb{F}_2^n$ is closed under Schur product, i.e., given two elements $c_i, \tilde{c}_i \in \mathcal{C}_i, c_i * \tilde{c}_i \in \mathcal{C}_{i+1}$, for all $i = 1, \ldots, L-1$.*
4. *$\Gamma_C = \Lambda_D$,*

III. Construction $C^\star$ over binary codes

This section is devoted to the introduction of a new method of constructing constellations from binary codes: Construction $C^\star$.

Definition 6. *(Construction $C^\star$) Let $\mathcal{C}$ be an $nL-$dimensional code in $\mathbb{F}_2^{nL}$. Then Construction $C^\star \in \mathbb{R}^n$ is defined as*

$$\begin{aligned} \Gamma_{C^\star} \quad := \quad & \{c_1 + 2c_2 + \cdots + 2^{L-1}c_L + 2^L z : (c_1, c_2, \ldots, c_L) \in \mathcal{C}, \\ & c_i \in \mathbb{F}_2^n, i = 1, \ldots, L, z \in \mathbb{Z}^n\}. \end{aligned} \tag{7}$$

Definition 7. *(Projection codes) Let $(c_1, c_2, ..., c_L)$ be a partition of a codeword $c = (b_1, ..., b_{nL}) \in \mathcal{C}$ into length$-n$ subvectors $c_i = (b_{(i-1)n+1}, ..., b_{in})$, $i = 1, \ldots, L$. Then, a projection code $\mathcal{C}_i$ consists of all vectors c_i that appear as we scan through all possible codewords $c \in \mathcal{C}$. Note that if $\mathcal{C}$ is linear, every projection code $\mathcal{C}_i, i = 1, \ldots, L$ is also linear.*

Remark 1. *If $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_L$ then Construction $C^\star$ coincides with Construction C, because the projection codes are independent. However, in general, the projection codes are dependent, i.e., not all combinations compose a codeword in the main code $\mathcal{C}$ so we get a subset of Construction C., i.e., $\Gamma_{C^\star} \subseteq \Gamma_C$.*

Definition 8. *(Associated Construction C) Given a Construction $C^\star$ defined by a linear binary code $\mathcal{C} \subseteq \mathbb{F}_2^{nL}$, we call the associated Construction C the constellation defined as*

$$\Gamma_C = \mathcal{C}_1 + 2\mathcal{C}_2 + \cdots + 2^{L-1}\mathcal{C}_L + 2\mathbb{Z}^n, \tag{8}$$

such that $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_L \in \mathbb{F}_2^n$ are the projection codes of $\mathcal{C}$ as in Definition 7.

One can observe that the immediate advantage of working with Construction $C^\star$ instead of Construction C lies in the fact that a code of block length nL typically has a larger minimum Hamming distance and may present a better immunity to noise than a code of block length n.

Example 1. *Consider a linear binary code $\mathcal{C}$ with length $nL = 4$, ($L = n = 2$), where $\mathcal{C} = \{(0,0,0,0), (1,0,0,1), (1,0,1,0), (0,0,1,1)\} \subseteq \mathbb{F}_2^4$. Thus, an element $w \in \Gamma_{C^\star}$ can be written as*

$$w = c_1 + 2c_2 + 4z \quad \in \Gamma_{C^\star}, \tag{9}$$

such that $(c_1, c_2) \in \mathcal{C}$ and $z \in \mathbb{Z}^2$. Geometrically, the resulting constellation is given by the blue points represented in Figure 1. Note that $\Gamma_{C^\star}$ is not a lattice because, for example, $(1,2), (3,0) \in \Gamma_{C^\star}$, but $(1,2) + (3,0) = (4,2) \notin \Gamma_{C^\star}$. However, if we consider the associated Construction C with codes $\mathcal{C}_1 = \{(0,0), (1,0)\}$ and $\mathcal{C}_2 = \{(0,0), (1,1), (0,1), (1,0)\}$, we have a lattice (Figure 1), because $\mathcal{C}_1$ and $\mathcal{C}_2$ satisfy the condition given by Theorem 1.

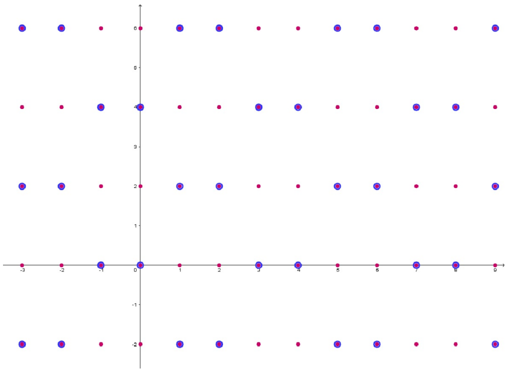


Fig. 1. (Nonlattice) Construction $C^\star$ constellation in blue and its associated (lattice) Construction C constellation in pink

The next example presents a case where both Constructions $C^\star$ and C are lattices, but they are not equal.

Example 2. *Let a linear binary code $\mathcal{C} = \{(0,0,0,0), (0,0,1,0), (1,0,0,1), (1,0,1,1)\} \subseteq \mathbb{F}_2^4$ ($nL = 4$, $L = n = 2$), so the projection codes are $\mathcal{C}_1 = \{(0,0), (1,0)\}$ and $\mathcal{C}_2 = \{(0,0), (1,0), (0,1), (1,1)\}$. An element $w \in \Gamma_{C^\star}$ can be described as*

$$w = \begin{cases} (0,0) + 4z, & \textit{if } c_1 = (0,0) \textit{ and } c_2 = (0,0) \\ (1,2) + 4z, & \textit{if } c_1 = (1,0) \textit{ and } c_2 = (0,1) \\ (2,0) + 4z, & \textit{if } c_1 = (0,0) \textit{ and } c_2 = (1,0) \\ (3,2) + 4z, & \textit{if } c_1 = (1,0) \textit{ and } c_2 = (1,1), \end{cases} \tag{10}$$

$z \in \mathbb{Z}^2$. This construction is represented by black points in Figure 2. Note that $\Gamma_{C^\star}$ is a lattice and $\mathcal{C} \neq \mathcal{C}_1 \times \mathcal{C}_2$, what implies that $\Gamma_{C^\star} \subsetneq \Gamma_C$. However, the associated Construction C is also a lattice (Figure 2).

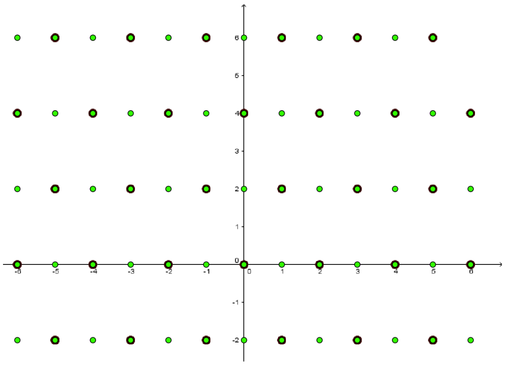


Fig. 2. (Lattice) Construction $C^\star$ constellation in black and its associated (lattice) Construction C constellation in green

To appreciate the advantage of $\Gamma_{C^\star}$ over the associated Γ_C, one can notice that the packing densities are, respectively $\Delta_{\Gamma_{C^\star}} = \frac{\Pi}{4} \approx 0.7853$ and $\Delta_{\Gamma_C} = \frac{\Pi}{8} \approx 0.3926$. Therefore, in this example, $\Gamma_{C^\star}$ has a higher packing density than Γ_C.

We can also describe the densest lattice in dimension 24, the Leech lattice Λ_{24}, in terms of Construction $C^\star$ constellation with $L = 3$ levels.

Example 3. *Based on the construction given by Conway and Sloane [7] (pp. 131-132) and Amrani et al [2], we start by considering three special linear binary codes*

- *$\mathcal{C}_1 = \{(0,\dots,0), (1,\dots,1)\} \subseteq \mathbb{F}_2^{24}$;*
- *$\mathcal{C}_2$ as a Golay code $\mathcal{C}_{24} \subset \mathbb{F}_2^{24}$ achieved by adding a parity bit to the original $[23, 12, 7]-$binary Golay code $\mathcal{C}_{23}$, which consists in a quadratic residue code of length 23;*
- *$\mathcal{C}_3 = \tilde{\mathcal{C}}_3 \cup \overline{\mathcal{C}}_3 = \mathbb{F}_2^{24}$, where $\tilde{\mathcal{C}}_3 = \{(x_1,\dots,x_{24}) \in \mathbb{F}_2^{24} : \sum_{i=1}^{24} x_1 \equiv 0 \mod 2\}$ and $\overline{\mathcal{C}}_3 = \{(y_1,\dots,y_{24}) \in \mathbb{F}_2^{24} : \sum_{i=1}^{24} y_1 \equiv 1 \mod 2\}$.*

Observe that $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ are linear codes. Consider a code $\mathcal{C} \subseteq \mathbb{F}_2^{72}$ whose codewords are described in one of two possible ways:

$$\begin{aligned} \mathcal{C} = \{&(0,\dots,0, \underbrace{a_1,\dots,a_{24}}_{\in \mathcal{C}_{24}}, \underbrace{x_1,\dots,x_{24}}_{\in \tilde{\mathcal{C}}_3}), \\ &(1,\dots,1, \underbrace{a_1,\dots,a_{24}}_{\in \mathcal{C}_{24}}, \underbrace{y_1,\dots,y_{24}}_{\in \overline{\mathcal{C}}_3})\}. \end{aligned} \tag{11}$$

Thus, we can define the Leech lattice Λ_{24} as a $3-$level Construction $C^\star$ given by

$$\Lambda_{24} = \Gamma_{C^\star} = \{c_1 + 2c_2 + 4c_3 + 8z : (c_1, c_2, c_3) \in \mathcal{C}, z \in \mathbb{Z}^{24}\}. \tag{12}$$

Observe that in this case $\Gamma_{C^\star} \neq \Gamma_C$.

In this case, the associated Construction C has packing density $\Delta_{\Gamma_C} \approx 0.00012 < 0.001929 \approx \Delta_{\Gamma_{C^\star}}$, which is the packing density of Λ_{24}, the best known packing density in dimension 24 [7].

IV. Conditions for latticeness of Construction $C^\star$

In general, it is possible to have a lattice $\Gamma_{C^\star}$, with $\Gamma_{C^\star} \subsetneq \Gamma_C$, as can be observed in Example 2. This fact motivated our search for a condition for a lattice Construction $C^\star$. In the upcoming discussion, we will exhibit some definitions and present a condition for $\Gamma_{C^\star}$ to be a lattice.

Definition 9. *(Antiprojection) The antiprojection (inverse image of a projection) $\mathcal{S}_i(c_1,\dots,c_{i-1},c_{i+1},\dots,c_L)$ consists of all vectors $c_i \in \mathcal{C}_i$ that appear as we scan through all possible codewords $c \in \mathcal{C}$, while keeping $c_1,\dots,c_{i-1},c_{i+1},\dots,c_L$ fixed:*

$$\begin{aligned} &\mathcal{S}_i(c_1,...,c_{i-1},c_{i+1},...,c_L) = \\ &\{c_i \in \mathcal{C}_i : (c_1,\dots,\underbrace{c_i}_{i\textit{-th posititon}},\dots,c_L) \in \mathcal{C}\}. \end{aligned} \tag{13}$$

The main contribution of this paper is the following:

Theorem 2. *(Lattice conditions for $\Gamma_{C^\star}$) Let $\mathcal{C} \subseteq \mathbb{F}_2^{nL}$ be a linear binary code with projection codes $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_L$ such that $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\dots,0) \subseteq \dots \subseteq \mathcal{C}_{L-1} \subseteq \mathcal{S}_L(0,\dots,0) \subseteq \mathcal{C}_L \subseteq \mathbb{F}_2^n$. Then the constellation given by $\Gamma_{C^\star}$ represents a lattice if and only if $\mathcal{S}_i(0,\dots,0)$ closes $\mathcal{C}_{i-1}$ under Schur product for all levels $i = 2,\dots,L$.*

The proof of Theorem 2 is given below, after a few motivational examples and related results.

While $\mathcal{S}_i(0,\dots,0) \subseteq \mathcal{C}_i$ by construction, note that the assumption that $\mathcal{C}_i \subseteq \mathcal{S}_{i+1}(0,\dots,0)$, for $i = 2,\dots,L$, is not always satisfied by a general Construction $C^\star$, sometimes even if this Construction $C^\star$ is a lattice; see Example 5.

Observe that when $\Gamma_{C^\star} = \Gamma_C$, i.e., when $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \dots \times \mathcal{C}_L$, we have that $\mathcal{S}_i(0,\dots,0) = \mathcal{C}_i, i = 1,\dots,L$ and our condition will coincide with the one presented in Theorem 1. Besides, if $\Gamma_{C^\star} \subsetneq \Gamma_C$ we also have that:

Corollary 1. *(Latticeness of associated Construction C) Let $\mathcal{C} \subseteq \mathbb{F}_2^{nL}$. If $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\dots,0) \subseteq \dots \subseteq \mathcal{C}_{L-1} \subseteq$*

$\mathcal{S}_L(0,\ldots,0) \subseteq \mathcal{C}_L$ and the constellation $\Gamma_{C^\star}$ is a lattice then also the associated Construction C is a lattice.

Proof. If $\Gamma_{C^\star}$ is a lattice, conditions presented in Theorem 2 holds. For associated Construction C, $\mathcal{S}_i(0,\ldots,0) = \mathcal{C}_i$, for all $i = 1,\ldots,L$. Thus, it follows that $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \cdots \subseteq \mathcal{C}_L$ is closed under Schur product and Γ_C is a lattice. □

Before presenting the proof of Theorem 2, we will see that the Leech lattice construction described in Example 3 satisfies its condition.

Example 4. *We want to examine whether the proposed codes $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ in Example 3 satisfy the conditions stated by Theorem 2.*

Observe that for these codes $\mathcal{S}_2(0,\ldots,0) = \mathcal{C}_2$ and $\mathcal{S}_3(0,\ldots,0) = \tilde{\mathcal{C}}_3 = \{(x_1,\ldots,x_{24}) \in \mathbb{F}_2^{24} : \sum_{i=1}^{24} x_1 \equiv 0 \mod 2\}$. Hence we need to verify that $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\ldots,0) \subseteq \mathcal{C}_2 \subseteq \mathcal{S}_3(0,\ldots,0) \subseteq \mathcal{C}_3$ and that $\mathcal{S}_i(0,\ldots,0)$ closes $\mathcal{C}_{i-1}$ under Schur product for $i = 2, 3$.

Indeed $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\ldots,0) = \mathcal{C}_2$, since $(0,\ldots,0) \in \mathcal{C}_2$ and if we consider the parity check matrix $H \in \mathbb{F}_2^{12\times 24}$ of the $[24,12,8]-$Golay code

$$H = \begin{pmatrix} B_{12\times 12} & | & I_{12\times 12} \end{pmatrix}, \tag{14}$$

where

$$B_{12\times 12} = \begin{pmatrix} 1&1&0&1&1&1&0&0&0&1&0&1\\ 1&0&1&1&1&0&0&0&1&0&1&1\\ 0&1&1&1&0&0&0&1&0&1&1&1\\ 1&1&1&0&0&0&1&0&1&1&0&1\\ 1&1&0&0&0&1&0&1&1&0&1&1\\ 1&0&0&0&1&0&1&1&0&1&1&1\\ 0&0&0&1&0&1&1&0&1&1&1&1\\ 0&0&1&0&1&1&0&1&1&1&0&1\\ 0&1&0&1&1&0&1&1&1&0&0&1\\ 1&0&1&1&0&1&1&1&0&0&0&1\\ 0&1&1&0&1&1&1&0&0&0&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&0 \end{pmatrix} \tag{15}$$

it is easy to check that $H \cdot (1,\ldots,1)^T = 0 \in \mathbb{F}_2^{12}$, so $(1,\ldots,1) \in \mathcal{C}_2$ which implies that $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\ldots,0)$.

Moreover, an element $c_2 \in \mathcal{C}_2$ can be written as $c_2 = G.h$, where $G = \begin{pmatrix} I_{12\times 12} \\ B_{12\times 12} \end{pmatrix}$ is the generator matrix of the Golay code and $h = (h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8, h_9, h_{10}, h_{11}, h_{12})^T \in \mathbb{F}_2^{12}$. Thus, when we sum all the coordinates of the resulting vector $c_2 = G.h$ we have $8h_1 + 8h_2 + 8h_3 + 8h_4 + 8h_5 + 8h_6 + 8h_7 + 8h_8 + 8h_9 + 8h_{10} + 8h_{11} + 12h_{12} \equiv 0 \mod 2 \Rightarrow c_2 \in \tilde{\mathcal{C}}_3 = \mathcal{S}_3(0,\ldots,0)$. Hence,

$$\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\ldots,0) \subseteq \mathcal{C}_2 \subseteq \mathcal{S}_3(0,\ldots,0) \subseteq \mathcal{C}_3. \tag{16}$$

We still need to prove that

- *$\mathcal{S}_2(0,\ldots,0)$ closes $\mathcal{C}_1$ under Schur product and this is clearly true because the Schur product of any elements in $\mathcal{C}_1$ belong to $\mathcal{S}_2(0,\ldots,0)$.*
- *$\mathcal{S}_3(0,\ldots,0)$ closes $\mathcal{C}_2$ under Schur product: if we consider $c_2 = G.h \in \mathcal{C}_2$ and $\tilde{c}_2 = G.\tilde{h} \in \mathcal{C}_2$, we have checked computationally that the sum of all coordinates of the Schur product $c_2 * \tilde{c}_2 \equiv 0 \mod 2 \Rightarrow c_2 * \tilde{c}_2 \in \mathcal{S}_3(0,\ldots,0) = \tilde{\mathcal{C}}_3$.*

We can have a lattice Construction $C^\star$ even when the nesting condition in Theorem 2 is not satisfied.

Example 5. *Consider the linear binary code $\mathcal{C} = \{(0,0,0,0,0,0), (1,0,1,1,0,1), (0,0,1,0,1,1), (1,0,0,1,1,0), (0,0,0,0,1,0), (0,0,1,0,0,1), (1,0,0,1,0,0), (1,0,1,1,1,1)\} \subseteq \mathbb{F}_2^6$ with $L = 3, n = 2$. Observe that $\mathcal{C}_1 \nsubseteq \mathcal{S}_2(0,0,0,0)$ and*

$$\begin{aligned}\Gamma_{C^\star} = \{&(0,0)+8z, (1,2)+8z, (2,4)+8z, (3,6)+8z,\\ &(4,0)+8z, (5,2)+8z, (6,4)+8z, (7,6)+8z\}\end{aligned} \tag{17}$$

with $z \in \mathbb{Z}^2$, is a lattice.

To prove Theorem 2 we need to introduce the following auxiliary result:

Lemma 1. *(Sum in $\Gamma_{C^\star}$) Let $\mathcal{C} \subseteq \mathbb{F}_2^{nL}$ be a linear binary code. If $x, y \in \Gamma_{C^\star}$ are such that*

$$x = c_1 + 2c_2 + \cdots + 2^{L-1}c_L + 2^L z \tag{18}$$

$$y = \tilde{c}_1 + 2\tilde{c}_2 + \cdots + 2^{L-1}\tilde{c}_L + 2^L \tilde{z}, \tag{19}$$

with $(c_1, c_2, \ldots, c_L), (\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_L) \in \mathcal{C}$ and $z, \tilde{z} \in \mathbb{Z}^n$, then

$$\begin{aligned} x + y = {} & c_1 \oplus \tilde{c}_1 + 2(s_1 \oplus (c_2 \oplus \tilde{c}_2)) + \cdots + \\ & +2^{L-1}(s_{L-1} \oplus (c_L \oplus \tilde{c}_L)) + 2^L(s_L + z + \tilde{z}), \end{aligned} \tag{20}$$

where

$$\begin{gathered} s_i = (c_i * \tilde{c}_i) \oplus r_i^1 \oplus r_i^2 \oplus \cdots \oplus r_i^{i-1} = (c_i * \tilde{c}_i) \bigoplus_{j=1}^{i-1} r_i^j, \\ r_i^1 = (c_i \oplus \tilde{c}_i) * (c_{i-1} * \tilde{c}_{i-1}), \quad r_i^j = r_i^{j-1} * r_{i-1}^{j-1}, \\ 2 \leq j \leq L-1, i = 1, \ldots, L. \end{gathered} \tag{21}$$

Proof. The proof is done by mathematical induction in the number of levels L and it will be provided in the full paper [4]. □

The mathematical intuition behind Theorem 2 lies in the fact that since $a + b = a \oplus b + 2(a * b)$ for $a, b \in \mathbb{F}_2^n$, when adding two points in Γ_C or $\Gamma_{C^\star}$, each level $i \geq 2$ has the form of $c_i \oplus \tilde{c}_i \oplus carry_{(i-1)}$, where $carry_{(i-1)}$ is the "carry" term from the addition in the lower level. Since the projection code $\mathcal{C}_i$ is linear, $c_i \oplus \tilde{c}_i$ is a codeword in the $i-$th level. Hence, closeness of $\Gamma_{C^\star}$ under addition amounts to the fact that $carry_{(i-1)}$ is also a codeword in $\mathcal{C}_i$, which is essentially the condition of the theorem. Formally,

Proof of Theorem 2. ($\Leftarrow$) For any $x, y \in \Gamma_{C^\star}$, written as in Equations (18) and (19), we have $x + y$ as given in Lemma 1 (Equations (20) and (21)) and we need to verify if $x+y \in \Gamma_{C^\star}$.

Clearly $x+y \in \mathcal{C}_1 + 2\mathcal{C}_2 + \cdots + 2^{L-1}\mathcal{C}_L + 2^L\mathbb{Z}^n$. It remains to demonstrate that $(c_1 \oplus \tilde{c}_1, s_1 \oplus c_2 \oplus \tilde{c}_2, \ldots, s_{L-1} \oplus c_L \oplus \tilde{c}_L) \in \mathcal{C}$.

Indeed, using the fact that the chains $\mathcal{C}_{i-1} \subseteq \mathcal{S}_i(0,\ldots,0)$ for all $i = 2,\ldots,L$ are closed under the Schur product, it is an element of $\mathcal{C}$ because it is a sum of elements in $\mathcal{C}$, i.e.,

$$\begin{aligned}&(c_1 \oplus \tilde{c}_1, s_1 \oplus c_2 \oplus \tilde{c}_2, \ldots, s_{L-1} \oplus c_L \oplus \tilde{c}_L) = \\ &\underbrace{(c_1 \oplus \tilde{c}_1, c_2 \oplus \tilde{c}_2, \ldots, c_L \oplus \tilde{c}_L)}_{\in \mathcal{C}} \oplus \underbrace{(0, s_1, \ldots, 0)}_{\in \mathcal{C}} \oplus \cdots \oplus \\ &\oplus \underbrace{(0, \ldots, 0, s_{L-1})}_{\in \mathcal{C}}.\end{aligned} \tag{22}$$

Observe that any $nL-$tuple $(0,\ldots,s_{i-1},\ldots,0)$ is in $\mathcal{C}$ because by hypothesis, the chain $\mathcal{S}_i(0,\ldots,0)$ closes $\mathcal{C}_{i-1}$ under Schur product, hence $S_i(0,\ldots,0)$ contains $(c_{i-1} * \tilde{c}_{i-1}), r_{i-1}^1, ..., r_{i-1}^{i-2}$ which is sufficient to guarantee that $s_{i-1} \in \mathcal{S}_i(0,\ldots,0)$ so $(0,\ldots,s_{i-1},\ldots,0) \in \mathcal{C}$, for all $i = 2,\ldots,L-1$.
($\Rightarrow$) For the converse, we know that $\Gamma_{C^\star}$ is a lattice, which implies that if $x, y \in \Gamma_{C^\star}$ then $x+y \in \Gamma_{C^\star}$. From the notation and result from Lemma 1, more specifically Equations (18), (19), (20) and (21), it means that

$$(c_1 \oplus \tilde{c}_1, s_1 \oplus (c_2 \oplus \tilde{c}_2), \ldots, s_{L-1} \oplus (c_L \oplus \tilde{c}_L)) \in \mathcal{C}. \tag{23}$$

We can write this $L-$tuple as

$$\begin{aligned}&\underbrace{(c_1 \oplus \tilde{c}_1, s_1 \oplus (c_2 \oplus \tilde{c}_2), \ldots, s_{L-1} \oplus (c_L \oplus \tilde{c}_L))}_{\in \mathcal{C}} = \\ &\underbrace{(c_1 \oplus \tilde{c}_1, c_2 \oplus \tilde{c}_2, \ldots, c_L \oplus \tilde{c}_L)}_{\in \mathcal{C},\text{ by linearity of } \mathcal{C}} \oplus (0, s_1, \ldots, s_{L-1}) \qquad (24)\\ \Rightarrow \quad &(0, s_1, \ldots, s_{L-1}) \in \mathcal{C}. \qquad (25)\end{aligned}$$

Notice that we have

$$\begin{aligned} s_1 &= c_1 * \tilde{c}_1 & (26)\\ s_2 &= ((c_1 * \tilde{c}_1) * (c_2 \oplus \tilde{c}_2)) \oplus (c_2 * \tilde{c}_2) & (27)\\ s_3 &= ((c_3 \oplus \tilde{c}_3) * (c_2 * \tilde{c}_2)) * (c_2 \oplus \tilde{c}_2 * (c_1 * \tilde{c}_1)) & \\ &\oplus ((c_3 \oplus \tilde{c}_3) * (c_2 * \tilde{c}_2)) \oplus (c_3 * \tilde{c}_3) & (28)\\ &\vdots & \end{aligned}$$

Due to the nesting $\mathcal{C}_1 \subseteq \mathcal{S}_2(0,\ldots,0) \subseteq \cdots \subseteq \mathcal{C}_{L-1} \subseteq \mathcal{S}_L(0,\ldots,0) \subseteq \mathcal{C}_L$, we can guarantee that there exist codewords whose particular Schur products $c_i * \tilde{c}_i = 0$, for $i = 1,\ldots,L-2$. Thus,

$$s_{L-1} = (c_{L-1} * \tilde{c}_{L-1}) \tag{29}$$

and from Equation (25), $(0, 0, \ldots, c_{L-1} * \tilde{c}_{L-1}) \in \mathcal{C}$, i.e., $S_L(0,\ldots,0)$ must close $\mathcal{C}_{L-1}$ under Schur product. Proceeding similarly, we demonstrate that $S_i(0,\ldots,0)$ must close $\mathcal{C}_{i-1}$, for all $i = 2,\ldots,L$ and it completes our proof.

□

V. Conclusion and future work

In this paper a new method of constructing constellations was introduced, denoted by Construction $C^\star$, which is subset of Construction C and is based on a modern coding scheme, the bit-interleaved coded modulation (BICM). It was proved when this construction is a lattice and how to describe the Leech lattice using this technique.

Our future work include examining on a comparative basis the advantages of Construction $C^\star$ compared to Construction C in terms of packing density. We also aim to change the natural labeling μ to the Gray map, the standard map used in BICM. Another direction to be completed is to find a more complete condition for the latticeness of Construction $C^\star$, that covers cases such as the one in Example 5.

Acknowledgment

CNPq (140797/2017-3, 312926/2013-8) and FAPESP (2013/25977-7) supported MFB and SIRC. Israel Science Foundation grant 676/15 supported RZ.

References

[1] E. Agrell and T. Eriksson, "Optimization of Lattice for Quantization". *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1814-1828, Sep. 1998.

[2] O. Amrani *et al*, "The Leech Lattice and the Golay Code: Bounded-Distance Decoding and Mu1tilevel Constructions". *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1030-1043, Jul. 1994.

[3] M. F. Bollauf and R. Zamir, "Uniformity properties of Construction C", in *2016 IEEE Inter. Symp. on Inform. Theory*, (Barcelona), 2016, pp. 1516-1520.

[4] M. F. Bollauf, R. Zamir and S.I.R. Costa, "Constructions C and $C^\star$: theoretical and practical approach", *in preparation*.

[5] A. Bonnecaze *et al*, "Quaternary Quadratic Residue Codes and Unimodular Lattices". *IEEE Trans. on Inform. Theory*, vol. 41, no. 2, pp. 366-377, Mar. 1995.

[6] G. Caire, G. Taricco and E. Biglieri, "Bit-interleaved coded modulation". *IEEE Trans. on Inform. Theory*, vol. 44, no. 3, pp. 927-946, May. 1998.

[7] J. H. Conway and N.J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. New York, USA: Springer, 1999.

[8] R. de Buda, "Fast FSK signals and their demodulation". *Can. Electron. Eng. Journal*, vol. 1, pp. 2834, Jan. 1976.

[9] G. D. Forney, "Coset codes-part I: introduction and geometrical classification". *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 1123-1151. Sep. 1988.

[10] W. Kositwattanarerk and F. Oggier, "Connections between Construction D and related constructions of lattices". *Designs, Codes and Cryptography*, v. 73, pp. 441-455, Nov. 2014.

[11] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel". *IEEE Trans. Commun.*, vol. 40, no. 3, pp. 873884, May 1992.

Semantically Secure Lattice Codes for Compound MIMO Channels

Antonio Campello, Cong Ling
Department of Electrical and Electronic Engineering
Imperial College London, U.K.
Email: {a.campello, c.ling}@imperial.ac.uk

Jean-Claude Belfiore
Mathematical and Algorithmic Sciences Lab
France Research Center, Huawei Technologies
belfiore@telecom-paristech.fr

Abstract—**We consider compound MIMO wiretap channels where minimal channel state information at the transmitter (CSIT) is assumed. Using the flatness factor for MIMO channels, we propose lattice codes universally achieving the secrecy capacity of compound MIMO wiretap channels up to a constant gap that is linear in the number of transmit antennas, independently on the number of eavesdropper antennas. The proposed approach improves upon existing works on secrecy coding for MIMO wiretap channels from an error probability perspective, and establishes information theoretical security (in fact semantic security). We also give an algebraic construction to reduce the code design complexity, as well as the decoding complexity of the legitimate receiver.**

I. Introduction

Due to the open nature of the wireless medium, wireless communications are inherently vulnerable to eavesdropping attacks. Information theoretic security offers additional protection for wireless data, since it only relies on the physical properties of wireless channels, thus representing a competitive/complementary approach to security compared to traditional cryptography.

In the information theory community, a commonly used secrecy notion is *strong secrecy*: the mutual information $\mathbb{I}(M; Z^N)$ between the confidential message M and the channel output Z^N should vanish when the code length $N \to \infty$. This assumption of uniformly distributed messages was dropped in [17], establishing *semantic security*: for *any* message distribution, the advantage obtained by an eavesdropper from its received signal vanishes for large block lengths. This notion is motivated by the fact that the plaintext can be fixed and arbitrary.

For the Gaussian wiretap channel, [15] introduced the *secrecy gain* of lattice codes while [11] proposed semantically secure lattice codes based on the lattice Gaussian distribution. To this aim, the *flatness factor* of a lattice was introduced in [11] as a fundamental criterion which implies that conditional outputs are indistinguishable for different input messages. Using a random coding argument, it was shown that there exist families of lattice codes which are *good for secrecy*, meaning that their flatness factor is vanishing, and achieve semantic security for rates up to $1/2$ nat from the secrecy capacity.

Compared to the Gaussian wiretap channel, the cases of fading and multi-input multi-output (MIMO) wiretap channels are more technically challenging. The fundamental limits of fading wireless channels with secrecy constraints have been investigated in [1], [3], [4], [9], [**?**] where the achievable rates, secrecy capacity, and the secrecy outage probability were given. Although CSIT is sometimes available for the legitimate channel, it is hardly possible that it would be available for the eavesdropping channel. Schaefer and Loyka [18] studied the secrecy capacity of the *compound* MIMO wiretap channel, where a transmitter has no knowledge of the realization of the eavesdropping channel, except that it belongs to a given set (the *compound set*). The compound model represents a well-accepted, reasonable approach to information theoretic security, which assumes minimal CSIT of the eavesdropping channel [10], [8], and can also model multicast scenarios.

In this paper, we propose universal codes for compound Gaussian MIMO wiretap channels. Previously, [14] has established strong secrecy over *ergodic* stationary MIMO wiretap channels for secrecy rates that are within a constant gap to the secrecy capacity. Besides different channel models (compound vs. ergodic channels) considered, we obtain a smaller (and different in nature) gap by employing a different construction as in [14].

For a compound channel formed by the set of all matrices with same white-input capacity, our lattice coding scheme universally achieves rates up to $(C_b - C_e - n_a)^+$, where C_b is the capacity of the legitimate channel, C_e is the capacity of the eavesdropper channel and n_a is the number of transmit antennas and $(x)^+ = \max\{x, 0\}$. We also show how to extend the analysis in order to accommodate number-of-antennas mismatch, i.e., security is valid *regardless* of the number of antennas at the eavesdropper. This is a very appealing property, since the number of receive antennas of an eavesdropper may be unknown to the transmitter. Notice that previous works [2], [14] required $n_e \geq n_a$.

II. Problem Statement

We consider the following wiretap model. A transmitter sends information through a MIMO channel to a legitimate receiver (Bob) and is eavesdropped by an illegitimate user (Eve). The channel equations for Bob and Eve read:

$$\underbrace{\mathbf{Y}_b}_{n_b \times T} = \underbrace{\mathbf{H}_b}_{n_b \times n_a} \underbrace{\mathbf{X}}_{n_a \times T} + \underbrace{\mathbf{W}_b}_{n_b \times T} \text{ and } \underbrace{\mathbf{Y}_e}_{n_e \times T} = \underbrace{\mathbf{H}_e}_{n_e \times n_a} \underbrace{\mathbf{X}}_{n_a \times T} + \underbrace{\mathbf{W}_e}_{n_e \times T}, \tag{1}$$

where n_a is the number of transmit antennas, n_b, n_e is the number of receive antennas for Bob and Eve, T is the coherence time, and W_b and W_e have circularly symmetric complex Gaussian iid entries with variance σ_b^2, σ_e^2 per complex dimension. We denote the signal-to-noise ratios by

$$\rho_b \triangleq \frac{P}{\sigma_b^2} \text{ and } \rho_e \triangleq \frac{P}{\sigma_e^2}.$$

We assume that the exact channel realizations $(\mathbf{H}_b, \mathbf{H}_e)$ are *unknown* to the transmitter but belong to a compound set $\mathcal{S} = \mathcal{S}_b \times \mathcal{S}_e \subset \mathbb{C}^{n_b \times n_a} \times \mathbb{C}^{n_e \times n_a}$. Suppose that $\mathcal{S}_b$ and $\mathcal{S}_e$ are the set of channels with same isotropic mutual information i.e.,

$$\begin{aligned} \mathcal{S}_b &= \left\{\mathbf{H}_b \in \mathbb{C}^{n_b \times n_a} : \left|\mathbf{I} + \rho_b \mathbf{H}_b^\dagger \mathbf{H}_b\right| = e^{C_b}\right\} \text{ and} \\ \mathcal{S}_e &= \left\{\mathbf{H}_e \in \mathbb{C}^{n_e \times n_a} : \left|\mathbf{I} + \rho_e \mathbf{H}_e^\dagger \mathbf{H}_e\right| = e^{C_e}\right\}, \end{aligned} \tag{2}$$

for fixed $C_b, C_e \geq 0$. In this case, it is known (e.g. [18]) that $C_s \geq (C_b - C_e)^+$. The worst case is achieved by taking a specific "isotropic" realization such that $\mathbf{H}_b^\dagger \mathbf{H}_b$ and $\mathbf{H}_e^\dagger \mathbf{H}_e$ are a multiple of the identity, from where we conclude that $C_s = C_b - C_e$. In what follows we construct lattice codes that approach the rate C_s with *semantic* security. As a corollary, the semantic security capacity and the strong secrecy capacity for the compound set $\mathcal{S}_b \times \mathcal{S}_e$ coincide.

A. Notions of Security

A secrecy code (or, more precisely, an (R, R', T) secrecy code) for the compound MIMO channel consists of: .

(i) A set of messages $\mathcal{M}_T = \left\{1, \ldots, e^{TR}\right\}$

(ii) An auxiliary source U taking values in $\mathcal{U}_T$ with entropy $R' = H(U)$.

(iii) An encoding function $f_T : \mathcal{M}_T \times \mathcal{U}_T \to \mathbb{C}^{n_a \times T}$ s.t.

$$\frac{1}{T}\text{tr}\left(\mathbb{E}\left[f_T(m,U)^\dagger f_T(m,U)\right]\right) \leq n_a P \tag{3}$$

(iv) A decoding function $g : \mathcal{S}_b \times \mathbb{R}^{n_b \times T} \to \mathcal{M}_T$.

A pair $(s_b, s_e) \in \mathcal{S}_b \times \mathcal{S}_e$ is referred to as a *channel state* (or *channel realization*). To ensure reliability for all channel states, we require a sequence of codes whose error probability for message M vanishes:

$$\mathbb{P}_{\text{err}|M}^{(T)} \triangleq \mathbb{P}(g(s_b, M) \neq M) \to 0, \forall s_b \in \mathcal{S}_b, \text{ as } T \to \infty. \tag{4}$$

Let p_M be a message distribution over $\mathcal{M}_T$. For channel coding, p_M is usually assumed to be uniform, however this assumption is not sufficient for modern security purposes. Let $\mathbf{Y}_e$ be the output of the channel to the eavesdropper, who is omniscient. In the limit of $T \to \infty$ the notion of *semantic security* coincides with the following [11],[17]:

$$\max_{m', m'' \in \mathcal{M}_T} \mathbb{V}(p_{\mathbf{Y}_e|m'}, p_{\mathbf{Y}_e|m''}) \to 0 \text{ for all } s_e \in \mathcal{S}_e, \tag{5}$$

where $\mathbb{V}$ stands for the l_1 variational distance between distributions. In other words the eavesdropper cannot distinguish the output of the channel for different messages. This notion also requires a sequence of codes to be *universally* secure for all channel states. We say that a sequence of codes of rate approaching R is semantically secure for compound MIMO if, for all $(s_b, s_e) \in \mathcal{S}$ it satisfies the reliability condition (4) and (5). In what follows we proceed to construct universally secure codes for the MIMO wiretap channel using lattice coset codes.

III. Correlated Discrete Gaussian Distributions

We exhibit in this subsection the important results and concepts for the definition and analysis of our lattice coding scheme.

A. Preliminary lattice definitions

A (complex) lattice Λ with generator matrix $\mathbf{B}_c$ is a discrete additive subgroup of $\mathbb{C}^{n_a}$ given by

$$\Lambda = \mathcal{L}(\mathbf{B}_c) = \left\{\mathbf{B}_c \mathbf{x} : \mathbf{x} \in \mathbb{Z}^{2n_a}\right\}. \tag{6}$$

A complex lattice has an equivalent real lattice generated by the matrix $\mathbf{B}_r$ obtained by stacking real and imaginary parts of matrix $\mathbf{B}_c$.

A *fundamental region* $\mathcal{R}(\Lambda)$ for Λ, is any interior-disjoint region that tiles $\mathbb{C}^{n_a}$ through translates by vectors of Λ. For any $\mathbf{y}, \mathbf{x} \in \mathbb{C}^{n_a}$ we say that $\mathbf{y} = \mathbf{x} \pmod \Lambda$ iff $\mathbf{y} - \mathbf{x} \in \Lambda$. By convention, we fix a fundamental region and denote by $\mathbf{y} \pmod \Lambda$ the unique representative $\mathbf{x} \in \mathcal{R}(\Lambda)$ such that $\mathbf{y} = \mathbf{x} \pmod \Lambda$. The volume of Λ is defined as the volume of a fundamental region for the equivalent real lattice, given by $V(\Lambda) = |\mathbf{B}_r|$. Notice that if $\mathbf{B}_c$ is full rank, then $V(\Lambda) > 0$.

B. The Flatness Factor

The discrete Gaussian distribution and the flatness factor will be used to measure bound the information leakage to an eavesdropper.

The pdf of a correlated Gaussian distribution with covariance matrix $\mathbf{\Sigma}$ is

$$f_{\sqrt{\mathbf{\Sigma}},\mathbf{c}}(\mathbf{x}) = \frac{1}{\pi^{n_a}|\mathbf{\Sigma}|} \exp\left\{-(\mathbf{x} - \mathbf{c})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{c})\right\}.$$

We write $f_{\sqrt{\mathbf{\Sigma}},\Lambda}(\mathbf{x})$ for the sum of $f_{\sigma,\mathbf{c}}(\mathbf{x})$ over $\mathbf{c} \in \Lambda$. Wen $\mathbf{c} = \mathbf{0}$ we omit the index. The *flatness factor* of a lattice quantifies the distance between $f_{\sigma,\Lambda}(\mathbf{x})$ and an uniform distribution over $\mathcal{R}(\Lambda)$.

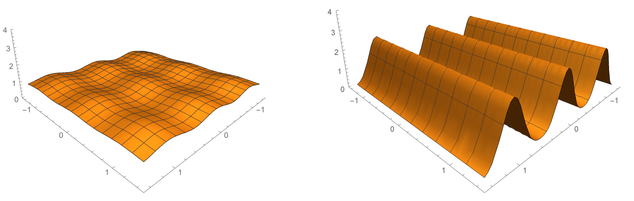
Definition 1 (Flatness factor for correlated Gaussian distributions)**.**

$$\epsilon_\Lambda(\sqrt{\mathbf{\Sigma}}) \triangleq \max_{\mathbf{x} \in \mathcal{R}(\Lambda)} |V(\Lambda) f_{\sqrt{\mathbf{\Sigma}},\Lambda}(\mathbf{x}) - 1|$$

where $\mathcal{R}(\Lambda)$ is a fundamental region of Λ.

When $\mathbf{c} = 0$ we ignore the index and write $f_{\sqrt{\mathbf{\Sigma}},\mathbf{0}}(\mathbf{x}) = f_{\sqrt{\mathbf{\Sigma}}}(\mathbf{x})$. For a co-variance matrix $\mathbf{\Sigma}$ we define the generalized-volume-to-noise ratio as $\gamma_\Lambda(\sqrt{\mathbf{\Sigma}}) = V(\Lambda)^{1/n_a}/|\mathbf{\Sigma}|^{1/n_a}$.

In our applications, the matrix $\mathbf{\Sigma}$ will be determined by the channel realization (1), and we will deal with lattices of dimension $n_a T$, where T is the coherence time. Figure 1 shows the effect of fading on the lattice Gaussian function. A function which is flat over the Gaussian channel (corresponding to $\mathbf{\Sigma} = \mathbf{I}$) (a) need not be flat for a channel in deep fading (corresponding to a ill-conditioned $\mathbf{\Sigma}$) (b), in which case an eavesdropper could clearly distinguish one dimension of the signal.



(a) $\mathbf{\Sigma} = 0.25\mathbf{I}$ (b) $\mathbf{\Sigma} = 0.25\mathrm{diag}(6, 1/6)$

Fig. 1: Illustration of the Gaussian periodic function for the lattice $\mathbb{Z}^2$ and different co-variance matrices with same determinant.

C. The discrete Gaussian Distribution

In order to define our coding scheme, we need a last element, which is the distribution of the sent signals. To this purpose, we define the *discrete Gaussian distribution* $\mathcal{D}_{\Lambda+\mathbf{c},\sqrt{\mathbf{\Sigma}}}$ as the distribution assuming values on $\Lambda + \mathbf{c}$, such that the probability of each point $\lambda + \mathbf{c}$ is given by

$$\mathcal{D}_{\Lambda+\mathbf{c},\sqrt{\mathbf{\Sigma}}}(\lambda + \mathbf{c}) = \frac{f_{\sqrt{\mathbf{\Sigma}}}(\lambda + \mathbf{c})}{f_{\sqrt{\mathbf{\Sigma}},\Lambda}(\mathbf{c})}.$$

Its relation to the continuous Gaussian distribution can be done via the smoothing parameter or the flatness factor. For instance, a vanishing flatness factor guarantees that the power per-dimension of $\mathcal{D}_{\Lambda+\mathbf{c},\sigma\mathbf{I}}$ is approximately σ^2.

The next proposition ([14, Appendix I-A]) says that the sum of a continuous Gaussian and a discrete Gaussian is approximately a continuous Gaussian, provided that the flatness factor is small.

Lemma 1. *Given $\mathbf{x}_1$ sampled from discrete Gaussian distribution $D_{\Lambda+\mathbf{c},\sqrt{\mathbf{\Sigma}_1}}$ and $\mathbf{x}_2$ sampled from continuous Gaussian distribution $f_{\sqrt{\mathbf{\Sigma}_2}}$. Let $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$ and let $\mathbf{\Sigma}_3^{-1} = \mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1}$. If $\sqrt{\mathbf{\Sigma}_3} \succeq \eta_\varepsilon(\Lambda)$ for $\varepsilon \leq \frac{1}{2}$, then the distribution g of $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ is close to $f_{\sqrt{\mathbf{\Sigma}_0}}$:*

$$g(\mathbf{x}) \in f_{\sqrt{\mathbf{\Sigma}_0}}(\mathbf{x})\left[1 - 4\varepsilon, 1 + 4\varepsilon\right].$$

IV. Coding Scheme and Analysis

A. Overview

Given a pair of nested lattices $\Lambda_e \subset \Lambda_b \in \mathbb{C}^{n_a T}$ such that

$$\frac{1}{T}\log|\Lambda_b/\Lambda_e| = R,$$

the transmitter maps m to a representative of Λ_b/Λ_e via a one-to-one map ϕ, such that $\phi(m) = \lambda_m$, and then samples the signal $\mathbf{x} \sim \mathcal{D}_{\Lambda_e+\lambda_m,\sigma_s}$, broadcasting it to the channels. A block diagram for the transmission until the front-end receivers Bob and Eve is depicted in Figure 2a.

In order to find pairs of sequences of nested lattices Λ_b and Λ_e we resort to constructions of lattices from error-correcting codes.

1) Construction A: A general "flexible" construction can be defined via "generalized reductions". For let $\phi_p : \Lambda_{\text{base}} \to \mathbb{F}_p^T$ be a surjective homomorphism from a base lattice Λ_{base} of complex dimension N to the vector space $\mathbb{F}_p^T$ (also referred to as a *reduction*). Define the lattice $\Lambda(\mathcal{C})$ as the pre-image of a linear code $\mathcal{C}$,

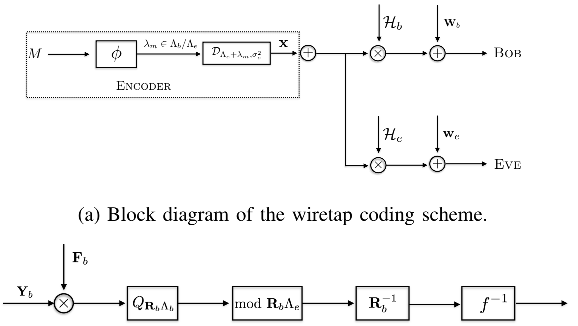
$$\Lambda(\mathcal{C}) = \phi^{-1}(\mathcal{C}).$$

If $\mathcal{C}$ has length T and dimension k, the volume of $\Lambda(\mathcal{C})$ equals to $p^{T-k}V(\Lambda_{\text{base}})$. For instance if $T = 2N$, $\Lambda_{\text{base}} = \mathbb{Z}[i]^T$ mapping ϕ is the reduction modulo p:

$$\begin{aligned}&\phi(a_1 + b_1 i, a_2 + b_2 i, \ldots, a_N + b_N i) = \\&(a_1 \pmod p), b_1 \pmod p), a_2 \pmod p), b_2 \\&\pmod p), \cdots, a_N \pmod p), b_N \pmod p)),\end{aligned} \tag{7}$$

we recover an analogue to Loeliger's (mod-p) Construction A [12]. In this case we obtain a nested lattice beween $\mathbb{Z}[i]^T$ and $p\mathbb{Z}[i]^T$. More refined "direct" constructions can be obtained by using number theory and prime ideals of $\mathbb{Z}[i]$. Notice that, for this construction, if $\mathcal{C}_1 \subset \mathcal{C}_2$, we obtain two nested lattices $\Lambda(\mathcal{C}_1) \subset \Lambda(\mathcal{C}_2)$, from where we can perform coset codes. We choose the"reliability lattice" $\Lambda_b = \Lambda(\mathcal{C}_2)$, the "secrecy lattice" $\Lambda_e = \Lambda(\mathcal{C}_1)$. The parameters of the code are chosen according to the achievable rates, and will be describe more carefully later on.

2) Main result: The lattice Λ_e controls the eavesdropper confusion, and has to be chosen in such a way that the flatness factor vanishes universally for any eavesdropper realization (universally good for secrecy), so that it does not leak any information. Our main result is the following theorem, stating the existence of schemes with vanishing probability of error and information leakage for universally any pair realizations in the compound set $\mathcal{S}_b \times \mathcal{S}_e$.



(a) Block diagram of the wiretap coding scheme.

(b) Block diagram of Bob's receiver, where $\mathbf{F}_b$ is the MMSE-GDFE matrix and $\mathbf{R}_b^{-1}$ is the inverse linear operator that maps cosets of $\mathbf{R}_b\Lambda_b/\mathbf{R}_b\Lambda$ into cosets of Λ_b/Λ_e.

Fig. 2: Encoding and decoding over the compound wiretap channel.

Theorem 1. *There exists a sequence of pairs of nested lattices $(\Lambda_b^T, \Lambda_e^T)_{T=1}^{\infty}$, $\Lambda_b^T \subset \Lambda_e^T \subset \mathbb{C}^{n_a T}$, such that, as $T \to \infty$, the lattice coding scheme universally achieves any secrecy rates*

$$R < (C_b - C_e - n_a)^+,$$

where m is the number of transmit antennas.

B. The Legitimate Channel: Reliability

It was shown in [6] that if $\mathbf{X} \sim \mathcal{D}_{\Lambda_b,\sigma_s}$, then the maximum-a-posteriori (MAP) decoder for the signal $\mathbf{Y}_b$ is equivalent to lattice decoding of $\mathbf{F}_b\mathbf{Y}_b$, where $\mathbf{F}_b$ is the MMSE-GDFE matrix to be defined in the sequel. We cannot claim directly that $\mathbf{X} \sim \mathcal{D}_{\Lambda_b,\sigma_s}$, since the message distribution in $\mathcal{M}$ need not be uniform. Nevertheless, we show that reliability is still possible for all individual messages

The full decoding process is depicted in Figure 2b. Bob first applies a filtering matrix $\mathbf{F}_b$ so that

$$\tilde{\mathbf{Y}}_b = \mathbf{F}_b\tilde{\mathbf{Y}}_b = \mathbf{R}_b\mathbf{X} + \mathbf{W}_{b,\text{eff}},$$

where $\mathbf{R}_b^\dagger\mathbf{R}_b = \mathbf{H}_b^\dagger\mathbf{H}_b + \rho_b^{-1}\mathbf{I}$ and $\mathbf{F}_b^\dagger\mathbf{R}_b = \rho_b^{-1}\mathbf{H}_b$. The effective noise is then

$$\mathbf{W}_{b,\text{eff}} = (\mathbf{F}_b\mathbf{H}_b - \mathbf{R}_b)\mathbf{X} + \mathbf{F}_b\mathbf{W}_b.$$

The next step is to decode $\tilde{\mathbf{Y}}_b$ in $\mathbf{R}_b\Lambda_b$, in order to obtain $Q_{\mathbf{R}_b\Lambda_b}(\tilde{\mathbf{Y}}_b)$, which is then remapped into the element of the coset $\mathbf{R}_b\Lambda_b/\mathbf{R}_b\Lambda_e$ through the operation mod $\mathbf{R}_b\Lambda_e$. We can then invert the linear transformation associated to $\mathbf{R}_b$ (notice that $\mathbf{R}_b$ is full rank) in order to obtain the coset in Λ_b/Λ_e and re-map it to the message space $\mathcal{M}$ through f^{-1}. Using the fact that the effective noise is sub-Gaussian [14] with parameter σ_b^2. Therefore, as long as $\varepsilon' \approx 0$ the probability of error tends to zero if we choose Λ_b to be AWGN good.

C. The Eavesdropper Channel: Semantic Security

For a *fixed* realization $\mathbf{H}_e$, the key element for bounding the information leakage is the following lemma [11, Lem 2]:

Lemma 2. *Suppose that there exists a random variable with density q taking values in $\mathbb{C}^{n_e\times T}$ such that $\mathbb{V}(p_{\mathbf{Y}_e|m}, q_{\mathbf{Y}_e}) \leq \varepsilon_T$ for all $m \in \mathcal{M}_T$. Then, for all message distributions*

$$\mathbb{I}(M;\mathbf{Y}_e) \leq 2T\varepsilon_T R - 2\varepsilon_T \log 2\varepsilon_T. \tag{8}$$

We show that if the distribution is sufficiently flat, then $\mathbf{Y}_e|m$ is statistically close to a multivariate Gaussian. Let us assume for now that $\mathbf{H}_e$ is an invertible square matrix. We next show how to reduce the other cases to this one. In this case, given a message M, we have $\mathcal{H}_e\mathbf{x} \sim \mathcal{D}_{\mathcal{H}_e(\Lambda_e+\lambda_m),\sqrt{(\mathcal{H}_e\mathcal{H}_e^\dagger)\sigma_s^2}}$.

According to Lemma 1, the distribution of $\mathcal{H}_e\mathbf{x} + \mathbf{w}_e$ is within variational distance $4\varepsilon_T$ from the normal distribution $\mathcal{N}(0,\sqrt{\mathbf{\Sigma}_0})$, where $\varepsilon_T = \varepsilon_{\mathcal{H}_e\Lambda_e}(\sqrt{\mathbf{\Sigma}_3})$ and

$$\mathbf{\Sigma}_0 = (\mathcal{H}_e\mathcal{H}_e^\dagger)\sigma_s^2 + \sigma_e^2\mathbf{I} \text{ and } \mathbf{\Sigma}_3^{-1} = (\mathcal{H}_e\mathcal{H}_e^\dagger)^{-1}\sigma_s^{-2} + \sigma_e^{-2}\mathbf{I}. \tag{9}$$

We have thus the following bound for the information leakage (Equation (8) with ε_T replaced by $4\varepsilon_T$).

$$\mathbb{I}(M;\mathbf{y}_e) \leq 8T\varepsilon_T R - 8\varepsilon_T \log 8\varepsilon_T. \tag{10}$$

Therefore, if the flatness factor $\varepsilon_T = \varepsilon_{\mathcal{H}_e\Lambda_e}(\sqrt{\mathbf{\Sigma}_3}) = O(1/T)$, the leakage vanishes as T increases *for the specific realization $\mathcal{H}_e$*. To achieve strong secrecy universally, we must, however, ensure the existence of a lattice with vanishing flatness factor for *all* possible $\mathbf{\Sigma}_3$. The universality discussion is omitted due to space constraints (full details are avaliable in [5]), but can be obtained similar to reliability in [6], with channel quantization. The secrecy condition, implies, in turn, that semantic security is possible for any VNR

$$\gamma_{\mathcal{H}_e\Lambda_e^T}(\sqrt{\mathbf{\Sigma}_3}) = \frac{|\mathcal{H}_e^\dagger\mathcal{H}_e|^{1/n_aT}V(\Lambda_e)^{1/n_aT}}{|\mathbf{\Sigma}_3|^{1/n_aT}} < \pi \text{ or} \tag{11}$$

$$V(\Lambda_e)^{1/n_aT} < \left|\mathbf{I} + \rho_e\mathbf{H}_e^\dagger\mathbf{H}_e\right|^{-1/n_a}\pi\sigma_s^2 = (\pi\sigma_s^2)e^{-C_e/n_a}.$$

Number-of-Antenna Mismatch. The last section assumed that the number of eavesdropper receive antennas an transmit antennas are *equal*. However, due to universality, the arguments can be extended to any number of eavesdropper antennas. We provide a sketch of the case $n_e < n_a$.

Let $\tilde{\mathbf{H}}_e \in \mathbb{C}^{(n_a-n_e)\times n_a}$ be a completion of $\mathbf{H}_e$ in (1) and consider the following surrogate (augmented) MIMO channel

$$\begin{pmatrix} \mathbf{Y}_e \\ \tilde{\mathbf{Y}}_e \end{pmatrix} = \begin{pmatrix} \mathbf{H}_e \\ \beta\tilde{\mathbf{H}}_e \end{pmatrix}\mathbf{X} + \begin{pmatrix} \mathbf{W}_e \\ \tilde{\mathbf{W}}_e \end{pmatrix},$$

where $\tilde{\mathbf{H}}_e$ is scaled so that the capacity of the new channel is arbitrarily close to the original one. Indeed for any full rank completion $\tilde{\mathbf{H}}_e$, from the matrix determinant lemma, we have $|\mathbf{I} + \rho_e\overline{\mathbf{H}}_e^\dagger\overline{\mathbf{H}}_e| \geq e^{C_e}$ Therefore, by making $\beta \to 0$, the left-hand side tends to e^{C_e}. For any signal $\mathbf{X}$, the information leakage of the surrogate channel is strictly greater then the original one (the the eavesdropper's original channel is stochastically degraded with respect to the augmented one). A universally secure code for the $n_a \times n_a$ MIMO compound channel will have vanishing information leakage for the surrogate $n_a \times n_a$ channel (for *any* completion) and therefore will also be secure for the original $n_e \times n_a$ channel.

D. Proof of Theorem 1: Achievable Secrecy Rates

From the previous subsections, semantic security can be achievable if Λ_b and Λ_e satisfy

1) Reliability (4): $\gamma_{\mathbf{R}_b\Lambda_b}(\sigma_b) > \pi e$
2) Secrecy (11): $\gamma_{\mathcal{H}_e\Lambda_e}(\sqrt{\mathbf{\Sigma}_3}) < \pi$
3) Sub-gaussianity of equivalent noise and power constraint: $\varepsilon_{\Lambda_e}(\sigma_s) \to 0$

The first two conditions can be satisfied for rates up to

$$\log|\mathbf{I} + \rho_b\mathbf{H}_b^\dagger\mathbf{H}_b| - \log|\mathbf{I} + \rho_e\mathbf{H}_e^\dagger\mathbf{H}_e| - n_a$$

nats per channel use, but the last conditions may, *a priori*, limit these rates to certain signal-to-noise ratio (SNR) regimes. However if condition 2) is satisfied, we automatically satisfy the condition for $\varepsilon\Lambda_e(\sigma_s) \to 0$, since

$$\frac{V(\Lambda_e)^{1/n_aT}}{\sigma_s^2} \leq \frac{V(\Lambda_e)^{1/n_aT}}{e^{-C_e/n_a}\sigma_s^2} < \pi.$$

Therefore if $(\Lambda_b^T, \Lambda_e^T)$ is a sequence of nested universally-good/universally-secure pairs lattices, then we can achieve rates up to $R \leq (C_b - C_e - n_a)^+$.

We conjecture that this gap can be reduced with better bounds on the variational distance with respect to the flatness factor. Theorem 1 is also a slight improvement on the main result of [11, Thm. 5] in the sense that one of the conditions on the SNR of Bob ($\mathrm{SNR}_b > e$) is no longer needed.

V. Algebraic Constructions

We close this paper with an alternative method for achieving semantic secrecy, assuming that the lattice admit an "algebraic reduction" and can absorb part of the channel state. In this method, inspired by previous works [7], [13] there is no increase in blocklength due to channel quantization and, in fact, any code which is good for the wiretap *Gaussian* channel can be coupled with this technique, as long as it also possesses an additional algebraic structure.

A. Algebraic Approach

Following [16], we define a lattice Λ_e^T admitting algebraic reduction. In the sequel we denote the Frobenius norm of a matrix by $\|\mathbf{M}\|_F \triangleq \sqrt{\mathrm{tr}(\mathbf{M}^\dagger \mathbf{M})}$.

Definition 2. *We say that Λ admits algebraic reduction if for any unit determinant matrix $\mathbf{M} \in \mathbb{C}^{n_a T \times n_a T}$ there exists a matrix decomposition of the form $\mathbf{M} = \mathbf{E}\mathbf{U}$, where $\mathbf{E}$ and $\mathbf{U}$ are also unit-determinant satisfying the following properties:*

1) $\mathbf{U}\Lambda = \Lambda$ *and*
2) $\left\|\mathbf{E}^{-1}\right\|_F \leq \alpha$ *for some absolute constant α that does not depend on $\mathbf{M}$.*

The Golden Code is one example of lattice that admits algebraic reduction [13]. Any lattice built from the generalized Construction A admits a similar reduction. Furthermore, if we can relax the requirement (1) to include equivalence instead of equality, it is possible to exhibit constructions that admit algebraic reduction for any number of antennas [6].The proof of the following lemma shows that lattices admitting algebraic reduction has bounded flatness factor. The proof is omitted due to space constraints.

Lemma 3. *Suppose that $\Lambda \subset \mathbb{C}^{n_a T}$ is such that its dual, Λ^*, admits algebraic reduction. Then*

$$\varepsilon_\Lambda(\sqrt{\mathbf{\Sigma}}) \leq \varepsilon_\Lambda\left(\sqrt{\alpha^{-1}(\det \mathbf{\Sigma})^{1/n_a T}}\right).$$

Therefore, for any channel realization, a sufficient condition for the flatness factor in (11), $\varepsilon_{\mathcal{H}_e \Lambda_e}(\sqrt{\mathbf{\Sigma}_3})$, to vanish is that the upper bound in Lemma 3 vanishes.

This can be achieved provided that:

$$V(\Lambda_e)^{1/n_a T} < \pi e^{-C_e/n_a} \alpha^{-1} \sigma_s^2. \tag{12}$$

Notice that this last expression depends only on the determinant of $\mathbf{\Sigma}_3$ or on the capacity of the eavesdropper channel, not on any individual realization. For this condition to hold, we only need a sequence of secrecy-good lattices for an eavesdropping AWGN channel with smaller noise variance (by factor α^{-1}). Therefore, the following result holds.

Theorem 2. *Let $(\Lambda_b^T, \Lambda_e^T)$ be a sequence of nested lattices where: (i) Λ_b^T is universally good for the compound MIMO channel and (ii) Λ_e^T satisfies Definition 2 and is secrecy good* ***for the AWGN channel*** *(Condition* (12)*). Then nested lattice Gaussian coding achieves any secrecy rates up to*

$$R \leq (C_b - C_e - n_a - n_a \log(\alpha))^+.$$

Notice the extra gap with respect to Theorem 1. Although we have conjectured that the gap in Theorem 1 can be essentially removed, this is not the case for $\log \alpha$ in Theorem 2. Indeed, since α cannot be smaller than $\sqrt{n_a}$, this gap is always larger than $n_a \log n_a$. However the code construction can be reduced to the problem of finding good lattices for the Gaussian wiretap channel (with some additional algebraic structure), making the design potentially more practical.

-

References

[1] J. Barros and M. R. D. Rodrigues. Secrecy capacity of wireless channels. In *2006 IEEE International Symposium on Information Theory*, pages 356–360, July 2006.

[2] J.-C. Belfiore and F. Oggier. An error probability approach to MIMO wiretap channels. *IEEE Trans. Commun.*, 61(8):3396–3403, August 2013.

[3] M. Bloch and J. Barros. *Physical Layer Security: From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[4] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin. Wireless information-theoretic security. *IEEE Transactions on Information Theory*, 54(6):2515–2534, June 2008.

[5] A. Campello, C. Ling, and J.-C. Belfiore. Semantically Secure Lattice Codes for Compound Mimo Channels. *preprint.*

[6] A. Campello, C. Ling, and J. C. Belfiore. Algebraic lattice codes achieve the capacity of the compound block-fading channel. In *IEEE International Symposium on Information Theory (ISIT)*, pages 910–914, July 2016.

[7] E. Viterbo G. Rekaya, J-C. Belfiore. A very efficient lattice reduction tool on fast fading channels. In *Proceedings of the Internation Symposium on Information Theory and its Applications (ISITA), Parma, Italy*, 2004.

[8] A. Khisti. Interference alignment for the multiantenna compound wiretap channel. *IEEE Transactions on Information Theory*, 57(5):2976–2993, May 2011.

[9] Y. Liang, H. V. Poor, and S. Shamai. Secrecy capacity region of fading broadcast channels. In *2007 IEEE International Symposium on Information Theory*, pages 1291–1295, June 2007.

[10] Yingbin Liang, Gerhard Kramer, H. Vincent Poor, and Shlomo Shamai. Compound wiretap channels. *Eurasip Journal on Wireless Communications and Networking*, 2009, 2009.

[11] C. Ling, L. Luzzi, J.-C. Belfiore, and D. Stehlé. Semantically secure lattice codes for the Gaussian wiretap channel. *IEEE Trans. Inform. Theory*, 60(10):6399–6416, Oct. 2014.

[12] H.-A. Loeliger. Averaging bounds for lattices and linear codes. *IEEE Transactions on Information Theory*, 43(6):1767–1773, Nov 1997.

[13] L. Luzzi, G. Rekaya-Ben Othman, and J.-C. Belfiore. Augmented lattice reduction for MIMO decoding. *IEEE Trans. Wireless Commun.*, 9:2853–2859, September 2010.

[14] Laura Luzzi, Roope Vehkalahti, and Cong Ling. Almost universal codes for MIMO wiretap channels. *CoRR*, abs/1611.01428, 2016.

[15] F. Oggier, P. Solé, and J.-C. Belfiore. Lattice codes for the wiretap gaussian channel: Construction and analysis. *IEEE Trans. Inform. Theory*, 62(10):5690–5708, Oct. 2016.

[16] G. Rekaya-Ben Othman, L. Luzzi, and J. C. Belfiore. Algebraic reduction for the Golden Code. In *2010 IEEE Information Theory Workshop on Information Theory*, pages 1–5, Jan 2010.

[17] Reihaneh Safavi-Naini and Ran Canetti, editors. *Semantic Security for the Wiretap Channel*, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[18] R. F. Schaefer and S. Loyka. The Secrecy Capacity of Compound Gaussian MIMO Wiretap Channels. *IEEE Transactions on Information Theory*, 61(10):5535–5552, 2015.

Precoding via Approximate Message Passing with Instantaneous Signal Constraints

Ali Bereyhi*, Mohammad Ali Sedaghat†, Ralf R. Müller*,
*Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), †Cisco Optical GmbH Nürnberg
ali.bereyhi@fau.de, mohammad.sedaghat@fau.de, ralf.r.mueller@fau.de

***Abstract*–This paper proposes a low complexity precoding algorithm based on the recently proposed Generalized Least Square Error (GLSE) scheme with generic penalty and support. The algorithm iteratively constructs the transmit vector via Approximate Message Passing (AMP). Using the asymptotic decoupling property of GLSE precoders, we derive closed form fixed point equations to tune the parameters in the proposed algorithm for a general set of instantaneous signal constraints. The tuning strategy is then utilized to construct transmit vectors with restricted peak-to-average power ratios and to efficiently select a subset of transmit antennas. The numerical investigations show that the proposed algorithm tracks the large-system performance of GLSE precoders even for a moderate number of antennas.**

I. Introduction

For a given precoding support $\mathbb{X} \subset \mathbb{C}$ and penalty function $u(\cdot) : \mathbb{X} \mapsto \mathbb{R}$, the Generalized Least Square Error (GLSE) precoder constructs the transmit vector $\boldsymbol{x} \in \mathbb{X}^N$ from the data vector $\boldsymbol{s} \in \mathbb{C}^K$ and the channel matrix $\mathbf{H} \in \mathbb{C}^{N \times K}$ as $\boldsymbol{x} = \mathrm{glse}\,(\boldsymbol{s}, \rho|\mathbf{H})$ where ρ is a power control factor and [1]

$$\mathrm{glse}\,(\boldsymbol{s}, \rho|\mathbf{H}) = \operatorname*{argmin}_{\boldsymbol{v} \in \mathbb{X}^N} \|\mathbf{H}\boldsymbol{v} - \sqrt{\rho}\,\boldsymbol{s}\|^2 + u(\boldsymbol{v}). \tag{1}$$

The generality of $\mathbb{X}$ and $u(\cdot)$ allows for addressing various forms of constraints on the transmit vector. Compared to the classical approaches for imposing such constraints, the studies in [1]–[4] have shown significant enhancements obtained via the GLSE precoding scheme. Nevertheless, the computational complexity of this scheme has been remained as the main challenge and is intended to be addressed in this paper.

The main motivation of this study comes from the great deal of interest being received recently by massive Multiple-Input Multiple-Output (MIMO) systems [5]. Form implementational points of view, however, these systems confront the problem of high Radio Frequency (RF)-cost which raises due to the vast number of RF-chains needed in such setups. The initial approach to overcome this issue is to restrict the Peak-to-Average Power Ratio (PAPR) of the transmit vector [6], [7]. In this case, nonlinear power amplifiers with lower dynamic ranges can be employed, and the total RF-cost can be significantly reduced. Another approach is Transmit Antenna Selection (TAS) [8], [9] in which a subset of transmit antennas is kept active at each transmission interval, and therefore, the number of required RF-chains is reduced. Although such approaches combat the issue of high RF-cost, the conventional algorithms significantly degrade the performance. In this case, GLSE precoders reduce this degradation by finding the optimal transmit vector which satisfies the constraints imposed by these approaches. In general, GLSE precoders solve an optimization problem in each transmission interval. This task is not trivial for choices of $u(\cdot)$ and $\mathbb{X}$ which are non-convex. For cases with convex optimization problems, the precoder can be implemented via generic linear programming algorithms. The high computational complexity of these algorithms for large dimensions, however, leaves the implementation of GLSE precoders as an issue in massive MIMO setups. Generalized Approximate Message Passing (GAMP) [10] proposes a low complexity iterative approach for several estimation problems based on approximating the loopy belief propagation algorithm in the large limit [11]. The algorithm is known to considerably outperform other available iterative approaches. The underlying estimation problems, which are addressed by GAMP, are mathematically similar to the GLSE precoding scheme, and therefore, the algorithm can be employed to design a class of iterative precoders based on the GLSE scheme.

The main contribution of this paper is to adopt and tune the GAMP algorithm to address the GLSE precoding scheme, recently proposed in [1]–[4]. The developed iterative scheme is referred to as "GLSE-GAMP" precoding and exhibits low complexity characteristic. Using the fact that the GLSE and GLSE-GAMP precoders consider same optimization problems, we further propose a tuning strategy based on the asymptotic results in [1]–[4] derived via the replica method. Our numerical investigations show that the performance of GLSE-GAMP precoders tuned by the proposed strategy is accurately consistent with asymptotics of corresponding GLSE precoders.

Notation

Throughout the paper, scalars, vectors and matrices are represented with non-bold, bold lower case and bold upper case letters, respectively. $\mathbf{I}_K$ is a $K \times K$ identity matrix, and $\mathbf{H}^\mathsf{H}$ is the Hermitian of $\mathbf{H}$. The set of real and integer numbers are denoted by $\mathbb{R}$ and $\mathbb{Z}$, and $\mathbb{C}$ represents the complex plane. For $s \in \mathbb{C}$, $\mathsf{Re}\{s\}$, $\mathsf{Im}\{s\}$ and $\mathsf{s} := [\mathsf{Re}\{s\}\ \mathsf{Im}\{s\}]^\mathsf{T}$ identify the real part, imaginary part and augmented vector, respectively, and the expression $\mathsf{s} \in \mathbb{S}$ indicates that s is the augmented version of $s \in \mathbb{S}$. For $\mathbf{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \dots, f_n(\boldsymbol{x})]^\mathsf{T}$, the gradient operator is defined as $\nabla_{\boldsymbol{x}} \mathbf{f}(\boldsymbol{x}) := [\nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}), \dots, \nabla_{\boldsymbol{x}} f_n(\boldsymbol{x})]^\mathsf{T}$. $\|\cdot\|$ and $\|\cdot\|_1$ denote the Euclidean and ℓ_1-norm, respectively. Considering the random variable x, p_x represents either the probability mass or density function. Moreover, $\mathbb{E}$ identifies the expectation. For sake of compactness, $\{1, \dots, N\}$ is abbreviated by $[N]$, and we define $\tilde{\phi}(x, \lambda) := \exp(-x^2/\lambda)$ and $\tilde{\mathrm{Q}}(x, \lambda) := \int_x^\infty \tilde{\phi}(u, \lambda)\mathrm{d}u/\lambda$ for a given non-negative real λ.

This work was supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under Grant No. MU 3735/2-1.

II. Problem Formulation

Consider a Gaussian broadcast MIMO setup in which a sequence of data symbols $\{s_k\}$ for $k \in [K]$ is transmitted to K single-antenna users simultaneously. The transmitter is equipped with N transmit antennas. The channel is considered to be quasi-static fading and perfectly known at the transmitter. By employing the GLSE precoding scheme given in (1) with some penalty $u(\cdot)$ and precoding support $\mathbb{X} \subseteq \mathbb{C}$, the transmit vector is constructed as $\boldsymbol{x}_{N\times 1} = \mathrm{glse}\left(\boldsymbol{s}, \rho | \mathbf{H}\right)$ where $\boldsymbol{s}_{K\times 1} := \left[s_1, \dots, s_K\right]^{\mathsf{T}}$ and ρ is a non-negative power control factor. For this setup, we assume that the following constraints hold.

(a) $\boldsymbol{s}_{K\times 1}$ has independent and identically distributed (i.i.d.) zero-mean complex Gaussian entries with unit variance.

(b) $u(\cdot)$ decouples meaning that $u(\boldsymbol{v}) = \sum_{j=1}^{N} u(v_j)$.

(c) N and K grow large, such that the load factor $\alpha := K/N$ is kept fixed in both N and K.

(d) $\mathbf{H}^{\mathsf{H}}\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathsf{H}}$ in which $\mathbf{U}$ is an $N \times N$ unitary matrix, and $\mathbf{D}$ is a diagonal matrix with asymptotic eigenvalue distribution $\mathrm{p}_{\mathbf{D}}$. For $\mathrm{p}_{\mathbf{D}}$, we define the Stieltjes transform as $\mathrm{G}_{\mathbf{D}}(s) = \mathbb{E}\left\{(d-s)^{-1}\right\}$ with the expectation being taken over $d \sim \mathrm{p}_{\mathbf{D}}$ and the R-transform as $\mathrm{R}_{\mathbf{D}}(\omega) = \mathrm{G}_{\mathbf{D}}^{-1}(-\omega) - \omega^{-1}$ where $\mathrm{G}_{\mathbf{D}}^{-1}(\cdot)$ denotes the inverse with respect to composition.

By proper choices of the support $\mathbb{X}$ and penalty $u(\cdot)$, the GLSE precoder can impose several constraints on the transmit vector.

- Setting $u(\boldsymbol{v}) = \lambda \|\boldsymbol{v}\|^2$ and $\mathbb{X} = \left\{x \in \mathbb{C} : |x|^2 < P\right\}$, the transmit vector is restricted to have a limited PAPR. In fact in this case, the peak power is set to P and a desired constraint on the PAPR is imposed by tuning λ such that the average power is accordingly restricted.
- Let $u(\boldsymbol{v}) = \lambda \|\boldsymbol{v}\|^2 + \mu \|\boldsymbol{v}\|_1$ and $\mathbb{X} = \mathbb{C}$; then, the number of active transmit antennas is constrained.

III. GLSE-GAMP Precoders

The GLSE scheme can be considered as a max-sum problem which can be addressed via the GAMP algorithm [10].

A. GAMP Algorithm

The GAMP algorithm, proposed in [10], intends to estimate $\boldsymbol{v}_{N\times 1}$ from $\boldsymbol{s}_{K\times 1}$ iteratively considering the following setup.

(a) Each entry of $\boldsymbol{v}$ is generated from the corresponding entry of some $\boldsymbol{a} \in \mathbb{A}^N$ via $\mathrm{p}_{v|a}$.

(b) The entries of $\boldsymbol{s}$ are obtained form the entries of the vector $\boldsymbol{z}_{K\times 1}$ through identical scalar channels with $\mathrm{p}_{s|z}$.

(c) $\boldsymbol{z}$ is a random linear transform of $\boldsymbol{v}$, i.e., $\boldsymbol{z} = \mathbf{H}\boldsymbol{v}$ for some random $K \times N$ matrix $\mathbf{H}$.

Depending on the estimation scheme, the GAMP algorithm is developed to address the "max-sum" or "sum-product" problems. The max-sum GAMP algorithm iteratively determines the Maximum-A-Posterior (MAP) estimation

$$\boldsymbol{x} = \operatorname*{argmax}_{\boldsymbol{v}} \sum_{n=1}^{N} f_{\mathrm{in}}(v_n, a_n) + \sum_{k=1}^{K} f_{\mathrm{out}}(z_k, y_k) \tag{2}$$

for some scalar functions $f_{\mathrm{in}}(\cdot,\cdot)$ and $f_{\mathrm{out}}(\cdot,\cdot)$ which represent the conditional distributions $\mathrm{p}_{v|a}$ and $\mathrm{p}_{s|z}$. The sum-product GAMP algorithm, moreover, addresses the Minimum Mean-Square-Error (MMSE) estimation where $\boldsymbol{x} = \mathbb{E}\left\{\boldsymbol{v} | \boldsymbol{s}, \boldsymbol{a}\right\}$.

B. The GAMP-GLSE Algorithm

By comparing GLSE precoding with (2), it is observed that the precoding scheme solves a max-sum problem in which $\boldsymbol{z} := \mathbf{H}\boldsymbol{v}$ with $\mathbf{H}$ being the channel matrix, $v_k \in \mathbb{X}$ for $k \in [K]$, and $f_{\mathrm{in}}(v_n, a_n) = -u(v_n)$ and $f_{\mathrm{out}}(z_k, s_k) = -|z_k - \sqrt{\rho} s_k|^2$. As the result, the GAMP algorithm can be applied to iteratively construct the transmit vector $\boldsymbol{x}$. By some lines of derivations, the max-sum GAMP algorithm can be adopted to the GLSE scheme in (1). The resulting algorithm is referred to as "GLSE-GAMP" algorithm and is represented in Algorithm 1 for the precoding support $\mathbb{X} \subseteq \mathbb{C}$ and the complex-valued matrix $\mathbf{H}$. The variables and functions in the algorithm, for $k \in [K]$ and $n \in [N]$, are defined as follows.

- The real two-dimensional vectors $\mathbf{w}_k$, $\mathbf{z}_k$, $\mathbf{y}_k$, $\mathbf{s}_k$, $\mathbf{u}_n$ and $\mathbf{x}_n$ are the augmented forms of the complex scalars w_k, z_k, y_k, s_k, u_n and x_n, respectively.
- The matrices $\mathbf{R}_k^{\mathrm{w}}$, $\mathbf{R}_k^{\mathrm{y}}$, $\mathbf{R}_n^{\mathrm{u}}$ and $\mathbf{R}_n^{\mathrm{x}}$ are real 2×2 matrices, and $\mathbf{Q}_{kn}$ is defined as
$$\mathbf{Q}_{kn} := \begin{bmatrix} \mathsf{Re}\left\{h_{kn}\right\} & -\mathsf{Im}\left\{h_{kn}\right\} \\ \mathsf{Im}\left\{h_{kn}\right\} & \mathsf{Re}\left\{h_{kn}\right\} \end{bmatrix} \tag{3}$$
with h_{kn} representing the entry (k,n) of $\mathbf{H}$.
- $\mathrm{g}_{\mathrm{out}}\left(\cdot\right)$ is the output thresholding function defined as
$$\mathrm{g}_{\mathrm{out}}\left(\mathbf{w}, \mathbf{s}, \mathbf{R}\right) := \nabla_{\mathbf{w}} \min_{\mathbf{z}\in\mathbb{C}} \mathcal{E}_{\mathrm{out}}(\mathbf{z}, \mathbf{w}, \mathbf{s}, \mathbf{R}) \tag{4}$$
where the function $\mathcal{E}_{\mathrm{out}}(\cdot)$ is determined by
$$\begin{aligned}\mathcal{E}_{\mathrm{out}}(\mathbf{z}, \mathbf{w}, \mathbf{s}, \mathbf{R}) = \frac{1}{2}(\mathbf{z}-\mathbf{w})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{w}) \\ + \|\mathbf{z} - \sqrt{\rho}\mathbf{s}\|^2 \end{aligned} \tag{5}$$
- $\mathrm{g}_{\mathrm{in}}\left(\cdot\right)$ is the input thresholding function being defined as
$$\mathrm{g}_{\mathrm{in}}\left(\mathbf{u}, \mathbf{R}\right) := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{X}} \mathcal{E}_{\mathrm{in}}(\mathbf{x}, \mathbf{u}, \mathbf{R}). \tag{6}$$
where the function $\mathcal{E}_{\mathrm{in}}(\cdot)$ is evaluated by
$$\mathcal{E}_{\mathrm{in}}(\mathbf{x}, \mathbf{u}, \mathbf{R}) = \frac{1}{2}(\mathbf{u}-\mathbf{x})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{u}-\mathbf{x}) + u(\mathbf{x}). \tag{7}$$
- The initial conditions are $\mathbf{x}_n(1) = \arg\min_{\mathbf{x}\in\mathbb{X}} u(\mathbf{x})$ and $\mathbf{R}_n^{\mathrm{x}}(1) = \left[\nabla_{\mathbf{x}}^2 u(\mathbf{u}_n(1))\right]^{-1}$.

The update rules in Algorithm 1 are derived by extending the sum-max GAMP algorithm to the case with a complex-valued matrix $\mathbf{H}$ and an arbitrary input support $\mathbb{X} \subseteq \mathbb{C}$. The extension is followed by determining the update rules for the corresponding loopy belief propagation algorithm and then taking some steps similar to [10, Appendix C]. The detailed derivations are skipped due to the page limit and is represented in the extended version of the manuscript.

Remark 1: One should distinguish between the GLSE scheme and the GLSE-GAMP algorithm. In fact, the former is a least square based scheme to design transmit signals which fulfill some desired constraints. The GLSE-GAMP algorithm, on the

```
Algorithm 1 GLSE-GAMP Precoding Algorithm
Initiate Start from $t = 1$ and for $k \in [K]$ let $\mathbf{y}_k(0) = \mathbf{0}$. Set
  $\mathbf{x}_n(1)$ and $\mathbf{R}^{\mathrm{x}}_n(1)$ for $n \in [N]$ to their initial conditions.
  while $t < T$
    for $k \in [K]$
      $\mathbf{R}^{\mathrm{w}}_k(t) = \sum_{n=1}^{N} \mathbf{Q}_{kn} \mathbf{R}^{\mathrm{x}}_n(t) \mathbf{Q}^{\mathsf{T}}_{kn}$    (8a)
      $\mathbf{z}_k(t) = \sum_{n=1}^{N} \mathbf{Q}_{kn} \mathbf{x}_n(t)$    (8b)
      $\mathbf{w}_k(t) = \mathbf{z}_k(t) - \mathbf{R}^{\mathrm{w}}_k(t) \mathbf{y}_k(t-1)$    (8c)
      $\mathbf{y}_k(t) = \mathbf{g}_{\mathrm{out}}(\mathbf{w}_k(t), \mathbf{s}_k, \mathbf{R}^{\mathrm{w}}_k(t))$    (8d)
      $\mathbf{R}^{\mathrm{y}}_k(t) = -\nabla_{\mathbf{w}}\, \mathbf{g}_{\mathrm{out}}(\mathbf{w}_k(t), \mathbf{s}_k, \mathbf{R}^{\mathrm{w}}_k(t))$    (8e)
    end for
    for $n \in [N]$
      $\mathbf{R}^{\mathrm{u}}_n(t) = \left[ \sum_{k=1}^{K} \mathbf{Q}^{\mathsf{T}}_{kn} \mathbf{R}^{\mathrm{y}}_k(t) \mathbf{Q}_{kn} \right]^{-1}$    (9a)
      $\mathbf{u}_n(t) = \mathbf{x}_n(t) + \mathbf{R}^{\mathrm{u}}_n(t) \left[ \sum_{k=1}^{K} \mathbf{Q}^{\mathsf{T}}_{kn} \mathbf{y}_k(t) \right]$    (9b)
      $\mathbf{x}_n(t+1) = \mathbf{g}_{\mathrm{in}}(\mathbf{u}_n(t), \mathbf{R}^{\mathrm{u}}_n(t))$    (9c)
      $\mathbf{R}^{\mathrm{x}}_n(t+1) = \left[ \nabla_{\mathbf{u}}\, \mathbf{g}_{\mathrm{in}}(\mathbf{u}_n(t), \mathbf{R}^{\mathrm{u}}_n(t)) \right] \mathbf{R}^{\mathrm{u}}_n(t)$    (9d)
    end for
  end while
Output: $\mathbf{x}_n(T)$ for $n \in [N]$.
```

other hand, proposes an iterative approach based on GAMP to address the GLSE scheme. For some choices of the penalty function, precoding support and channel matrix, the GLSE-GAMP algorithm converges to the transmit signal given by the GLSE scheme. There are however some particular cases in which the GLSE-GAMP algorithm does not converge. For these cases, Algorithm 1 does not give the desired transmit signal. To avoid the divergence in such cases, we need to modify the algorithm. This issue is briefly discussed in Section V.

In contrast to GLSE precoders, GLSE-GAMP precoders exhibit low complexity characteristic. Considering Algorithm 1 and noting that the matrices in (8a)-(9d) are fixed 2×2 matrices, it is straightforward to show that the total worst-case complexity of GLSE-GAMP precoders per iteration is $\mathcal{O}(KN)$. The number of iterations, moreover, does not grow with the dimensions. Therefore, one can conclude that the overall complexity of the precoding scheme is $\mathcal{O}(KN)$ as well.

C. Tuning GLSE-GAMP precoders

In order to impose a given set of constraint on the transmit signal, the corresponding GLSE-GAMP precoder should be tuned. As an example, consider the case in which the number of active transmit antennas, as well as the average transmit power, is desired to be restricted via a GLSE-GAMP precoder. In this case, one may set $\mathbb{X} = \mathbb{C}$ and $u(\boldsymbol{v}) = \lambda \|\boldsymbol{v}\|^2 + \mu \|\boldsymbol{v}\|_1$. The factors λ and μ in this case control the average transmit power and the fraction of active antennas, respectively. Consequently for given constraints, these factors need to be tuned. Nevertheless, the derivation of an exact tuning strategy is not a trivial problem as the constrained parameters, i.e., the average power or fraction of active antennas, cannot be derived in terms of the tuning factors straightforwardly. We therefore propose a tuning strategy based on the asymptotics of the GLSE-GAMP algorithm and its connection to the GLSE scheme. The large-system performance of GLSE-GAMP precoders is studied through asymptotic analyses of "state evolution" equations; see [12] and the references therein. Following the results in the literature, e.g. [13], [14], it is shown that for choices of $\mathbf{H}$, $\mathbb{X}$ and $u(\cdot)$, in which the GLSE-GAMP algorithm converges, the asymptotic performance of the algorithm coincides with the large-system performance of GLSE precoders investigated in [1], [4]. This result indicates that in the large-system limit, the tuning factors for GLSE-GAMP and GLSE precoders are the same. Therefore, for a given set of constraint, we derive the tuning factors of the GLSE-GAMP precoders by tuning the corresponding GLSE precoders.

Tuning Strategy: Assume that the constraints $f_j(\boldsymbol{x})/N = C_j$ are desired to be satisfied via a GLSE-GAMP precoder with penalty $u(\cdot)$ and support $\mathbb{X}$ which are controlled by λ_j for $j \in [J]$. Here, $f_j(\cdot)$ are decoupling functions meaning that $f_j(\boldsymbol{x}) = \sum_{n=1}^{N} f_j(x_n)$. To tune λ_j accordingly, we define

$$\mathrm{x} = \operatorname*{argmin}_{v \in \mathbb{X}} |v - s_0|^2 + \xi\ u(v) \tag{10}$$

where $s_0 \sim \mathcal{CN}\left(0, \sigma^2\right)$ with

$$\sigma^2 = \left[\mathrm{R}_{\mathbf{D}}(-\chi)\right]^{-2} \frac{\partial}{\partial \chi} \left[(\lambda_s \chi - \mathsf{p}) \mathrm{R}_{\mathbf{D}}(-\chi) \right]. \tag{11}$$

and $\xi = \left[\mathrm{R}_{\mathbf{D}}(-\chi)\right]^{-1}$ for χ and p which satisfy $\mathsf{p} = \mathbb{E}|\mathrm{x}|^2$ and

$$\frac{\sigma^2 \chi}{\xi} = \mathbb{E}\mathsf{Re}\left\{\mathrm{x}^* s_0\right\}. \tag{12}$$

The precoder is then accordingly tuned by choosing λ_j for $j \in [J]$ such that the equations $\mathbb{E} f_j(\mathrm{x}) = C_j$ are satisfied.

Derivation: The derivation follows the marginal decoupling property of the GLSE precoders presented in [1], [4]. In fact, using the property, it is concluded that $f_j(\boldsymbol{x})/N$ asymptotically converges to $\mathbb{E} f_j(\mathrm{x})$. By taking the approach illustrated at the beginning of the section, the tuning strategy is obtained.

The proposed tuning strategy evaluate the decoupled GLSE precoder[1] by finding χ and p form the fixed-point equations. The asymptotic constrained parameters are then determined by taking the expectation $\mathbb{E} f_j(\mathrm{x})$ and set it equal to C_j. One should note that the strategy in general is heuristic, since it tunes the precoders for the large-system limit. Nevertheless,

[1] See Proposition 2 in [1] for the decoupling property of GLSE precoders. A more general version of the property is represented in [4, Section II-A].

the numerical investigations show that for several cases, the GLSE-GAMP precoders are well tuned via this strategy.

IV. Applications of GLSE-GAMP Precoders

In this section, we investigate two special cases of GLSE-GAMP precoders with TAS and limited PAPR. Throughout the analyses, we assume that $\mathbf{H}$ represents an i.i.d. Rayleigh fading channel with variance $1/N$, i.e., $\mathrm{R}_{\mathbf{J}}(\omega) = \alpha(1-\omega)^{-1}$.

A. GLSE-GAMP Precoder with TAS

As it was discussed, TAS can be directly addressed at the transmit side by using GLSE scheme with $u(v) = \lambda|v|^2 + \mu|v|$. The corresponding GLSE-GAMP precoder is therefore given by Algorithm 1 where $\mathbb{X} = \mathbb{C}$, $\mathbf{g}_{\text{out}}(\mathbf{w}, \mathbf{s}, \mathbf{R}) = \mathbf{G}_{\text{w}}\mathbf{w} + \mathbf{G}_{\text{s}}\mathbf{s}$ and $\nabla_{\mathbf{w}}\, \mathbf{g}_{\text{out}}(\mathbf{w}, \mathbf{s}, \mathbf{R}) = \mathbf{G}_{\text{w}}$, respectively with

$$\mathbf{G}_{\text{w}} := -2\,\mathbf{A}^{\mathsf{T}}\mathbf{A} - (\mathbf{A} - \mathbf{I}_2)^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{A} - \mathbf{I}_2) \tag{13a}$$

$$\mathbf{G}_{\text{s}} := -2\left[2\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{R} - \mathbf{A}^{\mathsf{T}} + (\mathbf{A} - \mathbf{I}_2)^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{A}\mathbf{R}\right] \tag{13b}$$

and $\mathbf{A} := (\mathbf{I}_2 + 2\mathbf{R})^{-1}$. For the input thresholding function, the analytic evaluation of the function from the augmented form in (6) is not trivial. We thus employ the complex scalar form of the equation which results in $\mathbf{g}_{\text{in}}(\mathbf{u}, \mathbf{R}) = \mathbf{G}(\mathbf{u})\mathbf{f}(\mathbf{u})$ and $\nabla_{\mathbf{u}}\, \mathbf{g}_{\text{in}}(\mathbf{u}, \mathbf{R}) = \mathbf{G}(\mathbf{u})\mathbf{F}(\mathbf{u})$ where

$$\mathbf{G}(\mathbf{u}) = \begin{cases} \mathbf{G}_{\text{u}} & \|\mathbf{u}\| \geq \tau \\ 0 & \|\mathbf{u}\| < \tau \end{cases} \tag{14}$$

with $\tau := 2\mu\left[\mathrm{Tr}\{\mathbf{R}^{-1}\}\right]^{-1}$ and $\mathbf{G}_{\text{u}} := (\mathbf{I}_2 + 2\lambda\mathbf{R})^{-1}$, and

$$\mathbf{f}(\mathbf{u}) := \left[1 - \frac{\tau}{\|\mathbf{u}\|}\right]\mathbf{u}, \tag{15a}$$

$$\mathbf{F}(\mathbf{u}) := \frac{\tau}{\|\mathbf{u}\|^3}\mathbf{u}\mathbf{u}^{\mathsf{T}} + \left[1 - \frac{\tau}{\|\mathbf{u}\|}\right]\mathbf{I}_2. \tag{15b}$$

By setting $\mu = 0$, the GLSE scheme reduces to Regularized Zero Forcing (RZF) precoding, and thus, the GLSE-GAMP algorithm iteratively constructs the output of the RZF precoder.

Tuning Strategy: We employ the strategy in Section III-C to tune μ and λ such that the fraction of active antennas and the average transmit power are η and P, respectively. For this case, $J = 2$ and $f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ and $f_2(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$. Consequently, λ and μ are determined from the fixed-point equations $\tilde{\phi}(\xi\mu; \theta) = \eta$ for $\theta = (\rho + P)/\alpha$ and

$$(1 + 2\xi\lambda)^2 = \frac{\theta}{P}\left[\eta - 2\xi\mu\tilde{\mathrm{Q}}(\xi\mu; \theta)\right] \tag{16}$$

and ξ is determined in terms of λ and μ through

$$\alpha\xi = \frac{1}{2} + \frac{\xi}{1 + 2\xi\lambda}\left[\eta - \xi\mu\tilde{\mathrm{Q}}(\xi\mu; \theta)\right]. \tag{17}$$

B. GLSE-GAMP Precoder with PAPR Constraint

The precoder in Section IV-A can further take the PAPR constraint into account by setting $\mathbb{X} = \left\{x \in \mathbb{C} : |x|^2 < P_{\max}\right\}$. The support in this case imposes a peak power constraint on the transmit signal which along with the penalty function restricts both the PAPR and the number of active antennas[1]. Considering Algorithm 1, the output function for this setup remains unchanged , and the input function reads

$$\mathbf{g}_{\text{in}}(\mathbf{u}, \mathbf{R}) = \begin{cases} \dfrac{\mathbf{u}}{\|\mathbf{u}\|}\sqrt{P_{\max}} & \tilde{\tau} \leq \|\mathbf{u}\| \\ \mathbf{G}_{\text{u}}\mathbf{f}(\mathbf{u}) & \tau \leq \|\mathbf{u}\| < \tilde{\tau} \\ 0 & 0 \leq \|\mathbf{u}\| < \tau \end{cases} \tag{18}$$

with the corresponding gradient

$$\nabla_{\mathbf{u}}\, \mathbf{g}_{\text{in}}(\mathbf{u}, \mathbf{R}) = \begin{cases} \dfrac{\tilde{\mathbf{u}}\tilde{\mathbf{u}}^{\mathsf{T}}}{\|\mathbf{u}\|^3}\sqrt{P_{\max}} & \tilde{\tau} \leq \|\mathbf{u}\| \\ \mathbf{G}_{\text{u}}\mathbf{F}(\mathbf{u}) & \tau \leq \|\mathbf{u}\| < \tilde{\tau} \\ 0 & 0 \leq \|\mathbf{u}\| < \tau, \end{cases} \tag{19}$$

where $\tilde{\mathbf{u}} := [\mathrm{u}_2, -\mathrm{u}_1]^{\mathsf{T}}$, $\tau := 2\mu\left[\mathrm{Tr}\{\mathbf{R}^{-1}\}\right]^{-1}$, and

$$\tilde{\tau} := \left(1 + \frac{4\lambda}{\mathrm{Tr}\{\mathbf{R}^{-1}\}}\right)\sqrt{P_{\max}} + \frac{2\mu}{\mathrm{Tr}\{\mathbf{R}^{-1}\}}. \tag{20}$$

$\mathbf{G}_{\text{u}}$, $\mathbf{f}(\mathbf{u})$ and $\mathbf{F}(\mathbf{u})$ are moreover given as in Section IV-A. By setting $\mu = 0$, the precoder employs all the transmit antennas and restricts only the PAPR. In this case, $\mathbf{F}(\mathbf{u}) = \mathbf{I}_2$, $\mathbf{f}(\mathbf{u}) = \mathbf{u}$, and τ reduces to zero.

Tuning Strategy: Consider the same constraints as for the case without the PAPR restriction. From Section III-C, λ and μ for the average power P and the fraction of active antennas η are given by the fixed-point equations $\tilde{\phi}(\xi\mu; \theta) = \eta$ and

$$(1 + 2\xi\lambda)^2 = \frac{\theta}{P}\left[\Delta_1(\xi\mu) - 2\xi\mu\Delta_2(\xi\mu)\right]. \tag{21}$$

Here, ξ is a function of λ and μ which satisfies

$$\alpha\xi = \frac{1}{2} + \frac{\xi}{1 + 2\xi\lambda}\left[\Delta_1(\xi\mu) - 2\xi\mu\Delta_2(\xi\mu)\right]. \tag{22}$$

Moreover, $\theta = (\rho + P)/\alpha$ and we have defined

$$\Delta_1(\xi\mu) := \tilde{\phi}(\xi\mu; \theta) - \tilde{\phi}(\xi\mu + (1 + 2\xi\lambda)\sqrt{P_{\max}}; \theta), \tag{23a}$$

$$\Delta_2(\xi\mu) := \tilde{\mathrm{Q}}(\xi\mu; \theta) - \tilde{\mathrm{Q}}(\xi\mu + (1 + 2\xi\lambda)\sqrt{P_{\max}}; \theta). \tag{23b}$$

V. Numerical Investigations

To investigate the performance of GLSE-GAMP precoders, we define the distortion measure for a given ρ as

$$\mathsf{D}(\rho) := \frac{1}{K}\mathbb{E}\|\mathbf{H}\boldsymbol{x} - \sqrt{\rho}\boldsymbol{s}\|^2 \tag{24}$$

which determines the average distortion caused by the multiuser interference at receive terminals. It is moreover shown that the achievable ergodic rate per user can be bounded from below in terms of $\mathsf{D}(\rho)$ as proved in [2].

The circles in Fig. 1 show the distortion given by the GLSE-GAMP precoder presented in Section IV-A for various inverse load factors $\alpha^{-1} = N/K$ considering several constraints on the number of active antennas. The results have been given for $N = 64$ antennas and $T = 20$ iterations. The asymptotic performances of the corresponding GLSE precoders, derived via the replica method in [4], have been also sketched with solid lines. Here, $\rho = 1$ and λ is set such that $P = 0.3$. As the figure shows, the GLSE-GAMP precoder tracks accurately the

[1] See [1, Section IV-B] for further illustrations.

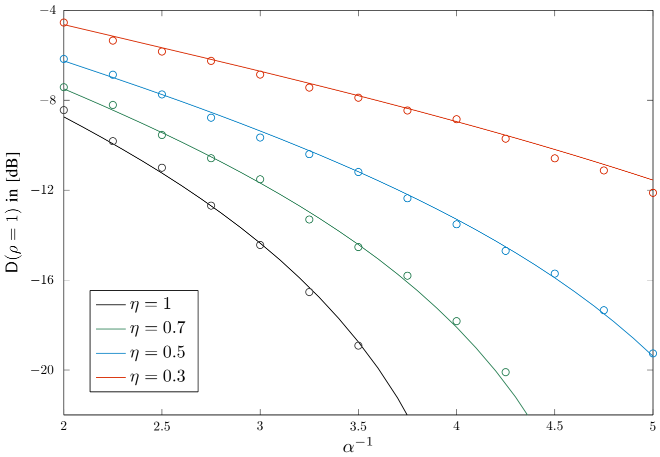


Fig. 1: Distortion at $\rho = 1$ vs. α^{-1} for $P = 0.3$ and various η. Circles depict the performance of the GLSE-GAMP precoder for $N = 64$ and $T = 20$. Solid lines denote the asymptotic performance of the corresponding GLSE scheme determined by the replica method.

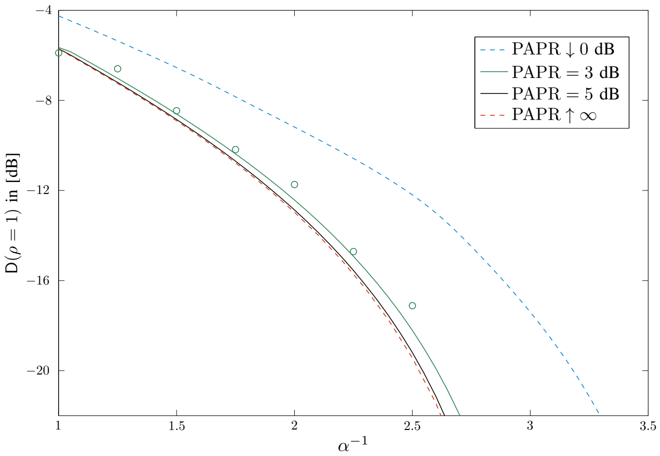


Fig. 2: Distortion at $\rho = 1$ vs. α^{-1} for several PAPRs. $P = 0.5$ and $\eta = 1$. Solid lines and circles respectively denote the results for GLSE and GLSE-GAMP algorithm with $N = 64$ and $T = 20$.

performance of the GLSE scheme, even for a practically moderate number of antennas. For the PAPR-limited precoder in Section IV-B, the distortion at $\rho = 1$ has been plotted in terms of α^{-1} in Fig. 2. The curves have been sketched for multiple PAPR constraints. Similar to Fig. 1, solid lines correspond to the GLSE scheme and circles denote the simulation results for the GLSE-GAMP precoder with $N = 64$ and $T = 20$ for PAPR $= 3$ dB. Here, we have considered $P = 0.5$, and $P_{\max}$ is tuned via the proposed strategy assuming all the antennas being active. The figure depicts that by increasing the PAPR up to 5 dB, the performance of the precoder is sufficiently close to the case without PAPR restriction. This observation suggests for employing the GLSE-GAMP precoder, in order to reduce the transmit PAPR without any significant performance loss. In this case, low efficiency power amplifiers can be utilized which can significantly reduce the RF-cost.

Remark 2: It is known that the GAMP algorithm converges for i.i.d. Gaussian matrices [13], [14]. However, by deviating from this assumption, the algorithm may diverge. This issue was recently addressed in [15] via the Vector Approximate Message Passing (VAMP) algorithm. Consequently, for channel models with ill-conditioned matrices, one can develop a precoding algorithm based on the GLSE scheme by taking a same approach while employing VAMP.

VI. Conclusion

This paper has proposed a class of low complexity precoders based on the GLSE scheme using the GAMP algorithm. The numerical investigations have been consistent with the replica results for the GLSE scheme given in [1]–[4]. This consistency demonstrates that various implementational limitations in massive MIMO systems can be effectively overcome using some low-complexity, but effective, algorithms. As indicated in Remark 2, the GLSE-GAMP precoders may fail in converging for channel models with ill-conditioned channel matrices, and therefore, an alternative algorithm can be proposed via VAMP. The extension under VAMP is however skipped and left as a possible future work.

References

[1] A. Bereyhi, M. A. Sedaghat, S. Asaad, and R. Müller, “Nonlinear precoders for massive MIMO systems with general constraints,” *International ITG Workshop on Smart Antennas (WSA)*, 2017.

[2] M. A. Sedaghat, A. Bereyhi, and R. Müller, “Least Square Error Precoders for Massive MIMO with Signal Constraints: Fundamental Limits,” *IEEE Transactions on Wireless Communications*, 2017.

[3] M. A. Sedaghat, A. Bereyhi, and R. Müller, “A New Class of Nonlinear Precoders for Hardware Efficient Massive MIMO Systems,” *International Conference on Communications (ICC)*, 2017.

[4] A. Bereyhi, M. A. Sedaghat, and R. Müller, “Asymptotics of nonlinear LSE precoders with applications to transmit antenna selection,” *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[5] J. Hoydis, S. Ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” *IEEE Journal on selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, 2013.

[6] S. K. Mohammed and E. G. Larsson, “Per-antenna constant envelope precoding for large multi-user MIMO systems,” *IEEE Transactions on Communications*, vol. 61, no. 3, pp. 1059–1071, 2013.

[7] J.-C. Chen, “Low-papr precoding design for massive multiuser MIMO systems via Riemannian manifold optimization,” *IEEE Communications Letters*, vol. 21, no. 4, pp. 945–948, 2017.

[8] H. Li, L. Song, and M. Debbah, “Energy efficiency of large-scale multiple antenna systems with transmit antenna selection,” *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 638–647, 2014.

[9] S. Asaad, A. Bereyhi, R. R. Müller, and A. M. Rabiei, “Asymptotics of transmit antenna selection: Impact of multiple receive antennas,” *International Conference on Communications (ICC)*, 2017.

[10] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” *IEEE Int. Sym. on Inf. Theory (ISIT)*, 2011.

[11] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[12] A. Javanmard and A. Montanari, “State evolution for general approximate message passing algorithms, with applications to spatial coupling,” *Information and Inference*, p. iat004, 2013.

[13] S. Rangan, P. Schniter, and A. Fletcher, “On the convergence of approximate message passing with arbitrary matrices,”*IEEE International Symposium on Information Theory (ISIT)*, 2014.

[14] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, “Fixed points of generalized approximate message passing with arbitrary matrices,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7464–7474, 2016.

[15] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” *IEEE International Symp. on Inf. Theory (ISIT)*, 2017.

Multilevel Codes in Lattice-Reduction-Aided Equalization

Robert F.H. Fischer[1], Johannes B. Huber[2], Sebastian Stern[1], Paulus M. Guter[2]

[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: johannes.huber@fau.de

***Abstract*—The application of multilevel codes in lattice-reduction-aided equalization is considered, i.e., the decoding of multilevel codes when (Gaussian) integer linear combinations of the codewords in signal space are present. Typically, multilevel codes do not generate lattice codes, hence arbitrary integer linear combinations are not directly decodable. We show that this lattice property is not required, which relaxes the constraints on the component codes significantly. A generalized version of multistage decoding which incorporates a "carry correction" is proposed; it circumvents the lattice property and any integer linear combinations are decodable. Numerical simulations are given to cover the performance of the proposed method.**

I. Introduction

In the literature, low-complexity but well-performing approaches for the equalization in *multiple-input/multiple-output (MIMO) multi-user uplink scenarios* are discussed for more than one decade. *Lattice-reduction-aided* (LRA) techniques [16], [15] and the tightly related concept of *integer-forcing* (IF) receivers [10], [17] are of special interest. In both schemes, the main idea is not to decode the transmitted signals, but *integer linear combinations* thereof; LRA and IF receivers differ in the way the (residual) integer interference is handled, cf. [4].

Right from the start, IF schemes were proposed as coded schemes—a strong coupling between integer equalization and decoding/code constraints is present. In contrast, LRA schemes are usually treated uncoded as a combination with coded modulation schemes seems to be easier. Here, the actual demand is that the code is linear in signal space; any integer linear combination of codewords has to be a codeword. Lattice codes obviously fulfill this demand.

In this paper, we are interested in the combination of *multilevel codes (MLC)* and LRA equalization, because multilevel coding together with *multistage decoding (MSD)* is in principle a capacity-achieving strategy [14]. Typically, multilevel codes do not generate lattice codes [9] and are not linear. However, we show that this property is indeed not required. We study the relevant design criteria on the component codes and propose a generalized version of multistage decoding which incorporates a "*carry correction*" such that integer linear combinations are decodable.[1]

The paper is organized as follows: In Sec. II the system model is introduced. The consequences of integer linear combinations on multilevel codes are studied in Sec. III and a new decoding scheme with carry correction is proposed. Sec. IV presents numerical examples; the paper is briefly summarized in Sec. V.

II. System Model

Throughout the paper, we assume K non-cooperating (single-antenna) users k, $k = 1, \dots, K$, communicating their binary source symbols[2] $\mathfrak{q}_k \in \mathbb{F}_2$ to a central receiver with $N_\mathrm{R} \geq K$ antennas. To that end, the symbols are encoded and mapped to complex-valued transmit symbols x_k, drawn from some signal constellation $\mathcal{A}$ with variance σ_x^2.

Denoting the K-dimensional transmit vector as $\boldsymbol{x}$, the $N_\mathrm{R} \times K$ (flat-fading) channel matrix as $\boldsymbol{H}$, and the N_R-dimensional noise vector (with zero-mean Gaussian noise components with variance σ_n^2 per dimension) as $\boldsymbol{n}$, the receive vector $\boldsymbol{y}$ (complex baseband notation) reads as usual

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \,. \tag{1}$$

We assume joint processing of the N_R components of the receive vector $\boldsymbol{y}$. The common, practicable approach is to first perform some form of joint equalization, followed by individual decoding of the codes.

Lattice-reduction-aided and integer-forcing equalization are low-complexity, well-performing approaches. In both variants, the main idea is to factorize the channel matrix as

$$\boldsymbol{H} = \boldsymbol{W}\,\boldsymbol{Z} \,, \tag{2}$$

where $\boldsymbol{Z} \in \mathbb{G}^{K\times K}$, $\mathbb{G} = \mathbb{Z} + \mathrm{j}\mathbb{Z}$, is a full-rank (Gaussian) integer matrix. Then, only the non-integer part $\boldsymbol{W}$ is equalized (here via MMSE linear equalization) [3], [17], [4]. The different factorization criteria (lattice or dual-lattice approach, ZF or MMSE criterion), constraints (unimodular vs. full-rank), and algorithms (shortest basis problem vs. shortest independent vector problem) are irrelevant for the scope of the paper; for an overview see, e.g., [4].

Including the linear (ZF or MMSE) equalizer frontend, the remaining part of the (LRA/IF) receiver has to deal with

$$\boldsymbol{r} = \boldsymbol{Z}\boldsymbol{x} + \bar{\boldsymbol{n}} \overset{\text{def}}{=} \bar{\boldsymbol{x}} + \bar{\boldsymbol{n}} \,, \tag{3}$$

where $\bar{\boldsymbol{n}}$ is the effective disturbance after equalization—here, for brevity we assume that all (complex-valued) components have the same variance $\sigma_{\bar{n}}^2$. Fig. 1 shows the effective end-to-end channel model.

The main difference between LRA and IF equalization is how the integer interference is resolved. In LRA schemes, the

[1] Shortly after the submission of this paper we became aware of a similar work [1], however, treating only coding over ASK constellations.

[2] The notation distinguishes quantities over the complex numbers (typeset as $\boldsymbol{x}$, $\boldsymbol{Z}$, ...), and over finite fields (typeset in Fraktur font; $\mathfrak{q}$, $\mathfrak{c}$, $\mathfrak{Z}_0$, ...).

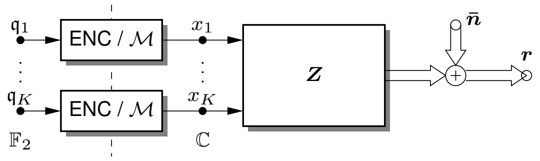


Fig. 1. End-to-end integer channel model (including linear equalization).

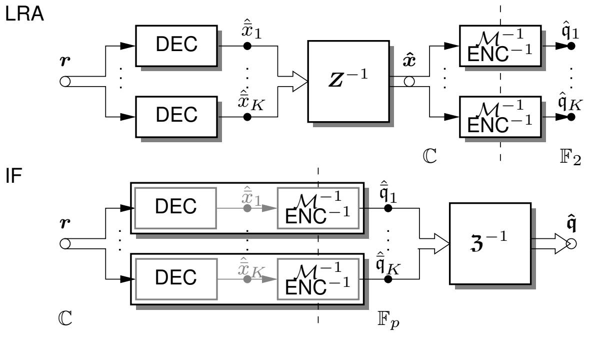


Fig. 2. System model of the receiver over end-to-end integer channels. Top: lattice-reduction-aided equalization; Bottom: integer-forcing receiver.

linear combinations $\bar{\boldsymbol{x}}$ of the transmit symbols in signal space are estimated by the decoders and the interference is undone via $\boldsymbol{Z}^{-1}$ over the complex numbers, cf. Fig. 2, top. Finally, via the encoder inverses, estimates $\hat{\mathfrak{q}}_k$ are produced.

In IF schemes (Fig. 2, bottom), the source symbols are drawn from a finite field $\mathbb{F}_p$, p a prime, and encoding is done over $\mathbb{F}_p$. At the receiver, linear combinations $\bar{\mathfrak{q}}_k$ of the source symbols are delivered by the decoders and the integer matrix $\mathfrak{Z}$ (the $\mathbb{F}_p$-equivalent to $\boldsymbol{Z}$) is inverted over $\mathbb{F}_p$.

These different orders of encoder inverse and inverse of $\boldsymbol{Z}$ impose different constraints on the codes. In LRA schemes, integer linear combinations in signal space have to be decodable. When lattice codes are used, this is directly fulfilled.

III. Multilevel Codes and Integer Interference

A particular strategy to coded modulation is *multilevel coding* [7], [14]. Via a set of *binary component codes* $\mathfrak{C}_l$, $l = 0, \ldots, m-1$, and a mapping from binary address labels of $m = \log_2(M)$ digits to M signal points (constituent constellation), a code in signal space is generated.

A. Mapping

We are interested in one-dimensional (amplitude-shift keying, ASK) and (square) two-dimensional (quadrature-amplitude modulation, QAM) constituent constellations and restrict ourselves to mapping according to the *set partitioning rule* [13]. In both cases, the mapping can be written as [5]

$$\mathcal{M}(\mathfrak{b}_{m-1} \ldots \mathfrak{b}_1 \mathfrak{b}_0) = \operatorname{mod}_B \left(\sum\nolimits_{l=0}^{m-1} \psi(\mathfrak{b}_l) \phi^l \right) - O \,, \quad (4)$$

where $\psi(\cdot)$ is the common mapping from the finite field ($\mathbb{F}_2$) elements "0" and "1" to integers "0" and "1", i.e., $\psi(0) = 0$ and $\psi(1) = 1$. $B = M$ for ASK and $B = \sqrt{M}$ for QAM defines the boundary region[3] and $O = \frac{M-1}{2}$ for ASK and $O = (1+\mathrm{j})\frac{\sqrt{M}-1}{2}$ for QAM is the offset for zero-mean constellations. Ignoring the modulo reduction and the offset, the point in signal space is given by its binary expansion w.r.t. the *base* ϕ. For ASK we have $\phi = 2$ and the usual binary expansion of an integer is present. For QAM we choose $\phi = -1 + \mathrm{j}$. This is due to the fact that the Gaussian integers $\mathbb{G}$ can be uniquely given via a binary representation where the base (radix) is the complex number $\pm 1 \pm \mathrm{j}$ [11], [6], cf. also [5].

B. Multilevel Codes and Lattices

Using this mapping, the multilevel code is defined by

$$\mathcal{C}_{\mathrm{MLC}} = \operatorname{mod}_B \left(\sum\nolimits_{l=0}^{m-1} \psi(\mathfrak{C}_l) \phi^l \right) - O \,, \quad (5)$$

where $\psi(\cdot)$, $\operatorname{mod}(\cdot)$, and the offset O are applied componentwise and $\mathfrak{C}_l$, $l = 0, \ldots, m-1$, are the binary component codes (at level l). Here, to simplify exposition, they are assumed to be block codes of equal length N.

Such a multilevel code is never a lattice. However, we can consider the following construction ("construction by code formula" [9], [5])

$$\mathcal{C} = \sum\nolimits_{l=0}^{m-1} \psi(\mathfrak{C}_l) \phi^l + \phi^m \mathbb{G}^N \,, \quad (6)$$

which differs from $\mathcal{C}_{\mathrm{MLC}}$ by i) eliminating the offset O and ii) inherently assuming an infinite number of "uncoded levels" via the addition of $\phi^m \mathbb{G}^N$ ($\phi^m \mathbb{Z}^N$ in case of ASK), i.e., periodic extension by adding integer multiples of ϕ^m to the coordinates of the codewords (note that $|\phi^m| = B$ is the size of the support per dimension of the multilevel code).

In [9] it is shown that $\mathcal{C}$ is a (real) lattice, i.e., has group structure under ordinary vector addition (over $\mathbb{R}$), if

a) $\mathfrak{C}_l$ are linear codes

b) $\mathfrak{C}_0 \subseteq \mathfrak{C}_1 \subseteq \cdots \subseteq \mathfrak{C}_{m-1}$ (7)

c) for all l and $\mathfrak{c}_i, \mathfrak{c}_j \in \mathfrak{C}_l$ it has to hold $\mathfrak{c}_i \odot \mathfrak{c}_j \in \mathfrak{C}_{l+1}$

where $\odot$ denotes the element-wise multiplication (in the arithmetic of $\mathbb{F}_2$, i.e., AND operation) of the words $\mathfrak{c}_i$ and $\mathfrak{c}_j$. The intuition behind the third demand is that the addition of two mapped words in real space (ASK) can be written as [9] ($\oplus$: element-wise addition in the arithmetic of $\mathbb{F}_2$)

$$\psi(\mathfrak{c}_i) + \psi(\mathfrak{c}_j) = \psi(\mathfrak{c}_i \oplus \mathfrak{c}_j) + 2\psi(\mathfrak{c}_i \odot \mathfrak{c}_j) \,, \quad (8)$$

where $\mathfrak{c}_i \odot \mathfrak{c}_j$ is the "carry" since[4] $1 + 1 = 2 = [1\,0]_2$.

For QAM signaling we have to resort to the arithmetic w.r.t. the base $-1 + \mathrm{j}$. Here, $2 = [1\,1\,0\,0]_{-1+\mathrm{j}}$ which gives

$$\psi(\mathfrak{c}_i) + \psi(\mathfrak{c}_j) = \psi(\mathfrak{c}_i \oplus \mathfrak{c}_j) + (\phi^2 + \phi^3)\psi(\mathfrak{c}_i \odot \mathfrak{c}_j) \,, \quad (9)$$

i.e., the "carry" is 110 [6] and $\mathfrak{c}_i \odot \mathfrak{c}_j \in \mathfrak{C}_{l+2} \subseteq \mathfrak{C}_{l+3}$ is the respective constraint ((7)c) is sufficient but not required).

[3] $\operatorname{mod}_B(x) \in 0, \ldots, B-1$ if x is real-valued; mod is applied separately to real and imaginary part when x is complex-valued.

[4] The sum of two codewords in signal space is only a valid codeword if the sum of two codewords in Hamming space is also in the code (linear codes) and if the "carry" is in the code of the next higher level.

If in ASK only the lowest level is coded (identical to construction A [2]), Constraints b) and c) are automatically fulfilled and a lattice is present. However, the (gross) coding gain is then limited to 6 dB by the first uncoded level. In QAM the levels are spaced only by $|\phi| = \sqrt{2}$ (3 dB) and not 2 as in ASK. Here, no carry to the next level but only the second and third next level is caused. Hence, the lower two levels can be coded and still all constraints are fulfilled and a lattice is present. In this case, the (gross) coding gain is also limited to 6 dB.

C. Multilevel Codes and Integer Linear Combinations

We are now interested in (Gaussian) integer linear combinations of multilevel codewords. To this end, we consider $\mathcal{C}$ from (6), i.e., ignore the boundary and the offset.

Let the MLC codewords $\boldsymbol{c}^{(k)}$ in signal space be obtained via the m codewords $\mathfrak{c}_l^{(k)}$, $l = 0, \dots, m-1$, $k = 1, 2, \dots$, in Hamming space, i.e.,

$$\boldsymbol{c}^{(k)} = \sum\nolimits_{l=0}^{m-1} \psi(\mathfrak{c}_l^{(k)}) \phi^l \ . \tag{10}$$

The action of integer linear combinations of these codewords is determined by the binary expansion (w.r.t. the base ϕ, $\mathfrak{z}_i^{(k)} \in \mathbb{F}_2$) of the respective (Gaussian) integer z_k

$$z_k = \left[\dots \mathfrak{z}_2^{(k)} \, \mathfrak{z}_1^{(k)} \, \mathfrak{z}_0^{(k)} \right]_\phi = \sum\nolimits_{i \geq 0} \psi(\mathfrak{z}_i^{(k)}) \, \phi^i \ . \tag{11}$$

Using (10) and (11) we have ($\psi(\mathfrak{a})\psi(\mathfrak{b}) = \psi(\mathfrak{ab})$, $\mathfrak{a}, \mathfrak{b} \in \mathbb{F}_2$)

$$\begin{aligned} \sum\nolimits_k z_k \boldsymbol{c}^{(k)} &= \sum\nolimits_k z_k \sum\nolimits_{l=0}^{m-1} \psi(\mathfrak{c}_l^{(k)}) \, \phi^l \\ &= \sum\nolimits_{l=0}^{m-1} \sum\nolimits_{i \geq 0} \sum\nolimits_k \psi(\mathfrak{z}_i^{(k)}) \, \phi^i \, \psi(\mathfrak{c}_l^{(k)}) \, \phi^l \\ &= \sum\nolimits_{l=0}^{m-1} \sum\nolimits_{i \geq 0} \sum\nolimits_k \psi(\mathfrak{z}_i^{(k)} \mathfrak{c}_l^{(k)}) \, \phi^{i+l} \\ &= \sum\nolimits_\ell \psi(\mathfrak{c}_\ell^{\text{eff}}) \, \phi^\ell \ , \end{aligned} \tag{12}$$

where $\mathfrak{c}_\ell^{\text{eff}}$ is the effective codeword at level ℓ. These words can be calculated by applying (8) or (9) (ASK or QAM) recursively. For example, for two codewords and QAM signaling we have

$$\begin{aligned} \mathfrak{c}_0^{\text{eff}} &= \mathfrak{z}_0^{(1)} \mathfrak{c}_0^{(1)} \oplus \mathfrak{z}_0^{(2)} \mathfrak{c}_0^{(2)} \\ \mathfrak{c}_1^{\text{eff}} &= \mathfrak{z}_0^{(1)} \mathfrak{c}_1^{(1)} \oplus \mathfrak{z}_0^{(2)} \mathfrak{c}_1^{(2)} \oplus \mathfrak{z}_1^{(1)} \mathfrak{c}_0^{(1)} \oplus \mathfrak{z}_1^{(2)} \mathfrak{c}_0^{(2)} \\ \mathfrak{c}_2^{\text{eff}} &= \mathfrak{z}_0^{(1)} \mathfrak{c}_2^{(1)} \oplus \mathfrak{z}_0^{(2)} \mathfrak{c}_2^{(2)} \oplus \mathfrak{z}_1^{(1)} \mathfrak{c}_1^{(1)} \oplus \mathfrak{z}_1^{(2)} \mathfrak{c}_1^{(2)} \\ &\qquad \oplus \mathfrak{z}_2^{(1)} \mathfrak{c}_0^{(1)} \oplus \mathfrak{z}_2^{(2)} \mathfrak{c}_0^{(2)} \oplus \mathfrak{z}_0^{(1)} \mathfrak{c}_0^{(1)} \odot \mathfrak{z}_0^{(2)} \mathfrak{c}_0^{(2)} \\ &\vdots \end{aligned} \tag{13}$$

Having a look at the appearing terms and keeping in mind that $\mathfrak{c}_\ell^{\text{eff}}$ has to be a valid codeword of code $\mathfrak{C}_\ell$ the three above demands (7) on the codes are directly clear.

D. Multistage Decoding

Usually, multilevel codes are decoded via *multistage decoding (MSD)*, cf. Alg. 1. Thereby, given $\boldsymbol{r} = \boldsymbol{c} + \boldsymbol{n}$ with $\boldsymbol{c} = \sum_{l=0}^{m-1} \psi(\mathfrak{c}_l)\phi^l \in \mathcal{C}$, one component code (estimate $\hat{\mathfrak{c}}_l$) after the other is decoded, taking into account the decoding results of lower levels but ignoring the codes of higher levels (i.e., treating them as uncoded). Each decoding step is hence identical to decoding a lattice constructed via lattice construction A [2] (multilevel code where only the lowest level is coded). Thus, in a multistage decoder it is immaterial whether the entire code forms (a translate of a subset of) a lattice.

Alg. 1 Multistage Decoding.

```
function $\hat{\boldsymbol{c}} = \texttt{MSD}(\boldsymbol{r})$
  $\ell = 0$; $\boldsymbol{r}_\ell = \boldsymbol{r}$                                  // init
  while $\ell < m$ {
     $\hat{\mathfrak{c}}_\ell = \text{DEC}_{\mathfrak{C}_\ell}\{\boldsymbol{r}_\ell\}$                       // decode level $\ell$
     $\boldsymbol{r}_{\ell+1} = (\boldsymbol{r}_\ell - \psi(\hat{\mathfrak{c}}_\ell))/\phi$              // eliminate known level
     $\ell = \ell + 1$ }
  $\hat{\boldsymbol{c}} = \sum_{l=0}^{m-1} \psi(\hat{\mathfrak{c}}_l)\phi^l$                          // codeword estimate
```

However, it is essential that the effective word at level ℓ is a valid codeword from $\mathfrak{C}_\ell$. If integer linear combinations of codewords are to be decoded and Constraint c) of (7) is violated, carries from lower levels destroy this property and MSD does not work any more.

E. Carry Correction

In the following, we assume that Constraints a) and b) on the component codes are fulfilled but Constraint c) is violated. We present a generalization of multistage decoding which decodes the K linear combinations and eliminates the effect of the carries.

Let $\boldsymbol{c}^{(1)}$, $\boldsymbol{c}^{(2)}$, …, $\boldsymbol{c}^{(K)}$ be the codewords of the K users in signal space (component codewords $\mathfrak{c}_0^{(k)}$, $\mathfrak{c}_1^{(k)}$, …, $\mathfrak{c}_{m-1}^{(k)}$, $k = 1, \dots, K$, in Hamming space) and $\boldsymbol{Z}$ the integer matrix. Ignoring the noise for the moment we have

$$\begin{bmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_K \end{bmatrix} = \boldsymbol{Z} \begin{bmatrix} \boldsymbol{c}^{(1)} \\ \vdots \\ \boldsymbol{c}^{(K)} \end{bmatrix} \ . \tag{14}$$

From the discussion above (cf. (12) and (13)), we see that the effective codewords at level $l = 0$ are obtained from the component codewords at this levels via

$$\begin{bmatrix} \mathfrak{c}_{1,0}^{\text{eff}} \\ \vdots \\ \mathfrak{c}_{K,0}^{\text{eff}} \end{bmatrix} = \mathfrak{Z}_0 \begin{bmatrix} \mathfrak{c}_0^{(1)} \\ \vdots \\ \mathfrak{c}_0^{(K)} \end{bmatrix} \tag{15}$$

where $\mathfrak{z}_0^{(i,j)}$ is the least significant bit (LSB) of $z_{i,j}$ w.r.t. to the basis ϕ and $\mathfrak{Z}_0 = [\mathfrak{z}_0^{(i,j)}]$. Having estimates for $\mathfrak{c}_{k,0}^{\text{eff}}$, $k = 1, \dots, K$, the original codewords at level 0 can hence be obtained by solving (over $\mathbb{F}_2$) the set of linear equations (15).

Once the codewords at level 0 are known, the contributions (over $\mathbb{C}$) of the superposition of these levels into the higher levels (carries) can be calculated via[5]

$$\begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_K \end{bmatrix} = \mathcal{Q}_\phi \left\{ \boldsymbol{Z} \begin{bmatrix} \psi(\mathfrak{c}_0^{(1)}) \\ \vdots \\ \psi(\mathfrak{c}_0^{(K)}) \end{bmatrix} \right\} \ . \tag{16}$$

[5] $\mathcal{Q}_\phi\{z\}$ denotes the quantization operation which nulls the LSB in the binary expansion w.r.t. ϕ, i.e., with $z = [\dots \mathfrak{z}_2 \, \mathfrak{z}_1 \, \mathfrak{z}_0]_\phi$ we obtain $\mathcal{Q}_\phi\{z\} = [\dots \mathfrak{z}_2 \, \mathfrak{z}_1 \, 0]_\phi$. This nulling is required since only the contribution to the higher levels but not the current level is of interest.

Alg. 2 Multistage Decoding with Carry Correction.

```
function $[\hat{c}^{(1)}, \dots, \hat{c}^{(K)}] = \mathtt{MSD}(\boldsymbol{r})$
    $\ell = 0$; $\boldsymbol{r}_{k,\ell} = \boldsymbol{r}_k$, $k = 1, \dots, K$        // init
    while $\ell < m$ {
        $\hat{\mathfrak{c}}_{k,\ell} = \mathrm{DEC}_{\mathfrak{C}_\ell}\{\boldsymbol{r}_{k,\ell}\}$, $k = 1, \dots, K$        // decode level $\ell$
        solve (15) and calculate $\boldsymbol{s}_k$ via (16)        // calculate carries
        $\boldsymbol{r}_{k,\ell+1} = (\boldsymbol{r}_{k,\ell} - \psi(\hat{\mathfrak{c}}_{k,\ell}) - \boldsymbol{s}_{k,\ell})/\phi$        // eliminate known interference
        $\ell = \ell + 1$ }
    $\hat{c}^{(k)} = \sum_{l=0}^{m-1} \psi(\hat{\mathfrak{c}}_{k,l})\phi^l$        // codeword estimates
```

This known "interference" caused by the codewords at level 0 can now be eliminated. Hence, in a generalized version of MSD, i) the influence of the effective codeword at level 0 is eliminated by subtraction of $\psi(\mathfrak{c}_{i,0}^{\mathrm{eff}})$ and ii) the influence (carries) of the linear combinations of the codewords onto higher levels is eliminated by subtraction of $\boldsymbol{s}_k$. This procedure is iterated over the levels. In Alg. 2, a pseudo-code description of this generalized version of multistage decoding is given.

The additional complexity for the calculation of the correction terms $\boldsymbol{s}_k$ is negligible compared to the decoding operations.

F. Conditions on $\boldsymbol{Z}$

The final question is under which conditions does the carry correction work. Since the effective codewords are generated via the binary matrix $\mathfrak{Z}_0$ (the matrix of LSBs of the (Gaussian) integer entries of the matrix $\boldsymbol{Z}$), this matrix has to have full rank.[6]

In order that $\mathfrak{Z}_0$ has full rank, its determinant (over $\mathbb{F}_2$) has to be non-zero. We have $\psi(\det(\mathfrak{Z}_0)) = \mathrm{mod}_2(\det(\psi(\mathfrak{Z}_0)))$ and using the Leibniz formula for the expansion of the determinant, we obtain $\det(\psi(\mathfrak{Z}_0)) = \det(\boldsymbol{Z}) + \phi\mathbb{G}$.

If $\det(\boldsymbol{Z}) \in \mathbb{G}$ is "odd" (LSB w.r.t. base ϕ equal to one) then $\det(\psi(\mathfrak{Z}_0))$ is also odd and hence non-zero. Unimodular matrices have an "odd" determinant. If $\det(\boldsymbol{Z})$ is "even" (LSB equal to zero) then $\det(\psi(\mathfrak{Z}_0))$ is also even and thus $\det(\mathfrak{Z}_0) = 0$ and $\mathfrak{Z}_0$ is not invertible.

Thus it is proven that for all matrices for which $\det(\boldsymbol{Z})$ is "odd" ($\in 1 + \phi\mathbb{G}$), hence in particular for unimodular matrices, carry correction works. This imposes restrictions on $\boldsymbol{Z}$. However, it has been shown that using a suited lattice reduction algorithm (Minkowski reduction) for i.i.d. Gaussian channel matrices even the restriction to unimodular matrices causes almost no loss compared to the set of full-rank matrices [12].

IV. Numerical Results

In oder to study the effect of integer linear combinations and carry correction, numerical simulations have been conducted. We assume a 16 QAM constellation; the code has four levels.

First, a "toy example" is presented. Linear block codes of length $N = 13$ are employed, the generator matrices of the $(13, 2, 8)$ code at level 0 and that of the $(13, 4, 4)$ code level 1 are

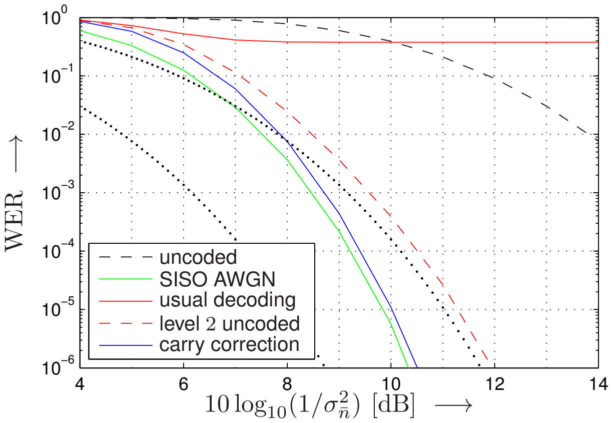


Fig. 3. Word error rate over the inverse noise power (in dB). 16 QAM. Component codes: linear block codes of length $N = 13$. Rate 31/13 bit/symbol. $K = 3$ users. Fixed integer matrix $\boldsymbol{Z} = [2{-}2\mathrm{j}\ \ 1\ \ {-2}{-}\mathrm{j};\ \ 1\ \ {-1}{-}\mathrm{j}\ \ 0;\ \ {-1}\ \ \mathrm{j}\ \ 0]$; $\det(\boldsymbol{Z}) = 2 + \mathrm{j}$. Dotted: asymptotic behavior.

$$\mathfrak{G}_0 = \begin{bmatrix} 1111111100000 \\ 0000011111111 \end{bmatrix}, \quad \mathfrak{G}_1 = \begin{bmatrix} 1111111100000 \\ 1100000111111 \\ 0011111100011 \\ 0000011111111 \end{bmatrix}. \quad (17)$$

At level 2 a $(13, 12, 2)$ single-parity-check code is used and level 3 is uncoded ($(13, 13, 1)$ code). This construction does not fulfill Condition c) of (7), hence no lattice is present and carry correction is required for correct decoding.

Fig. 3 shows the results (word error rate averaged over the users) for $K = 3$ and the arbitrarily selected integer matrix $\boldsymbol{Z} = [2{-}2\mathrm{j}\ \ 1\ \ {-2}{-}\mathrm{j};\ 1\ \ {-1}{-}\mathrm{j}\ \ 0;\ \ {-1}\ \ \mathrm{j}\ \ 0]$ with $\det(\boldsymbol{Z}) = 2 + \mathrm{j}$. Uncoded transmission and coded transmission over the single-input/single-output (SISO) AWGN channel are shown for reference. The code has an asymptotic (gross) coding gain of 9 dB (dotted; curve for uncoded transmission shifted by the asymptotic gain). If no carry correction is applied, decoding of the multilevel code fails completely.[7] With the proposed carry correction the performance of the code over the SISO channel can almost be achieved (some error multiplication is present). If level 2 is (treated as) uncoded, Constraint c) is fulfilled and no carry correction is required. However, the (asymptotic, gross) coding gain is then limited to 6 dB. Via carry correction we can break this 6 dB barrier.

Next, a multilevel code with low-density parity-check (LDPC) codes as component codes is considered. The rates of the codes are adjusted according to the capacity design rule for multilevel codes [14]. For a target rate of 3 bits per QAM symbol, the rates of the component codes have to be selected as $R_0/R_1/R_2/R_3 = .282/.753/.964/1$. We use a code length $N = 5000$, which gives code dimensions $\kappa_0/\kappa_1/\kappa_2/\kappa_3 = 1412/3766/4820/5000$.

[6] If $\mathfrak{Z}_0$ has columns which are the all-zero vector, the corresponding codewords do not influence the effective codewords and do not contribute to carry generation. Hence, only the relevant part (matrix $\mathfrak{Z}_0$ with all-zero columns removed) has to have full rank.

[7] Note that the component decoders always return a valid codeword. Since, due to the carry, the effective word at level 2 is not a valid codeword, wrong correction occurs, leading to a high floor in the error rate.

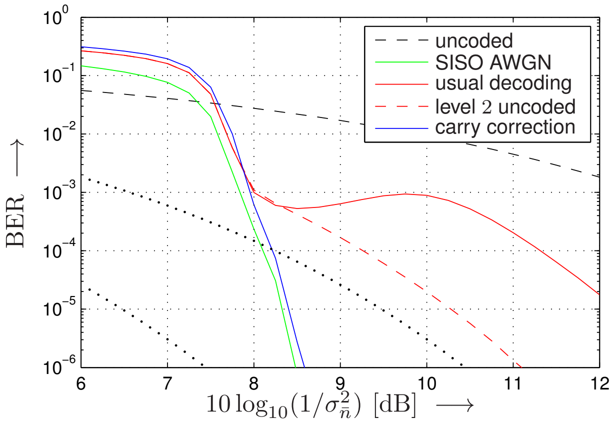


Fig. 4. Bit error rate over the inverse noise power (in dB). 16 QAM. Component codes: LDPC codes of length $N = 5000$. Rate 3 bit/symbol. $K = 3$ users. Fixed integer matrix $\boldsymbol{Z} = [1\ \ \mathrm{j}\ \ -1{+}\mathrm{j};\ \ 1\ \ 0\ \ -2\mathrm{j};\ \ 2\mathrm{j}\ \ -1\ \ -\mathrm{j}]$; $\det(\boldsymbol{Z}) = \mathrm{j}$. Dotted: asymptotic behavior.

As LDPC codes *irregular repeat-accumulate codes* [8] are used. The required subcode property $\mathfrak{C}_0 \subset \mathfrak{C}_1 \subset \mathfrak{C}_2 \subset \mathfrak{C}_3 = \mathbb{F}_2^N$ is guaranteed by constructing the parity-check matrices $\mathfrak{H}_l$ in such a way that $\mathfrak{H}_{l+1}$ is a subset of the rows of $\mathfrak{H}_l$. To that end, first the $(N - \kappa_2) \times N$ matrix $\mathfrak{H}_2$ is constructed with an ultra-sparse left part (column weight 2) and a right staircase part. This matrix is extended to $\mathfrak{H}_1$ by adding $\kappa_2 - \kappa_1$ rows in such a way that the newly added left part is ultra-sparse and that the staircase construction is seamless continued in the right part. Given $\mathfrak{H}_1$, the final parity-check matrix $\mathfrak{H}_0$ is constructed in the same way by adding $\kappa_1 - \kappa_0$ rows. Given the parity-check matrices $\mathfrak{H}_l$, generator matrices $\mathfrak{G}_l$ for systematic encoding are calculated. Noteworthy, this construction guarantees the subcode property (Constraint b)) of the linear (Constraint a)) component codes. However, Constraint c) is neither taken into account nor fulfilled.

Fig. 4 shows the error rate of the information bits over the noise level (in dB).[8] Here, $\boldsymbol{Z} = [1\ \ \mathrm{j}\ \ -1{+}\mathrm{j};\ \ 1\ \ 0\ \ -2\mathrm{j};\ \ 2\mathrm{j}\ \ -1\ \ -\mathrm{j}]$ with $\det(\boldsymbol{Z}) = \mathrm{j}$ is chosen randomly. Message-passing decoders using log-likelihood ratios based on nearest-neighbor approximation are used; at maximum 10 iterations are performed. Basically, the same behavior as in the example above is visible. The exact shape of the flattening in case of conventional (independent) multistage decoding depends on the operation of the decoder in case of non-converges (here: the current variable-node values are quantized and output). If level 2 is treaded as uncoded, the (gross) coding gain is again limited to 6 dB. Via the proposed generalized decoding scheme, a performance extremely close to that of the code over the SISO AWGN channel can be achieved.

V. Summary and Conclusions

In this paper, we have studied the application of multilevel codes in LRA equalization, i.e., the decoding of multilevel codes when (Gaussian) integer linear combinations of the codewords are present. A generalized version of multistage decoding incorporating a carry correction has been proposed which has only marginal additional complexity compared to independent decoding. Thereby, the constraints on the component codes are significantly relaxed and no lattice code has to be generated. Numerical results for arbitrary but fixed integer matrices have been given.

The next step is the performance evaluation over random (i.i.d. Gaussian) channel matrices. The factorization algorithm in [4] can straightforwardly be generalized such that only $\boldsymbol{Z}$ matrices with "odd" determinant are returned. However, almost no loss occurs if Minkowski reduction with its restriction to unimodular matrices is used. In summary, via the new decoding procedure, a simple combination of multilevel codes with low-complexity LRA equalization is enabled.

References

[1] S H. Chae, M. Jang, S.K. Ahn, J. Park, C. Jeong. Multilevel Coding Scheme for Integer-Forcing MIMO Receivers With Binary Codes. *IEEE Trans. Wireless Comm.*, vol. 16, no. 8, pp. 5428–5441, Aug. 2017.

[2] J.H. Conway, N.J.A. Sloane. *Sphere Packings, Lattices and Groups*. Springer Verlag, New York, Berlin, 3rd edition, 1999.

[3] 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*, vol. 65, no. 8, pp. 688–693, Aug. 2011.

[4] R.F.H. Fischer, M. Cyran, S. Stern. Factorization Approaches in Lattice-Reduction-Aided and Integer-Forcing Equalization. In *2016 Int. Zurich Seminar on Communications*, Zurich, Switzerland, March 2016.

[5] G.D. Forney. Coset Codes. I. Introduction and Geometrical Classification. *IEEE Trans. Information Theory*, vol. 34, no. 5, pp. 1123–1151, Sep. 1988.

[6] W.J. Gilbert. Arithmetic in Complex Bases. *Mathematics Magazine*, vol. 57, No. 2, pp. 77–81, March 1984.

[7] H. Imai and S. Hirakawa. A New Multilevel Coding Method Using Error Correcting Codes. *IEEE Trans. Information Theory*, vol. 23, no. 3, pp. 371–377, May 1977.

[8] H. Jin, A. Khandekar, R. McEliece. Irregular Repeat-Accumulate Codes. In *2nd Int. Symp. on Turbo Codes and Rel. Topics*, Brest, France, 2000.

[9] W. Kositwattanarerk, F. Oggier. Connections Between Construction D and Related Constructions of Lattices. *Designs, Codes and Cryptography*, vol. 73, no. 2, pp. 441–455, Nov. 2014.

[10] B. Nazer, M. Gastpar. Compute-and-Forward: Harnessing Interference Through Structured Codes. *IEEE Trans. Information Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[11] W. Penney. A "Binary" System for Complex Numbers. *Journal of the Association for Computing Machinery (JACM)*, vol. 12, no. 2, pp. 247–248, April 1965.

[12] S. Stern, R.F.H. Fischer Optimal Factorization in Lattice-Reduction-Aided and Integer-Forcing Linear Equalization. In *11. Int. ITG Conf. on Systems, Communications, and Coding (SCC)*, Hamburg, Germany, February 2017

[13] G. Ungerböck. Channel Coding with Multilevel/Phase Signals. *IEEE Trans. Information Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[14] U. Wachsmann, R.F.H. Fischer, J.B. Huber. Multilevel Codes: Theoretical Concepts and Practical Design Rules. *IEEE Trans. Information Theory*, vol. 45, no. 5, pp. 1361–1391, Jul 1999.

[15] C. Windpassinger, R.F.H. Fischer. Low-Complexity Near-Maximum-Likelihood Detection and Precoding for MIMO Systems using Lattice Reduction. In *IEEE Information Theory Workshop*, pp. 345–348, Paris, France, March/April 2003.

[16] H. Yao, G. Wornell. Lattice-Reduction-Aided Detectors for MIMO Communication Systems. In *IEEE Global Telecommunications Conference*, pp. 424–428, Taipei, Taiwan, Nov. 2002.

[17] J. Zhan, B. Nazer, U. Erez, M. Gastpar. Integer-Forcing Linear Receivers. *IEEE Trans. Information Theory*, vol. 60, no. 12, pp. 7661–7685, Dec. 2014.

[8] Please note the different scaling of the x-axis compared to Fig. 3.

Approximate ML Detection for MIMO Phase Noise Channels

Richard Combes and Sheng Yang
CentraleSupélec, L2S
91190 Gif sur Yvette, France
Email: {richard.combes, sheng.yang}@l2s.centralesupelec.fr

***Abstract*—We consider the MIMO communication channel impaired by phase noises at the transmitter and receiver. We focus on maximum likelihood detection for uncoded single-carrier transmission. We derive an approximation of the likelihood function, and propose the so-called *self-interference whitening* (SIW) algorithm. While the exact ML solution is computationally intractable, we construct a simulation-based lower bound on the error probability of ML detection. Numerical experiments demonstrate that SIW is, in most cases of interest, very close to optimal with moderate phase noise. Surprisingly, such near-ML performance can be achieved by applying only twice the nearest neighbor detection algorithm.**

I. Introduction

We consider the signal detection problem for the following discrete-time multiple-input multiple-output (MIMO) channel

$$\boldsymbol{y} = \mathrm{diag}\big(e^{j\theta_{\mathrm{r},1}}, \dots, e^{j\theta_{\mathrm{r},n_\mathrm{r}}}\big)\boldsymbol{H}\mathrm{diag}\big(e^{j\theta_{\mathrm{t},1}}, \dots, e^{j\theta_{\mathrm{t},n_\mathrm{t}}}\big)\boldsymbol{x}+\boldsymbol{z}, \quad (1)$$

where $\boldsymbol{H} \in \mathbb{C}^{n_\mathrm{r}\times n_\mathrm{t}}$ is the channel matrix known to the receiver; $\boldsymbol{z} \in \mathbb{C}^{n_\mathrm{r}\times 1}$ represents a realization of the additive noise whereas $\theta_{\mathrm{t},l}$ and $\theta_{\mathrm{r},k}$ are the phase noises at the l th transmit antenna and the k th receive antenna, respectively; the input vector $\boldsymbol{x} \in \mathbb{C}^{n_\mathrm{t}\times 1}$ is assumed to be carved from a quadratic amplitude modulation (QAM). The goal is to estimate $\boldsymbol{x}$ from the observation $\boldsymbol{y} \in \mathbb{C}^{n_\mathrm{r}\times 1}$, with only statistical knowledge on the additive noise and the phase noises.

When phase noise is absent, the problem is well understood, and the maximum likelihood (ML) solution can be found using a nearest neighbor detection (NND) algorithm (see [1] and references therein). For instance, the sphere decoder [2] is an efficient NND with low expected complexity dimension with respect to the dimension n_t [3]. Further, there also exist approximate NND algorithms, e.g., based on lattice reduction, with near-ML performance when used for MIMO detection [4].

The presence of phase noise in (1) is a practical, long-standing problem in communication. In the seminal [5] back in the 70's, Foschini *et al.* used this model to capture the residual phase jitter at the phase-locked loop of the receiver side, and investigated both the receiver performance and the constellation design in the scalar case ($n_\mathrm{t} = n_\mathrm{r} = 1$). In fact, most communication systems feature phase noise due to the phase and frequency instabilities in the carrier frequency oscillators at both the transmitter and the receiver [6]. The channel (1) is a valid mathematical model when the phase noise varies slowly as compared to the symbol duration.[1] While phase noise can be practically ignored in conventional MIMO systems, its impact becomes prominent at higher carrier frequencies since it can be shown that phase noise power increases *quadratically* with carrier frequency [6], [9]. The performance degradation due to phase noise becomes even more severe with the use of higher order modulations for which the angular separation between constellation points can be small. At medium to high SNR, phase noise dominates additive noise, becoming the capacity bottleneck [10], [11]. As for signal detection, finding the ML solution for the MIMO phase noise channel (1) is hard in general. Indeed, unlike for conventional MIMO channels, the likelihood function of the transmitted signal cannot be obtained in closed form.

In this work, we propose an efficient MIMO detection algorithm which finds an approximate ML solution in the presence of phase noise. The main contributions of this work are summarized as follows. First we derive a tractable approximation of the likelihood function of the transmitted signal. While the exact likelihood does not have a closed-form expression, the proposed approximation has a simple form and turns out to be accurate for weak to medium phase noises. Then we propose a heuristic method that finds an approximate solution by applying twice the nearest neighbor detection algorithm. The proposed algorithm, called *self-interference whitening* (SIW), has a simple geometric interpretation. Intuitively, the phase noise perturbation generates self-interference that depends on the transmitted signal through the covariance matrix. The main idea is to first estimate the covariance of the self-interference with a potentially inaccurate initial signal solution, then perform the whitening with the estimated covariance, followed by a second detection. From the optimization point of view, our algorithm can be seen as a (well-chosen) concave approximation to a non-concave objective function. Finally we assess the performance of SIW and competing algorithms in different communication scenarios. Since the error probability of ML decoding is unknown, we propose a simulation-based lower bound which we use as a benchmark. Simulation re-

[1] As pointed out in [7] and the references therein, an effective discrete-time channel is usually obtained from a waveform phase noise channel after filtering. When the continuous-time phase noise varies rapidly during the symbol period, the filtered output also suffers from amplitude perturbation. See the full version of this paper [8] for further discussion.

sults show that SIW achieves near ML performance in most scenarios. In this sense, our work reveals that near optimal MIMO detection with phase noise can be done as efficiently as without phase noise.

II. Assumptions and Problem Formulation

Notation: For random quantities, we use upper case letters, e.g., X, for scalars, upper case letters with bold and non-italic fonts, e.g., $\mathbf{V}$, for vectors, and upper case letter with bold and sans serif fonts, e.g., $\boldsymbol{\mathsf{M}}$, for matrices. Deterministic quantities are denoted in a rather conventional way with italic letters, e.g., a scalar x, a vector $\boldsymbol{v}$, and a matrix $\boldsymbol{M}$. The Euclidean norm of a vector $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|$. The transpose and conjugated transpose of $\boldsymbol{M}$ are $\boldsymbol{M}^T$ and $\boldsymbol{M}^H$, respectively.

We assume a MIMO channel with n_t transmit and n_r receive antennas. Let $\boldsymbol{H}$ denote the channel matrix, where the (k,l)-th element of $\boldsymbol{H}$, denoted as $h_{k,l}$, represents the channel gain between the l th transmit antenna and k th receive antenna. The transmitted vector is denoted by $\boldsymbol{x} = [x_1, \dots, x_{n_\mathrm{t}}]^T$, where $x_l \in \mathcal{X}, l = 1, \dots, n_\mathrm{t}$, $\mathcal{X}$ being typically a QAM constellation with normalized average energy, i.e., $\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} |x|^2 = 1$. The base-band output is the following random vector

$$\mathbf{Y} = \boldsymbol{\mathsf{\Lambda}}_\mathrm{R} \boldsymbol{H} \boldsymbol{\mathsf{\Lambda}}_\mathrm{T} \boldsymbol{x} + \mathbf{Z},$$

where the diagonal matrices $\boldsymbol{\mathsf{\Lambda}}_\mathrm{R} := \operatorname{diag}\left(e^{j\Theta_{\mathrm{r},1}}, \dots, e^{j\Theta_{\mathrm{r},n_\mathrm{r}}}\right)$ and $\boldsymbol{\mathsf{\Lambda}}_\mathrm{T} := \operatorname{diag}\left(e^{j\Theta_{\mathrm{t},1}}, \dots, e^{j\Theta_{\mathrm{t},n_\mathrm{t}}}\right)$ capture the phase perturbation at the receiver and transmitter, respectively; $\mathbf{Z} \sim \mathcal{CN}(0, \gamma^{-1}\boldsymbol{I})$ is the additive white Gaussian noise (AWGN) vector, where γ is the nominal signal-to-noise ratio (SNR). The phase noise $\mathbf{\Theta} := [\Theta_{\mathrm{t},1} \ \cdots \ \Theta_{\mathrm{t},n_\mathrm{t}} \ \Theta_{\mathrm{r},1} \ \cdots \ \Theta_{\mathrm{r},n_\mathrm{r}}]^T$ is jointly Gaussian with $\mathbf{\Theta} \sim \mathcal{N}(0, \boldsymbol{Q}_\theta)$ where the covariance matrix $\boldsymbol{Q}_\theta$ can be arbitrary. This model includes as a special case the uplink channel in which n_t is the number of single-antenna users. In such a case, the transmit phase noises are independent. We consider uncoded transmission so that each symbol x_l can take any value from $\mathcal{X}$ with equal probability.

Further, we assume that the channel matrix can be random but is perfectly known at the receiver, whereas such knowledge at the transmitter side is irrelevant in uncoded transmission. We also define $\boldsymbol{\mathsf{H}}_\Theta := \boldsymbol{\mathsf{\Lambda}}_\mathrm{R} \boldsymbol{H} \boldsymbol{\mathsf{\Lambda}}_\mathrm{T}$ and accordingly $\boldsymbol{H}_\theta$ for some realization of $\mathbf{\Theta} = \boldsymbol{\theta}$, thus, $\boldsymbol{H}_0 = \boldsymbol{H}$. Finally, we ignore the temporal correlation of the phase noise process and the channel process, and focus on the spatial aspect of the problem.

With AWGN, we have the following conditional probability density function (pdf)

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{H}) = \frac{\gamma^{n_\mathrm{r}}}{\pi^{n_\mathrm{r}}} e^{-\gamma\|\boldsymbol{y} - \boldsymbol{H}_\theta \boldsymbol{x}\|^2},$$

and the likelihood function by integrating over $\mathbf{\Theta}$

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}, \boldsymbol{H}) = \ln\left(\mathbb{E}_{\mathbf{\Theta}}\left[e^{-\gamma\|\boldsymbol{y} - \boldsymbol{H}_{\mathbf{\Theta}} \boldsymbol{x}\|^2}\right]\right) + \ln\frac{\gamma^{n_\mathrm{r}}}{\pi^{n_\mathrm{r}}}.$$

The ML detector finds an input vector from the alphabet $\mathcal{X}^{n_\mathrm{t}}$ such that the likelihood function is maximized. In practice, it is often more convenient to use the log-likelihood function as the objective function, i.e., after removing a constant term,

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_\theta) := \ln\left(\mathbb{E}_{\mathbf{\Theta}}\left[e^{-\gamma\|\boldsymbol{y} - \boldsymbol{H}_{\mathbf{\Theta}} \boldsymbol{x}\|^2}\right]\right),$$

where the arguments γ and $\boldsymbol{Q}_\theta$ can be omitted whenever confusion is not likely. Thus,

$$\hat{\boldsymbol{x}}_\mathrm{ML}(\boldsymbol{y}, \boldsymbol{H}) := \arg\max_{\boldsymbol{x}\in\mathcal{X}^{n_\mathrm{t}}} f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{H}). \tag{2}$$

From (2) we see two main challenges to compute the optimal solution. First, the expectation in (2) cannot be obtained in closed form. A numerical implementation is equivalent to finding the numerical integral in $n_\mathrm{t} + n_\mathrm{r}$ dimensions. This can be extremely hard in high dimensions. Second, the size of the optimization space, $|\mathcal{X}|^{n_\mathrm{t}}$, can be prohibitively large when the modulation size $|\mathcal{X}|$ and the input dimension n_t become large. In the full paper [8], we examine in more details why both of these issues are indeed challenging.

In a conventional MIMO channel, finding the ML solution is reduced to solving the following problem

$$\hat{\boldsymbol{x}}^0_\mathrm{ML}(\boldsymbol{y}, \boldsymbol{H}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_\mathrm{t}}} \|\boldsymbol{y} - \boldsymbol{H}_0 \boldsymbol{x}\|^2, \tag{3}$$

which is also called the minimum Euclidean distance detection or nearest neighbor detection. Although the search space in (3) remains large, the expectation is gone. Furthermore, since the objective function is the Euclidean distance, efficient algorithms (e.g., sphere decoder [2] or lattice decoder [1]) exploiting the geometric structure of the problem can be applied without searching over the whole space $\mathcal{X}^{n_\mathrm{t}}$. Indeed, the sphere decoder has a polynomial average complexity with respect to the input dimension n_t when the channel matrix is drawn i.i.d. from a Rayleigh distribution [3].

In practice, one may simply ignore the existence of phase noise and still apply (3) to obtain $\hat{\boldsymbol{x}}^0_\mathrm{ML}$ which we refer to as the *naive ML* solution hereafter. While this can work relatively well when the phase noise is close to 0, it becomes highly suboptimal with stronger phase noise which is usually the case in high frequency bands with imperfect oscillators. In this paper, we provide a near ML solution by circumventing the two challenges mentioned earlier. We first propose an approximation of the likelihood function. Then we propose an algorithm to solve approximately the optimization problem (2).

III. Proposed Scheme

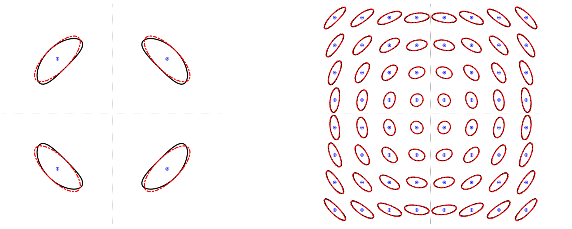
A. Proposed Approximation of the Likelihood Function

We derive an approximation of the likelihood when the phase noise is small. Indeed, in practice, the standard deviation of the phase noise is typically smaller than 10 degrees ≈ 0.174 rad. For stronger phase noises, it is not reasonable to use QAM and the problem should be addressed differently. Consider the following approximation:

$$\begin{aligned}\boldsymbol{\mathsf{\Lambda}}_\mathrm{R}^H \boldsymbol{y} - \boldsymbol{H}\boldsymbol{\mathsf{\Lambda}}_\mathrm{T}\boldsymbol{x} &= [-\boldsymbol{H}\boldsymbol{D}_x \ \ \boldsymbol{D}_y] \begin{bmatrix} e^{j\boldsymbol{\theta}_t} \\ e^{-j\boldsymbol{\theta}_r} \end{bmatrix} \\ &\approx (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}) - j[\boldsymbol{H}\boldsymbol{D}_x \ \ \boldsymbol{D}_y]\boldsymbol{\theta}, \end{aligned} \tag{4}$$

with $\boldsymbol{D}_x := \operatorname{diag}(x_1, \dots, x_{n_\mathrm{t}})$, $\boldsymbol{D}_y := \operatorname{diag}(y_1, \dots, y_{n_\mathrm{r}})$, and $\boldsymbol{\theta} := \begin{bmatrix}\boldsymbol{\theta}_t^T & \boldsymbol{\theta}_r^T\end{bmatrix}^T$; (4) is from the linear approximation[2]

[2] Here we use, with a slight abuse of notation, $e^{j\boldsymbol{\theta}}$ to denote the vector obtained from the element-wise complex exponential operation. Similarly, the little-o Landau notation $o(\boldsymbol{\theta})$ is element-wise.



(a) 4-QAM, $f = \hat{f} = -10$. (b) 64-QAM, $f = \hat{f} = -1.6$.

Fig. 1: Proposed approximate likelihood in the scalar case: $\gamma = 30$dB and phase noise has standard deviation 3°.

$e^{j\boldsymbol{\theta}} = 1 + j\boldsymbol{\theta} + o(\boldsymbol{\theta})$. Thus the Euclidean norm has the real approximation $\|\boldsymbol{y} - \boldsymbol{\Lambda}_{\mathrm{R}}\boldsymbol{H}\boldsymbol{\Lambda}_{\mathrm{T}}\boldsymbol{x}\|^2 \approx \|\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{b}\|^2$ where $\boldsymbol{A} \in \mathbb{R}^{2n_{\mathrm{r}} \times (n_{\mathrm{t}}+n_{\mathrm{r}})}$ and $\boldsymbol{b} \in \mathbb{R}^{2n_{\mathrm{r}} \times 1}$ are defined as

$$\boldsymbol{A} := \begin{bmatrix} \Im[\boldsymbol{H}\boldsymbol{D}_x] & \Im[\boldsymbol{D}_y] \\ -\Re[\boldsymbol{H}\boldsymbol{D}_x] & -\Re[\boldsymbol{D}_y] \end{bmatrix}, \quad \boldsymbol{b} := \begin{bmatrix} \Re[\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}] \\ \Im[\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}] \end{bmatrix}. \quad (5)$$

The above approximation leads to the following result.

Proposition 1. *Let $\boldsymbol{A}$ and $\boldsymbol{b}$ be defined as in (5). Then we have the following approximation of the log-likelihood function $\ln\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[e^{-\gamma\|\boldsymbol{y}-\boldsymbol{H}_{\theta}\boldsymbol{x}\|^2}\right]\right) \approx \hat{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_{\theta})$ with*

$$\hat{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{H}) := -\gamma\,\boldsymbol{b}^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{x}}^{-1}\boldsymbol{b} - \frac{1}{2}\ln\det\left(\boldsymbol{W}_{\boldsymbol{x}}\right), \text{ and} \quad (6)$$

$$\boldsymbol{W}_{\boldsymbol{x}} := \boldsymbol{I} + 2\gamma\boldsymbol{A}\boldsymbol{Q}_{\theta}\boldsymbol{A}^{\mathsf{T}}. \quad (7)$$

Hence, the proposed approximate ML (aML) solution is

$$\hat{\boldsymbol{x}}_{\mathrm{aML}}(\boldsymbol{y}, \boldsymbol{H}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_t}}\left\{\gamma\,\boldsymbol{b}^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{x}}^{-1}\boldsymbol{b} + \frac{1}{2}\ln\det\left(\boldsymbol{W}_{\boldsymbol{x}}\right)\right\}. \quad (8)$$

Proof. The proof is straightforward after applying the above approximation. Details can be found in the full paper [8]. □

In Fig. (1), we illustrate the proposed approximation for 4- and 64-QAM, we plot for each constellation point a level set of the likelihood function with respect to "$\boldsymbol{y}$" in solid line. The level sets of the approximated likelihood function are plotted similarly in dashed line. While the likelihood function is evaluated using numerical integration, the approximation is in closed form given by (6). In this figure, we observe that the approximation is quite accurate, especially for signal points with smaller amplitude. Further, the resemblance of the level sets for the approximate likelihood to ellipsoids suggests that the main contribution in the right hand side of (6) comes from the first term $-\gamma\,\boldsymbol{b}^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{x}}^{-1}\boldsymbol{b}$. We shall exploit this feature later on to construct the proposed algorithm.

While the proposed approximation simplifies significantly the objective function, the optimization problem (8) remains hard when the search space is large. For instance, with 64-QAM and 4 × 4 MIMO, the number of points in $\mathcal{X}_{\mathrm{t}}^{n}$ is more than 10^7! Therefore, we need further simplification by exploiting the structure of the problem.

B. The Self-Interference Whitening Algorithm

The difficulty of the optimization (8) is mainly due to the presence of the matrix $\boldsymbol{W}_{\boldsymbol{x}}$ that depends on $\boldsymbol{x}$. Let us first assume that the $\boldsymbol{W}_{\boldsymbol{x}}$ corresponding to the optimal solution $\hat{\boldsymbol{x}}_{\mathrm{aML}}$ were somehow known, and is denoted by $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$. Then the optimization problem (8) would be equivalent to

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathrm{aML}}(\boldsymbol{y}, \boldsymbol{H}) &= \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_{\mathrm{t}}}}\left\{\gamma\,\boldsymbol{b}^{\mathsf{T}}\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-1}\boldsymbol{b} + \frac{1}{2}\ln\det\left(\boldsymbol{W}_{\hat{\boldsymbol{x}}}\right)\right\} \\ &= \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_{\mathrm{t}}}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-1}\boldsymbol{b} \\ &= \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_{\mathrm{t}}}}\left\|\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}})\right\|^2, \quad (9)\end{aligned}$$

where $\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}}$ is any matrix such that $\left(\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}}\right)^{\mathsf{H}}\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}} = \boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-1}$;

$$\tilde{\boldsymbol{x}} := \begin{bmatrix}\Re[\boldsymbol{x}] \\ \Im[\boldsymbol{x}]\end{bmatrix}, \ \tilde{\boldsymbol{y}} := \begin{bmatrix}\Re[\boldsymbol{y}] \\ \Im[\boldsymbol{y}]\end{bmatrix}, \ \tilde{\boldsymbol{H}} := \begin{bmatrix}\Re[\boldsymbol{H}] & -\Im[\boldsymbol{H}] \\ \Im[\boldsymbol{H}] & \Re[\boldsymbol{H}]\end{bmatrix}.(10)$$

Note that for a given $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$, (9) can be solved with any NND algorithm. Unfortunately, without knowing the optimal solution $\hat{\boldsymbol{x}}_{\mathrm{aML}}$, the exact $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$ cannot be found. Therefore, the idea is to first estimate the matrix $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$ with some suboptimal solution $\hat{\boldsymbol{x}}$, and then solve the optimization problem (9) with a NND. We call this two-step procedure *self-interference whitening* (SIW). For instance, we can use the naive ML solution $\hat{\boldsymbol{x}}_{\mathrm{ML}}^{0}$ as the initial estimate to obtain $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$, and have $\hat{\boldsymbol{x}}'_{\mathrm{aML}}(\boldsymbol{y}, \boldsymbol{H}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_{\mathrm{t}}}}\left\|\boldsymbol{W}_{\hat{\boldsymbol{x}}_{\mathrm{ML}}^{0}}^{-\frac{1}{2}}\tilde{\boldsymbol{y}} - \boldsymbol{W}_{\hat{\boldsymbol{x}}_{\mathrm{ML}}^{0}}^{-\frac{1}{2}}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}}\right\|^2$.

Remark 1. *The intuition behind the SIW scheme is as follows. From the definition of $\boldsymbol{W}_{\boldsymbol{x}}$ in (7) and $\boldsymbol{A}$ in (5), we see that $\boldsymbol{W}_{\boldsymbol{x}}$ depends on $\boldsymbol{x}$ only through $\boldsymbol{H}\boldsymbol{D}_x$. First, the column space of $\boldsymbol{H}\boldsymbol{D}_x$ does not vary with $\boldsymbol{x}$ since $\boldsymbol{D}_x$ is diagonal. Second, a small perturbation of $\boldsymbol{x}$ does not perturb $\boldsymbol{W}_{\boldsymbol{x}}$ too much. Since the naive ML point $\hat{\boldsymbol{x}}_{\mathrm{ML}}^{0}$ is close to the actual point $\boldsymbol{x}$ in the column space of $\boldsymbol{H}$, it provides an accurate estimate of $\boldsymbol{W}_{\boldsymbol{x}}$. This can also be observed on Fig. (1b), where we see that the ellipsoid-like dashed lines have similar sizes and orientations for constellation points that are close to each other.*

Remark 2. *Another possible initial estimate is the naive linear minimum mean square error (LMMSE) solution. As the naive ML, the naive LMMSE ignores the phase noise and returns*

$$\hat{\boldsymbol{x}}_{\mathrm{LMMSE}}^{0}(\boldsymbol{y}, \boldsymbol{H}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}^{n_t}}\|\boldsymbol{H}^{H}(\gamma^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^{H})^{-1}\boldsymbol{y} - \boldsymbol{x}\|^2. \quad (11)$$

The SIW algorithm is described in Algorithm 1. Here, the complex function NND$(\boldsymbol{y}, \boldsymbol{H}, \mathcal{X})$ finds among the points from the alphabet $\mathcal{X}$ the closest one to $\boldsymbol{y}$ in the column space of $\boldsymbol{H}$; the function realNND$(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{H}}, \tilde{\mathcal{X}})$ is the real counterpart of NND. The function "complex$(\tilde{\boldsymbol{x}}')$" embeds the real vector $\tilde{\boldsymbol{x}}'$ to the complex space by taking the upper half as the real part and the lower half as the imaginary part. The SIW outputs the newly obtained point only if it has a higher approximate likelihood value than the naive ML point does. An example of the scalar case using 256-QAM is shown in Fig. (2). The transmitted point is x and the received point is y. The solid line is the level set of the likelihood function. If

Algorithm 1 Self-interference whitening

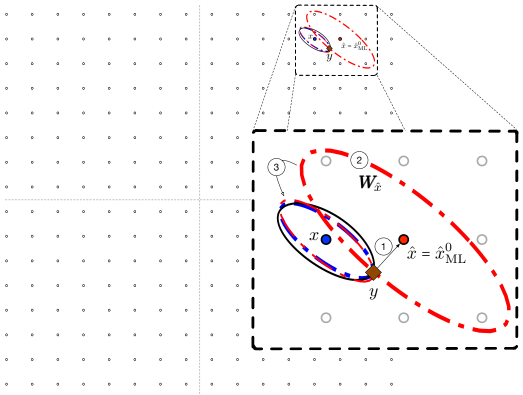
```
Input: $\boldsymbol{y}$, $\boldsymbol{H}$, $\gamma$, $\boldsymbol{Q}_\theta$
Find $\hat{\boldsymbol{x}}^0_{\mathrm{LMMSE}}$ from (11)
Find $\hat{\boldsymbol{x}}^0_{\mathrm{ML}} \leftarrow \mathrm{NND}(\boldsymbol{y}, \boldsymbol{H}, \mathcal{X})$
if $\hat{f}(\hat{\boldsymbol{x}}^0_{\mathrm{LMMSE}}, \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_\theta) > \hat{f}(\hat{\boldsymbol{x}}^0_{\mathrm{ML}}, \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_\theta)$ then
    $\hat{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{x}}^0_{\mathrm{LMMSE}}$
else
    $\hat{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{x}}^0_{\mathrm{ML}}$
end if
Generate $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$ from $\hat{\boldsymbol{x}}$ using (5) and (7)
Find $\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{\frac{1}{2}}$ using the Cholesky decomposition
Generate $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{H}}$ according to (10)
$\tilde{\boldsymbol{x}}' \leftarrow \mathrm{realNND}(\boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}}\tilde{\boldsymbol{y}}, \boldsymbol{W}_{\hat{\boldsymbol{x}}}^{-\frac{1}{2}}\tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}}, \tilde{\mathcal{X}})$
$\hat{\boldsymbol{x}}' \leftarrow \mathrm{complex}(\tilde{\boldsymbol{x}}')$
if $\hat{f}(\hat{\boldsymbol{x}}', \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_\theta) > \hat{f}(\hat{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{H}, \gamma, \boldsymbol{Q}_\theta)$ then
    $\hat{\boldsymbol{x}}'_{\mathrm{aML}} \leftarrow \hat{\boldsymbol{x}}'$
else
    $\hat{\boldsymbol{x}}'_{\mathrm{aML}} \leftarrow \hat{\boldsymbol{x}}$
end if
Output: $\hat{\boldsymbol{x}}'_{\mathrm{aML}}$
```

the likelihood function were computed for each point in the constellation (this is computationally hard), one would recover x from y successfully. While Euclidean detection outputs the wrong point $\hat{x}$, SIW can "correct" the error by first estimating the unknown matrix $\boldsymbol{W}_{\boldsymbol{x}}$, and then computing the matrix $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$ which is represented by the red dashed ellipse around $\hat{x}$. The estimate $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$ is very close to the correct value $\boldsymbol{W}_{\boldsymbol{x}}$, given by the actual x (blue dashed line). Then, SIW searches for the closest constellation point to y in the coordinate system generated by $\boldsymbol{W}_{\hat{\boldsymbol{x}}}$, so that x is recovered successfully. Also, computationally efficient NND algorithms can be used to perform the search.

The complexity of the SIW algorithm is essentially twice that of the NND algorithm used, since the other operations including the LMMSE detection have at most cubic complexity with respect to the dimension of the channel. The complexity of the NND algorithm depends directly on the conditioning of the given matrix. If the columns are close to orthogonal, then channel inversion is almost optimal. However, in the worse case, when the matrix is ill-conditioned, the NND algorithm can be slow and its complexity is exponential in the problem dimension. There exist approximate NND algorithms, e.g., based on lattice reduction, that can achieve near optimal performance with much lower complexity.

IV. Numerical Experiments

We now compare the performance[3] of SIW to ML and other baseline schemes: i) the naive LMMSE (11), ii) the naive ML (3), and iii) the selection between the two where the receiver outputs the one whose approximate likelihood value is higher. We derive a lower bound on the performance of



x
y
$\hat{x} = \hat{x}^0_{\mathrm{ML}}$
$\boldsymbol{W}_{\hat{x}}$
1
2
3

Fig. 2: SIW in the scalar case: 256-QAM, PN 2°. Dashed lines are the ellipses defined by $\boldsymbol{W}_{\hat{x}}$ (in red) and $\boldsymbol{W}_{x}$ (in blue).

ML detection since it is hard to implement. A scheme minimizing (6) admits the following performance lower bound:

$$P_{\mathrm{e}}^{\mathrm{aML}} \geq \mathbb{P}\left\{\hat{f}(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\boldsymbol{x} \in \mathcal{X}^{n_{\mathrm{t}}}} \hat{f}(\boldsymbol{x}, \mathbf{Y}, \mathbf{H})\right\} \tag{12}$$

$$\geq \mathbb{P}\left\{\hat{f}(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\boldsymbol{x} \in \mathcal{L} \subseteq \mathcal{X}^{n_{\mathrm{t}}}} \hat{f}(\boldsymbol{x}, \mathbf{Y}, \mathbf{H})\right\}, \tag{13}$$

where (12) holds by definition and (13) by monotonicity. While (13) holds for all $\mathcal{L}$, one has equality if $\mathcal{L}$ contains *all* the points in $\mathcal{X}^{n_{\mathrm{t}}}$ that have a higher approximate likelihood value than $\mathbf{X}$ does. Here, we consider a large set around $\mathbf{X}$ to compute the lower bound (13), but do not study its tightness. Similarly, for ML detection, we have

$$P_{\mathrm{e}}^{\mathrm{ML}} \geq \mathbb{P}\left\{f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\boldsymbol{x} \in \mathcal{X}^{n_{\mathrm{t}}}} f(\boldsymbol{x}, \mathbf{Y}, \mathbf{H})\right\} \tag{14}$$

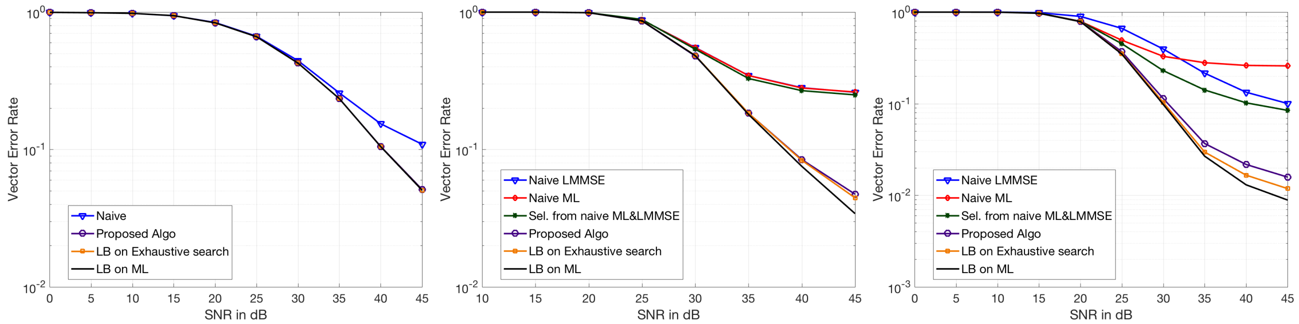
$$\geq \mathbb{P}\left\{f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H})\right\}, \quad \forall \mathbf{X}' \in \mathcal{X}^{n_{\mathrm{t}}} \tag{15}$$

$$= \mathbb{P}\left\{\mathbf{X} \neq \mathbf{X}',\ f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H})\right\}, \tag{16}$$

where (14) holds by definition and (16) holds since $\mathbf{X} \neq \mathbf{X}'$ is a consequence of $f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H})$. Also, (15) holds for any $\mathbf{X}' \in \mathcal{X}^{n_{\mathrm{t}}}$, with equality if $\mathbf{X}'$ is the exact ML solution. Since the ML solution is unknown, one may use any suboptimal solution instead and to obtain a lower bound. Indeed (16) is much easier to evaluate than (14), as the latter requires to minimize over $\mathcal{X}^{n_{\mathrm{t}}}$. Intuitively, if $\mathbf{X}'$ is a near ML solution, then the lower bound should be tight enough. We need to perform twice the numerical integration only when $\boldsymbol{x}' \neq \boldsymbol{x}$. If $\boldsymbol{x}' \neq \boldsymbol{x}$ with small probability, evaluating (16) can be done quickly.

In Fig. (3a) we consider point-to-point Rayleigh fading single-antenna, i.e., single-input single-output (SISO), channels. We consider 1024-QAM with phase noise of standard deviation 1° at both the transmitter and receiver sides. Here the naive ML scheme is in fact a simple threshold detection for the real and imaginary parts. First, we see that ignoring the existence of phase noise incurs a significant performance loss. Second, if exhaustive search is done with the proposed likelihood approximation, then it achieves the ML performance,

[3] Our performance metric is the vector detection error rate: detection is considered successful only when all the symbols in $\boldsymbol{x}$ are recovered correctly.



(a) SISO, 1024-QAM, PN 1°.

(b) 4×4 LoS-MIMO, 1024-QAM, PN 1°.

(c) 4×4 Uplink, 256-QAM, PN 2°.

Fig. 3: Simulation results for different communication scenarios.

since the simulation-based lower bound overlaps with the curve with exhaustive search. This confirms the accuracy of the approximation (6) in the SISO case. Even more remarkably, the SIW algorithm almost achieves the ML performance without exhaustive search.

In Fig. (3b) we consider point-to-point line-of-sight (LoS) MIMO, commonly deployed as microwave backhaul links [10], [12], [13]. We assume that the channel is constant over time but each antenna is driven by its own oscillator. This is the worst-case assumption, and motivated in practice by the fact that the communication distance is large and thus the distance between antenna elements is increased so that the channel matrix is well conditioned [12], [13]. We use the model of [13] with two transmit and two receive antennas each with dual polarizations, effectively a 4×4 MIMO channel. The optimal distance between the antenna elements at each side can be derived as a function of the communication distance [12], for which the channel matrix is unitary. As above we consider 1024-QAM with phase noise of standard deviation $1°$. As in the SISO case, phase noise mitigation substantially improves the performance and the proposed likelihood approximation remains accurate as shown by the comparison between the exhaustive search (8) and the lower bound on ML detection.

In Fig. (3c) we consider the uplink cellular communication channel with four single-antenna users and one multi-antenna base station receiver. We assume i.i.d. phase noises at the users' side with standard deviation $2°$ and no phase noise at the receiver side. This is a reasonable assumption since the oscillators at the base station are usually of higher quality than those used by mobile devices. We assume i.i.d. Rayleigh fading. Unlike in the previous scenarios, the naive ML is (surprisingly) dominated by the naive LMMSE at high SNR. Indeed, without receiver phase noise, inverting the channel yields spatial parallel channels. Although this incurs a power loss in general, phase noises across the parallel sub-channels are independent, so the demodulation only suffers from a scalar self-interference. On the other hand, naive ML suffers from the aggregated perturbation from all the phase noises. So naive LMMSE beats naive ML detection at high SNRs where phase noise dominates the additive noise. If both the transmitter and receiver have comparable phase noises, this does not occur, as channel inversion amplifies the receiver phase noises. The gain of SIW over the other schemes is clear.

V. Conclusions

We have studied the ML detection problem for uncoded MIMO phase noise channels, and proposed an approximation of the likelihood function that has been shown to be accurate in the regimes of practical interest. More importantly, using the geometric interpretation of the approximate likelihood function, we have designed SIW, an efficient approximate algorithm requiring only two nearest neighbor detections.

References

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, Aug. 2002.

[2] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channel," *IEEE Trans. Inf. Theory*, vol. 45, no. 10, Jul. 1999.

[3] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Sig. Proc.*, vol. 53, no. 8, Aug. 2005.

[4] J. Jaldén and P. Elia, "DMT optimality of LR-aided linear decoders for a general class of channels, lattice designs, and system models," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, Oct. 2010.

[5] G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "On the selection of a two-dimensional signal constellation in the presence of phase jitter and gaussian noise," *Bell Labs Tech. J.*, vol. 52, no. 6, 1973.

[6] A. Hajimiri and T. H. Lee, "A general theory of phase noise in electrical oscillators," *IEEE J. of Solid-State Circuits*, vol. 33, no. 2, Feb. 1998.

[7] H. Ghozlan and G. Kramer, "Models and information rates for Wiener phase noise channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, Apr. 2017.

[8] R. Combes and S. Yang, "An approximate ML detector for MIMO channels corrupted by phase noise," to appear in *IEEE Trans. Commun.*

[9] A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: A unifying theory and numerical methods for characterization," *IEEE Trans. Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 5, May 2000.

[10] G. Durisi, A. Tarable, C. Camarda, R. Devassy, and G. Montorsi, "Capacity bounds for MIMO microwave backhaul links affected by phase noise," *IEEE Trans. Commun.*, vol. 62, no. 3, Mar. 2014.

[11] S. Yang and S. Shamai (Shitz), "On the multiplexing gain of discrete-time MIMO phase noise channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, Apr. 2017.

[12] F. Bøhagen, P. Orten, and G. E. Øien, "Design of optimal high-rank line-of-sight MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, Apr. 2007.

[13] P. Ferrand and S. Yang, "Blind precoding in line-of-sight MIMO channels," in *IEEE SPAWC*, Edinburgh, UK, 2016.

Generalized BP Decoder with Cycle Decomposition for Short-Length LDPC Codes

Iryna Andriyanova
University of Paris Seine (University of Cergy-Pontoise/ENSEA/CNRS)
ETIS-UMR8051 Laboratory
95015 Cergy-Pontoise, France
Email: iryna.andriyanova@ensea.fr

***Abstract*—This work presents a new version of the generalized belief propagation (GBP) decoding algorithm for regular LDPC ensembles, where a cycle decomposition of the underlying Tanner graph is used to define GBP clusters. Such an approach allows to take advantage of the presence of cycles in the Tanner graph of a code, and not to ignore them, as the usual BP algorithm does. The new GBP algorithm improves the decoding performance of regular LDPC codes of lengths from several dozens to several hundreds bits.**

Owing to multiple upcoming applications in the world of Internet of Things, there is a growing interest in design of short-length linear codes, able to serve for real-time forward error correction. The codes are expected to have codelengths from several dozens to several hundreds of bits and to have a real-time decoding algorithm. Graph codes, and LDPC codes in particular, could be good candidates for these applications, thanks to the low complexity of their belief propagation (BP) decoding algorithm [1]. Unfortunately, BP is known to have a bad performance over short codelengths, and the reason for this is the presence of short cycles in the underlying Tanner graph for a given graph code.

Considering LDPC codes, the problem of cycles is not new and has created a number of modifications of the original BP decoding procedure, either with a post-processing of messages over the cycles in the residual Tanner graph (see, e.g., [2], [3]) or, as in case of structured codes (e.g., Repeat-Accumulate codes), by performing a MAP decoding over some particular cycle(s) in the graph. For quantum LDPC codes, for which 4-cycles are an issue, an interesting approach has been taken in case of the Kitaev's toric code [4] by using clusters and renormalisation groups. Moreover the idea of BP decoding over clusters has been proposed in [5] and is called a Generalized Belief Propagation (GBP). Unfortunately, the GBP has several main drawbacks [6]: 1) it usually worsens the iterative decoding threshold; 2) it is difficult to find an appropriate cluster characterisation in a general case, for an arbitrary LDPC code structure and 3) the complexity of the GBP decoding is usually much larger compared to the BP decoding.

In this work, a new GBP decoding algorithm is investigated, with the aim to circumvent the drawbacks 2 and 3 of the GBP for short-length LDPC codes (when the iterative decoding threshold degradation is not an issue). Given a Tanner graph of an LDPC code, one uses a cycle decomposition of the graph in order to create clusters, and an iterative algorithm is further defined in order to exchange extrinsic messages between the clusters. A cluster is related to a cycle in the Tanner graph of given LDPC code, and can therefore be represented by a tailbiting convolutional code. Thanks to the 2-state trellis representation of this convolutional component code, each cluster can be decoded with a low-complexity MAP decoding algorithm.

Our first results are obtained for $(2, 4)$ LDPC codes over the binary erasure channel. As the degrees of both variable and check nodes are even, the cycle decomposition has disjoint cycles. Figure 1 shows the word error rates (WERs) of three $(2, 4)$ LDPC codes of respective lengths 100, 500 and 1000, decoded by the standard BP algorithm (blue curves) and our GBP algorithm (red curves). The numerical results are also compared with lower bounds on the WER of a linear code under ML decoding (Theorem 2 of [7]), shown by black dashed curves. Our results will further be extended to the case of odd variable and/or check node degrees.

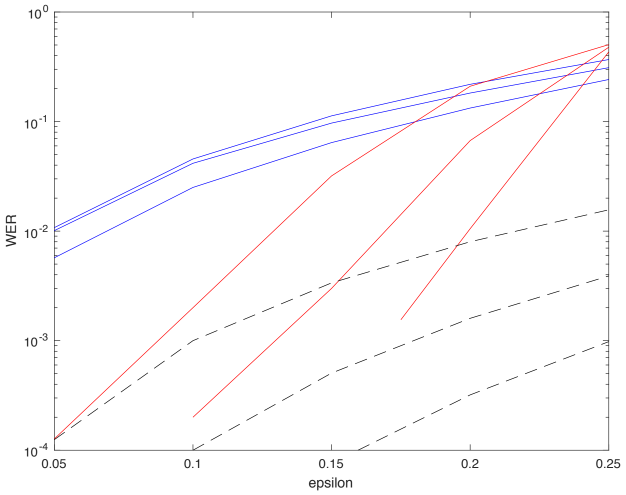


Fig. 1: WERs vs. erasure probability of the BEC for $(2, 4)$ LDPC codes of lengths 100, 500 and 1000 (from top to bottom) resp., under BP decoding (blue) and under our GBP decoding (red), compared with lower bounds on the MP decoding (dashed curves).

References

[1] T. Richardson and R. Urbanke, *Modern Coding Theory,* Cambridge University Press, 2008.

[2] P. M. Olmos, J. J. Murillo-Fuentes, F. Pérez-Cruz, *Tree-structure expectation propagation for decoding LDPC codes over binary erasure channels*, in Proc. of IEEE ISIT'2010, June 2010.

[3] T. Nozaki, K. Kasai and K. Sakaniwa, *Message passing algorithm with MAP decoding on zigzag cycles for non-binary LDPC codes*, in Proc. of IEEE ISIT'2013, July 2013.

[4] G. Duclos-Cianci and D. Poulin, *A renormalization group decoding algorithm for topological quantum codes*, ArXiv submission 1006.1362, June 2010. Available: https://arxiv.org/pdf/1006.1362.pdf.

[5] J.S. Yedidia, W.T. Freeman, and Y. Weiss, *Generalized belief propagation*, In Proc. of NIPS, pp. 689–695, 2000.

[6] J.-C. Sibel, *Region-based approximation to solve inference in loopy factor graphs: decoding LDPC codes by Generalized Belief Propagation*, PhD thesis, University of Cergy-Pontoise, June 2013. Available: http://biblioweb.u-cergy.fr/theses/2013CERG0629.pdf.

[7] I.E. Bocharova, B.D. Kudryashov, V. Skachek, E. Rosnes and Ø. Ytrehus, *ML and Near-ML Decoding of LDPC Codes Over the BEC: Bounds and Decoding Algorithms*, ArXiv submission 1709.01455, Sept. 2017. Available: http://arxiv.org/abs/1709.01455.

Codes on Graphs, Trellises and Spatial Coupling: another look at self-concatenated convolutional codes

Michael Lentmaier†, Saeedeh Moloudi†, and Alexandre Graell i Amat‡
†Department of Electrical and Information Technology, Lund University, Lund, Sweden
‡Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
{michael.lentmaier, saeedeh.moloudi}@eit.lth.se, alexandre.graell@chalmers.se

***Abstract*—We present recent results on spatially coupled turbo-like codes and point out a connection to the self-concatenated trellis construction proposed by Loeliger in 1997 [1].**

I. Codes on graphs with trellis constraints

We use the term turbo-like codes for generalized LDPC codes with convolutional component codes. In the corresponding factor graphs, the factor nodes represent trellis constraints instead of single parity-check equations. To efficiently describe the structure of different ensembles, similarly to protographs, we use a compact graph representation, illustrated in Fig. 1.

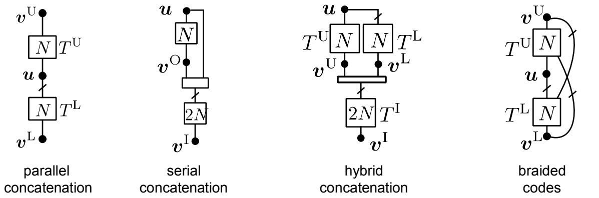


Fig. 1. A compact graph representation of turbo-like code ensembles [2].

II. Spatial coupling and threshold saturation

Spatially coupled (SC) turbo-like code ensembles can then be obtained from a sequence of compact graphs, analogously to SC-LDPC codes. The graphs are coupled by connecting the variable nodes at position t to factor nodes at positions $t, \ldots, t+m$, where m denotes the coupling memory. For the BEC, it can be shown using exact density evolution (DE) recursions that with spatial coupling the decoding threshold of an iterative BP decoder can be improved to the threshold of an optimal MAP decoder (threshold saturation) [2] [3]. For the AWGN channel, where exact DE recursions are not available, the thresholds can be estimated with Monte Carlo methods or predicted from the BEC thresholds [4]. For the ensembles shown in Fig. 1, the predicted thresholds are given in Table I.

TABLE I
Predicted AWGN channel thresholds (E_b/N_0[dB]) for different turbo-like code ensembles of rate $R = 1/3$.

	BP	MAP	$\mathrm{BP}_{\mathrm{SC}}^{m=1}$	$\mathrm{BP}_{\mathrm{SC}}^{m=3}$	$\mathrm{BP}_{\mathrm{SC}}^{m=5}$
Parallel	-0.1052	-0.3070	-0.3070	-0.3070	-0.3070
Serial	1.4023	-0.4740	-0.1196	-0.4673	-0.4740
Braided	1.2139	-0.4723	-0.4690	-0.4723	-0.4723
Hybrid	3.8846	-0.4941	0.2809	-0.4706	-0.4941

Observe that, while the classical turbo codes (parallel concatenation) have the best BP threshold, their MAP threshold is actually the worst. Vice versa, hybrid concatenated and braided codes have poor BP thresholds without spatial coupling. On the other hand, hybrid codes have the best MAP threshold, and braided codes are best for small coupling memory $m = 1$.

III. Self-concatenated convolutional codes

The idea of self-concatenated codes goes back to 1997 and has been proposed independently (at the same conference) by Loeliger [1] and by Divsalar / Pollara [5]. Compared to LDPC codes, a characteristic feature of turbo-like codes is that their factor graph contains a small number of long trellis constraints instead of a large number of short constraints. But what is the advantage of having more than one trellis? In [6] we have presented a unified self-concatenated ensemble, shown in Fig. 2, which contains all the ensembles in Fig. 1 as special cases. A novel element in this ensemble is the feedback of parity bits, which is required for representing the serial, hybrid, and braided ensembles. The various instances of the ensemble differ in the amount of feedback, puncturing and the structure of the permutations. Spatial coupling can be achieved by imposing a causality condition on the permutation matrices.

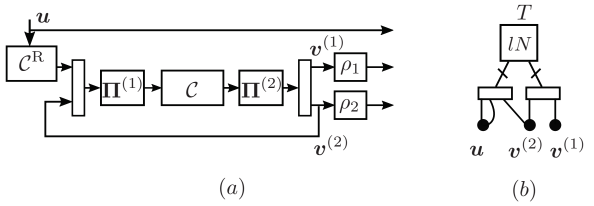


Fig. 2. A unified ensemble based on a single trellis (a) encoder block diagram (b) compact graph.

References

[1] H. A. Loeliger, "New turbo-like codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Ulm, Germany, June 1997.

[2] S. Moloudi, M. Lentmaier, and A. Graell i Amat, "Spatially coupled turbo-like codes," *IEEE Trans. on Inf. Theory*, vol. 63, no. 10, Oct. 2017.

[3] S. Moloudi, M. Lentmaier, and A. Graell i Amat, "Spatially coupled hybrid concatenated codes," in *Proc. 11th Int. ITG Conf. on Systems, Communications and Coding (SCC)*, Hamburg, Germany, Feb. 2017.

[4] M. U. Farooq, S. Moloudi, and M. Lentmaier, "Thresholds of braided convolutional codes on the AWGN channel," submitted to *IEEE Int. Symp. Inf. Theory (ISIT)*, 2018.

[5] D. Divsalar and F. Pollara, "Hybrid concatenated codes and iterative decoding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Ulm, Germany, June 1997.

[6] S. Moloudi, M. Lentmaier, and A. Graell i Amat, "A unified ensemble of concatenated convolutional codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, June 2017.

Exact mutual information of sparse superposition codes from the adaptive path interpolation method

Jean Barbier and Nicolas Macris
EPFL - IC - LTHC - Station 14 - CH-1015 Lausanne, Switzerland

Extended Abstract. Sparse superposition (SS) codes, or sparse regression codes, were first introduced by Barron and Joseph in 2010 for reliable communication over the additive white Gaussian noise channel and proven to be capacity-achieving under adaptive successive decoding along with power allocation [1], [2] and under the approximate message passing algorithm [3]. The *information word* or *message* is a vector made of L *sections*, $\mathbf{s} = [\mathbf{s}_1, \dots, \mathbf{s}_L]$. Each section is a B-dimensional vector with a single component equal to 1 and $B-1$ components equal to 0. We set $N = LB$ for the block length. The message $\mathbf{s}$ can be seen as a one-to-one mapping from an original message $\mathbf{u} \in \{0,1\}^{L\log_2(B)}$, where the position of the non-zero component in $\mathbf{s}_l$ is specified by the binary representation of $\mathbf{u}_l$ (i.e. $\mathbf{s}$ is obtained from $\mathbf{u}$ using a simple position modulation scheme). We consider random codes generated by a fixed *coding matrix* $\mathbf{F} \in \mathbb{R}^{M\times N}$ with i.i.d real Gaussian entries distributed as $\mathcal{N}(0, 1/L)$. The *codeword* $\mathbf{Fs} \in \mathbb{R}^M$ has a normalized average power $\mathbb{E}||\mathbf{Fs}||_2^2/M = 1$. The cardinality of this code is B^L and the length of the codeword is M. Hence, the (design) rate is defined as $R = (N\log_2 B)/(MB)$. The code is thus specified by (M, R, B) where R is the code rate, M the block length, B the section size. Codewords are transmitted through a known memoryless channel W. This requires a mapping π to map codeword components $[\mathbf{Fs}]_\mu \in \mathbb{R}$, $\mu \in \{1, \dots, M\}$, onto the input alphabet of W. The concatenation of π and W can be seen as an *effective memoryless channel* P_{out}, such that $P_{\text{out}}(\mathbf{y}|\mathbf{Fs}) := \prod_{\mu=1}^M W(y_\mu|\pi([\mathbf{Fs}]_\mu))$.

Decoding amounts to recover $\mathbf{s}$ from channel outputs $\mathbf{y}$. For Gaussian channels it can be interpreted as a "standard" *compressed sensing* problem with structured sparsity where $\mathbf{y}$ would be the compressed measurements. The rate R can be linked to the "measurement rate" α, used in the compressed sensing literature, by $\alpha = M/N = (\log_2 B)/(BR)$. For other channels it is more akin to a *generalized linear estimation* problem (again with structured sparsity for $\mathbf{s}$). Thus, the same algorithms and analysis used in compressed sensing and generalized linear estimation theory like the generalized approximate message passing algorithms algorithm and state evolution [5] can be used in the present context. See [4]–[6]. The optimal ensemble threshold for reliable communication is determined by the mutual information. This quantity is given (up to a trivial channel dependent term) by the average *free energy* $f = \lim_{N\to+\infty} N^{-1}\mathbb{E}\ln Z(\mathbf{y}, \mathbf{F})$ where $Z(\mathbf{y}, \mathbf{F})$ is the normalizing factor of the posterior distribution $P_0(\mathbf{s})P_{\text{out}}(\mathbf{y}|\mathbf{Fs})/Z(\mathbf{y}, \mathbf{F})$. The replica method of statistical mechanics conjectures an exact expression for f from which the optimal ensemble threshold can be read off [7], [8].

Recently it has been possible to prove the replica formulas for generalized estimation problems [9] (and consequently also for the SS code ensemble described above). Such formulas are given by "single letter" variational problems of the form $f = \sup_{q\in[0,\rho]} \inf_{r>0}\{\psi_{P_0}(r) + \alpha\psi_{P_{\text{out}}}(q;\rho) - rq/2\}$ where $\psi_{P_0}(r) = \mathbb{E}\ln\int dP_0(x)e^{\sqrt{r}Y_0x - rx^2/2}$ is the free entropy of a scalar Gaussian channel $Y_0 = \sqrt{r}X_0 + Z_0$, $Z_0 \sim \mathcal{N}(0,1)$ and $\psi_{P_{\text{out}}}(q;\rho) = \mathbb{E}\ln\int \mathcal{D}wP_{\text{out}}(\widetilde{Y}_0|\sqrt{q}\,V + \sqrt{\rho-q}\,w)$, $V \sim \mathcal{N}(0,1)$, $\widetilde{Y}_0 \sim P_{\text{out}}(\cdot|\sqrt{q}\,V + \sqrt{\rho-q}\,\widetilde{W})$, $\widetilde{W} \sim \mathcal{N}(0,1)$, $\mathcal{D}w = dw(2\pi)^{-1/2}e^{-w^2/2}$, $\rho = \mathbb{E}[X_0^2]$. The proof proceeds by an *adaptive path interpolation method* recently developed for a number of simpler estimation problems (e.g. matrix and tensor factorization problems) [10]. Proofs of replica fromulas have a long history by now which started with the fundamental work of Guerra and Toninelli on the Sherrington-Kirkpatrick spin glass. See [11] for an introduction. Roughly speaking, the new element introduced in the adaptive path method exploits remarkable (Nishimori) identities generally valid in Bayesian inference which imply concentration of overlap parameters and vanishing of the Guerra-Toninelli remainder terms.

References

[1] A. Joseph, A. Barron, *Fast sparse superposition codes have exponetial error probability for $R < C$,* IEEE Transactions on Information Theory, vol 60, no 2. pp. 919-942 (2014).

[2] A. Barron, S. Cho, *High-rate sparse superposition codes with iteratively optimal estimates*, ISIT (2012) pp. 120-124.

[3] C. Rush, A. Greig, R. Venkataramanan *Capacity-achieving sparse superposition codes via approximate message passing decoding*, IEEE Transactions on Information Theory, vol 63, no 3, pp. 1476-1500 (2017)

[4] J. Barbier and F. Krzakala, *Approximate message-passing decoder and capacity-achieving sparse superposition codes*, IEEE Transactions on Information Theory, vol 63, no 8, pp. 4894-4927 (2017)

[5] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, ISIT (2011) pp. 21682172.

[6] A. Javanmard and A. Montanari, *State evolution for general approximate message passing algorithms, with applications to spatial coupling*, Journal of Information and Inference, vol. 2, no. 2, pp. 115144 (2013).

[7] J. Barbier, M. Dia, N. Macris, *Universal sparse superposition codes with spatial coupling and GAMP decoding*, arXiv:1707.04203

[8] E. Biyik, J. Barbier, M. Dia, *Generalized approximate message passing decoder for universal sparse superposition codes*, arXiv:1701.03590

[9] J. Barbier, F. Krzakala, N. Macris, L. Miolane, L. Zdeborovà, *Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models*, arXiv:1708.03395

[10] J. Barbier, N. Macris, *The stochastic interpolation method: A simple scheme to prove replica formulas in Bayesian inference* , arXiv:1705.02780

[11] F. Guerra, *An introduction to mean field spin glass theory: methods and results*, Proc. of the 2005 Les Houches Summer School on Mathematical Statistical Physics pp. 243271 (2006).

Stabilizer Quantum Codes and their Factor Graphs

July X. Li and Pascal O. Vontobel
Department of Information Engineering
The Chinese University of Hong Kong
{july.x.li, pascal.vontobel}@ieee.org

Abstract—**Graphical notations like factor graphs have proven to be very useful toward expressing relationships between (random) variables and toward formulating low-complexity algorithm for either exactly or approximately computing quantities of interest. Two particular classes of normal factor graphs (NFGs) have been used in the context of stabilizer quantum codes; in this paper, we review these two different classes of NFGs and explain a connection between them.**

Stabilizer quantum error-correction codes (QECC) are a popular approach to correct certain types of errors that can happen to quantum states. (For an introduction to this topic, see, e.g., [1, Ch. 10].) In this paper, we focus on two classes of normal factor graphs (NFGs) [2]–[4] that have been used in the context of stabilizer QECCs: on the one hand, the class of NFGs in [5], on the other hand, the class of NFGs in [6], [7]. In this paper we link the two approaches.

In order to proceed, we introduce some notation. Let I be the 2×2 identity matrix and let

$$X \triangleq \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Y \triangleq \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix}, \quad \text{and} \quad Z \triangleq \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

be the Pauli matrices. The group $\langle \mathcal{G}_1, \cdot \rangle$ with group elements

$$\mathcal{G}_1 \triangleq \big\{ c \cdot \mathcal{P} \;\big|\; c \in \{\pm 1, \pm \mathrm{i}\},\ \mathcal{P} \in \{I, X, Y, Z\} \big\}$$

and group operation given by matrix multiplication is known as the single-qubit Pauli group. The group

$$\langle \mathcal{G}_n, \cdot \rangle \triangleq \big(\langle \mathcal{G}_1, \cdot \rangle \big)^{\otimes n}$$

with group elements

$$\mathcal{G}_n \triangleq \big\{ c \cdot \mathcal{P}_1 \mathcal{P}_2 \cdots \mathcal{P}_n \;\big|\; c \in \{\pm 1, \pm \mathrm{i}\},\ \mathcal{P}_i \in \{I, X, Y, Z\} \big\},$$

and group operation given by matrix multiplication is known as the n-qubit Pauli group. (Here we have used the shorthand notation $\mathcal{P}_1 \mathcal{P}_2 \cdots \mathcal{P}_n \triangleq \mathcal{P}_1 \otimes \mathcal{P}_2 \otimes \cdots \otimes \mathcal{P}_n$ for Pauli operators.)

Let $\mathcal{N}_n$ be the set of unitary matrices of size $2^n \times 2^n$ such that

$$U \cdot \mathcal{G}_n \cdot U^{\mathsf{H}} = \mathcal{G}_n \quad \text{for all } U \in \mathcal{N}_n. \tag{1}$$

This set of matrices forms a group under matrix multiplication and is known as the normalizer of $\mathcal{G}_n$ (see, e.g., [1, Ch. 10]) or as the Clifford group (see, e.g., [8]). It turns out that any $U \in \mathcal{N}_n$ can be expressed, up to a global phase, in terms of $O(n^2)$ Hadamard, phase, and controlled-NOT gates (see, e.g., [1, Th. 10.6]).

Consider a k-qubit system $|\psi\rangle$ that needs to be protected w.r.t. so-called Pauli errors, i.e., errors where qubits can be affected by unitary matrices in the Pauli group (as, e.g., happens in the case of a quantum depolarization channel). Encoding of such a state in terms of an $[\![n, k]\!]$ stabilizer QECC can be expressed in terms of a suitable unitary matrix $U \in \mathcal{N}_n$ applied to $|\phi\rangle \otimes |0\rangle^{\otimes(n-k)}$, where $|0\rangle^{\otimes(n-k)}$ are $n-k$ ancilla qubits in the $+1$ eigenstate of the Z operator. Decoding, on the other hand, can be accomplished by applying the unitary matrix U^{H}, measuring the ancilla qubits with the measurement matrices $\frac{1}{2} \cdot (I \pm Z)$, figuring out the most likely error based on the measurement outcomes (with the help of a classical computer), and applying a suitable unitary transformation to the k qubits. It is rather straightforward to express this procedure in terms of an NFG from the class of NFGs that was introduced in [5]. (See also the reformulation of such NFGs in terms of so-called double-edge NFGs in [9].)

The stabilizer group of a stabilizer QECC is a subgroup of $\mathcal{G}_n$ containing all the operators that stabilize an arbitrary state $|\psi\rangle$ after encoding, i.e., applying a matrix from the stabilizer group to an encoded state leaves the encoded state invariant. (Note that the stabilizer group is commutative.) The papers [6], [7] introduced a class of NFGs that allow one to express the stabilizer group of a stabilizer QECC, more precisely, that allow one to express a vector representation of the stabilizer group. In particular, the stabilizer group of various types of stabilizer QECC can be expressed in terms of these NFGs, and relatively simple proofs can be used to show the commutativity of the stabilizer group. (Algebraically, this is expressed as a certain sub-orthogonality property under the symplectic inner product of the vector representation of the stabilizer group.) Such NFGs are of interest toward low-complexity algorithms for decoding stabilizer QECCs.

Although the two above-mentioned classes of NFGs are superficially rather different, a connection between them can be achieved by applying suitable closing-the-box operations [4], [10] and using (1). The details are explained in the presentation.

Acknowledgment

P.O.V. kindly acknowledges discussions related to the topic of this paper with Dave Forney and Andi Loeliger.

References

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press, 2000.

[2] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[3] G. D. Forney, Jr., "Codes on graphs: normal realizations," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001.

[4] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Sig. Proc. Mag.*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[5] H.-A. Loeliger and P. O. Vontobel, “Factor graphs for quantum probabilities,” *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5642–5665, Sep. 2017.

[6] P. O. Vontobel, “Interior-point algorithms for linear programming decoding,” in *Proc. Inf. Theory Appl. Workshop*, UC San Diego, La Jolla, CA, USA, Jan. 27–Feb. 1 2008.

[7] J. X. Li and P. O. Vontobel, “Factor-graph representations of stabilizer quantum codes,” in *Proc. 54th Allerton Conf. on Communication, Control, and Computing*, Allerton House, Monticello, IL, USA, Sep. 28–30 2016, pp. 1046–1053.

[8] D. Poulin, J.-P. Tillich, and H. Ollivier, “Quantum serial turbo codes,” *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2776–2798, Jun. 2009.

[9] M. X. Cao and P. O. Vontobel, “Double-edge factor graphs: definition, properties, and examples,” in *Proc. IEEE Inf. Theory Workshop*, Kaohsiung, Taiwan, Nov. 6–10 2017.

[10] P. O. Vontobel and H.-A. Loeliger, “On factor graphs and electrical networks,” in *Mathematical Systems Theory in Biology, Communication, Computation, and Finance, IMA Volumes in Math. & Appl.*, D. Gilliam and J. Rosenthal, Eds. Springer Verlag, 2003.

Adaptive Weighted Signal Detection for Nanoscale Molecular Communications

Arzhang Shahbazi
School of Electrical and
Computer Engineering
Shiraz University
Shiraz, Iran
Email: a.shahbazi@shirazu.ac.ir

Ali Jamshidi
School of Electrical and
Computer Engineering
Shiraz university
Shiraz, Iran
Email: jamshidi@shirazu.ac.ir

***Abstract*—Molecular Communication is a rising paradigm to transfer message between nano-machines. Due to the specific characteristics of these systems, the channel noise and memory significantly influence the system performance. In this paper, a new adaptive threshold detector is proposed, which utilized the inter-symbol-interference. In contrast to other detection algorithms with high complexity, the proposed detector is more practical when the channel conditions are not easy to find or not known at the receiver side. Numerical results show that the proposed detector achieves lower bit error rate than the common threshold detectors. Furthermore, the comparison between detectors is given, which is based on the variation of distance, symbol period and number of molecules.**

I. Introduction

ADVANCES in design and development of nanomachines inspires the study of communication between such units. Molecular communication (MC) has been introduced as one of the most promising paradigm for communication between devices at nanoscale which inspired by nature. Information can be transmitted by changing in number, type, and the timing of the molecules [1], [2], [3]. Nanomachines (NM) are very limited in terms of complexity [4]- [5] and it is stated that a single NM can only perform simple tasks. The connection of several NMs via MC channels increase the capabilities of one NM and realize the construction of nanonetworks. In molecular communications NMs transmit information through the diffusion of chemical molecules. The diffusion channel is fundamentally different in every aspect to a classical wireless communications channel. Since the diffusion channel impulse response (CIR) has a long tail, molecules from the previous transmitted symbols will interfere with current transmitted symbol, resulting in the inter-symbol interference (ISI) which makes a unreliable transmission [6]. However, a simplified transmitter model, channel model and receiver detection methods are generally assumed in the literature for making the analysis tractable.

The MC system with fixed threshold was first introduced in [7]. In [8], two different detection methods are classified as sampling-based (amplitude) and strength-based (energy). In sampling-based the detector simply measures the amount of molecules arriving at a certain instant of time while in strength-based detector, receiver accumulates the number of molecules arriving at a certain time period. In [9], the value of the threshold is designed to maximize a posteriori probability. In [6], detection process is based on multiple observation with a weight assigned to each observation. Sequence detection methods based on maximum a posteriori (MAP) and maximum likelihood (ML) criterions introduced in [10]. It had been shown in [11] that the effect of ISI can be utilized in a new adaptive-threshold algorithm which results better performance in certain situations. While by means of adaptive-threshold the detector accumulates the number of received molecules in each symbol period and compare it with number of previous symbol molecules.

As the main contribution of this paper, a new detection algorithm is proposed which optimize the threshold by applying weights to current and previous symbols energies. This new decoding method while being low-complex as possible, achieves a reasonable performance compare to common detectors. Numerical results and a comparison between detectors in terms of distance, symbol duration and the number of molecules will be represented.

The remainder of this paper is classified as follows: In section II the communication system model is introduced. The proposed detector presented in section III. The numerical results and comparison to common detectors provided in section IV. Finally, in section V, the paper is concluded.

II. System Model

As depicted in Fig. 1, the MC model that is studied in this paper consists of two NMs, one of which represented as Transmitter Nano-machine (TN) and the other, represented as Receiver Nano-machine (RN). The transmitter is located at the origin of an infinite three-dimensional fluid environment, d denotes the distance between the transmitter to the center of the receiver whose radius is r. At the beginning time of each transmission time slot, TN releases N molecules to represent the transmission of a symbol. The diffusion process is represented as Brownian motion and therefore is simulated as a three-dimensional random walk. Once a molecule arrives at the receiver, it will be removed from the medium. The averaged Channel impulse response of MC system for distances $25\mu m$ and $35\mu m$ is depicted in Fig. 2. At time $t = 0$, 1000 molecules

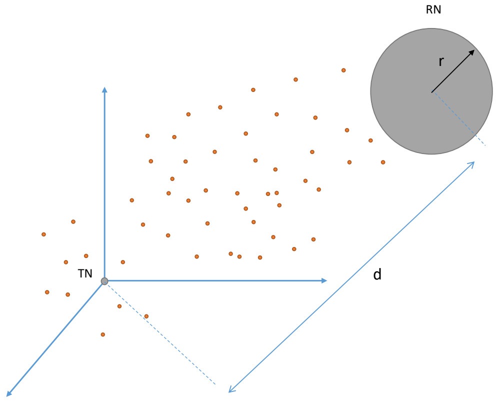


Fig. 1. Diffusion-based molecular communication system model.

is released by transmitter, which represent bit 1. The hitting probability function, which described as a probability per unit of time for a molecule to arrive at RN, is an inverse Gaussian function [11]:

$$f_{hit}(t) = \frac{r(d-r)}{d\sqrt{4\pi Dt^3}} \exp - \frac{(d-r)^2}{4Dt} \tag{1}$$

f_{hit} can be interpreted as the impulse response of a diffusion channel. In (1) D denotes the diffusion coefficient, which is given by:

$$D = \frac{k_b T}{6\pi\eta r} \tag{2}$$

Where k_b is the Boltzmann constant, T is the temperature, η is the viscosity of the fluid medium, and r is the radius of the molecule. Therefore, the fraction of molecules absorbed by receiver until time t, can be derived by intergrating f_{hit}:

$$F_{hit}(t) \;=\; \int\limits_0^t f_{hit}(x)dx \;=\; \frac{r}{d} erfc\left(\frac{d-r}{\sqrt{4Dt}}\right) \tag{3}$$

For simplicity, we assume that the radius for all molecules are equal so that the diffusion coefficients are the same. Furthermore, the probability for a molecule to reach RN during the ith bit duration $[iT_b, (i+1)T_b]$ after release can be computed as :

$$p_i = F_{hit}\left((i+1)T_b\right) - F_{hit}\left(iT_b\right) \tag{4}$$

The scenario that a single molecule arrives the RN during certain bit duration can be represented by a bernoulli experiment with two possible events of either hitting RN or not. For n molecules released at the same time under the assumption that molecules propagate independently and not change the hitting probability at RN, the n Bernoulli experiments can be described by the binomial distribution

$$y_i \propto \; B\left(n, p_i\right) \tag{5}$$

where y_i is the number of received molecules during the ith bit duration after their release. When transmitting a bit sequence, the number of received molecules at time instant k is summation over all y_i :

$$y[k] = \sum_{i=0}^{I} B\left(x[k-i], p_i\right) \tag{6}$$

here i is the length of channel memory and $x[k]$ is the OOK modulated symbol :

$$x(k) = \begin{cases} 0 & \text{if } u[k] = 0, \\ \text{N} & \text{if } u[k] = 1. \end{cases} \tag{7}$$

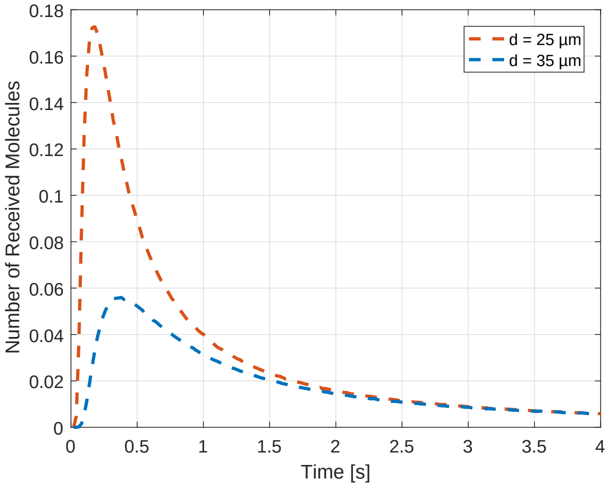


Fig. 2. Channel impulse response of molecular communication system. The environment examined here is the same as the one studied in section IV.

Moreover, RN performs energy detection and assumed to be a perfectly absorbing receiver. Noise sources that considered throughout this paper will be noise from the random propagation of the molecules, ISI molecules and background noise. In the first stage of the communication process, the TN encodes information symbols into the concentration of the molecules. One type of molecule and on-off keying modulation is utilized throughout this paper. Therefore, at the beginning of the bit interval to transmit bit '1', TN releases a certain amount of molecules and to transmit bit '0', TN releases no molecules. Where $x[k]$ stands for the kth data bit and $s(k)$ is corresponding released signal. The RN accumulate the number of received molecules from each symbol duration and determines incoming messages by comparing it to a threshold. As already mentioned, the design of threshold can be divided in two categories, fixed threshold [7] or adaptive threshold [11]. In fixed threshold detector (FTD), the optimal threshold is found based on minimizing BER and stays constant throughout the communication process. In adaptive threshold detector (ATD), the designed threshold varies depending on previously received symbols.

III. Proposed Detector

In this section, we propose the adaptive weighted threshold detector (AWTD). The idea of an adaptive threshold is inspired by the phenomenon of short-term synaptic plasticity [11]. Because of the limitation on NMs, it is necessary to introduce a simple and low-complex detection method for molecular communication. The concept of AWTD is motivated from the forgetting factor in RLS algorithm, in which we assign weights to two previously received symbols and compare it to the number of received molecules in current symbol. So, the decision rule can be described as :

$$y(k) = \begin{cases} 1 & \text{if } r(k) > \alpha r(k-1) + \alpha^2 r(k-2), \\ 0 & \text{if } r(k) \leq \alpha r(k-1) + \alpha^2 r(k-2). \end{cases} \quad (8)$$

Where $y(k)$ stands for the kth decoded bit, $r(k)$ as the number of molecules arrives within kth symbol period and α is the weight we assign. The AWTD detector decodes the kth bit by comparing the number of received molecules during current bit period with energy of two previous symbols with a weight assigned to each one. The optimal value of α determined by means of minimizing BER.

Since our simulation results show that the value of α is independent of system parameters like distance and symbol duration. In many applications that require the deployment of mobile NMs, when only transmitter moving or in channel condition when both of transmitter and receiver moving randomly, AWTD is the reasonable approach for detection. On the other hand, since AWTD do not need any knowledge about channel conditions, in channel models with time varying diffusion coefficient, it is the detection method that we can rely on. Other detection methods, like ML sequence detector or optimal threshold detector, are based on the knowledge of the channel characteristics. Obtaining channel information introduce an additional complexity and overhead. Our proposed detector is independent of channel knowledge and adapts its threshold only with respect to two previously symbols. Consequently any training or generation of test statistics are not required. So the memory and computational requirements of AWTD will be low, which can simply be implemented in molecular communication systems. More discussion will be represented in section IV.

IV. Numerical Results

In this section, we conduct simulation to determine the performance of the detector described in this paper. The impact of distance, number of transmitted molecules and symbol duration on BER will be studied in this section. We consider the system parameters that are summarized in Table I, Which is the same one applied in [11].

In Fig. 3, the BER for different values of α is shown. As already mentioned, the value of α is independent of other system parameters. Consequently, we considered three different scenarios to show this independency. With the help of these results, we take the optimal value as $\alpha = 0.6$.

TABLE I
Simulation System Parameters

Parameter	Value	unit
RX Radius	4.5	$[\mu m]$
Diffusion coeff.	4.37×10^{-10}	$[m^2/s]$
Distance	$25, 30, 35, 40, 45$	$[\mu m]$
T$_s$	$0.25, 0.5, 0.75, 1, 1.25, 1.5$	$[s]$
Sampling period	0.05	$[s]$
Molecules Released	$100, 500, 1000, 2500, 5000, 7500, 10000$	-
Number of Realizations	1000	-
Sequence Length	100	-

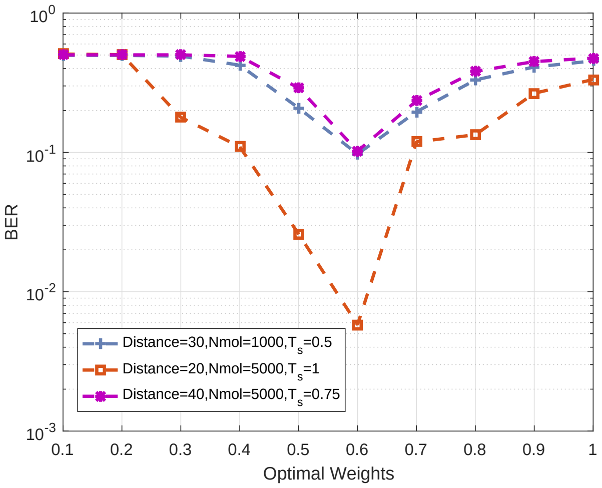


Fig. 3. Optimal weight versus BER.

In Fig. 4 the performance of detectors as a function of the number of transmitted molecules is investigated. For scenario under investigation number of molecules is varying from $N_{mol} = 100$ to $N_{mol} = 10000$ with fixed $d = 30\mu m$ and $t_s = 0.75s$. Simulation results show that FTD algorithm receives highest BER, this is caused by the fact that in contrast to adaptive threshold methods which utilized ISI effect, the FTD method struggle with higher distance and eventually higher ISI. As can be seen, both adaptive methods achieve better performance when the number of molecules increased, while our proposed detector AWTD outperforms ATD algorithm over the whole range of molecule numbers.

In Fig. 5 the impact of symbol duration on BER has been illustrated. The symbol duration is varying from $t_s = 0.25s$ to $t_s = 1.5s$ with fixed $N_{mol} = 1000$ and $d = 30\mu m$. We observe that FTD algorithm obtain lower BER when the symbol duration is increased and consequently ISI decreased. With referring to Fig. 5, our proposed detector achieves better performance for T_s under $1s$ in compare to FTD. Comparing to ATD, our detector nearly archives same BER for $t_s < 0.6s$ and outperform ATD for $t_s > 0.6s$.

As shown in section II, distance have a large impact on channel impulse response. As we observer in Fig. 2, the peak value for $d = 25\mu m$ happens earlier than $d = 35\mu m$. So the peak value and consequently the number of received molecules strictly depends on the distance between the transmitter and

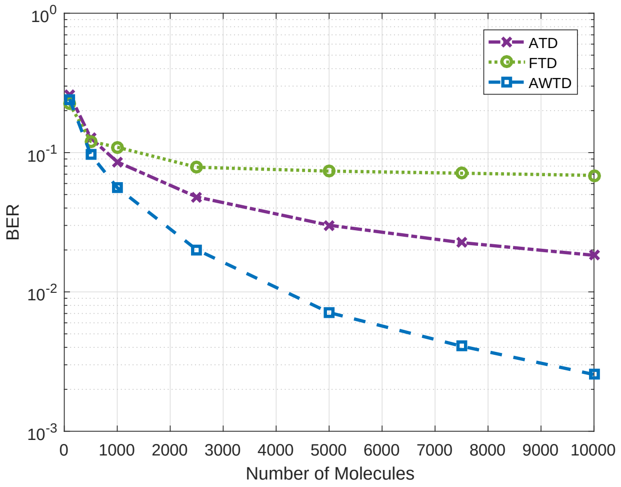


Fig. 4. Number of transmitted molecules versus BER. Distance and t_s are fixed at $d = 30\mu m$ and $t_s = 0.75s$, respectively.

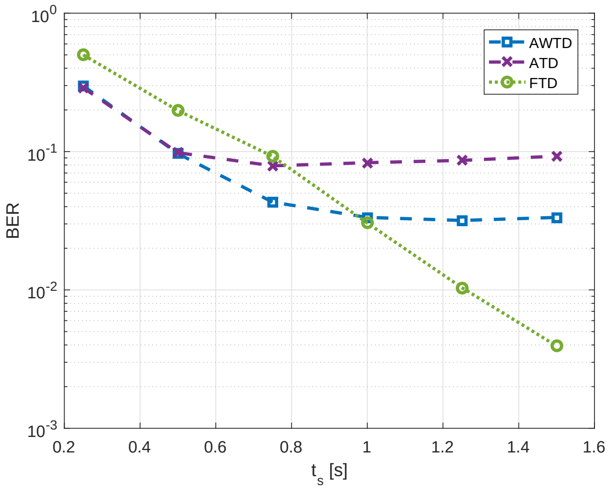


Fig. 5. Impact of Symbol duration on BER. N_{mol} and distance are fixed at 1000 and $30\mu m$, respectively.

the receiver. In Fig. 6 the impact of distance on BER has been illustrated. The distance is varying from $d = 25\mu m$ up to $d = 45\mu m$, with fixed $N_{mol} = 1000$ and $t_s = 0.75s$. For scenario under investigation, the new proposed algorithm outperforms ATD for distances under $d = 35\mu m$. In comparison to FTD, FTD achieves better performance for distances lower than $d = 25\mu m$ and AWTD more precisely is superior to FTD for distances higher than 30 μm.

V. Conclusion

A diffusion-based molecular communication system has been investigated. Specifically a system with single transmitter and single receiver with OOK modulation has been analyzed. Our proposed AWTD scheme will utilized the two previous symbols for evaluation the optimal threshold. The impact of number of molecules, distance and symbol duration on BER has been studied. It has been proved that proposed detector outperforms ATD and FTD for whole range on molecules. It achieves lower BER compare to ATD for specific symbol durations distances. Future works include theoretical derivation of optimal weight and BER. Finally, investigation of other modulation schemes and different scenarios can be analyzed for new proposed detector.

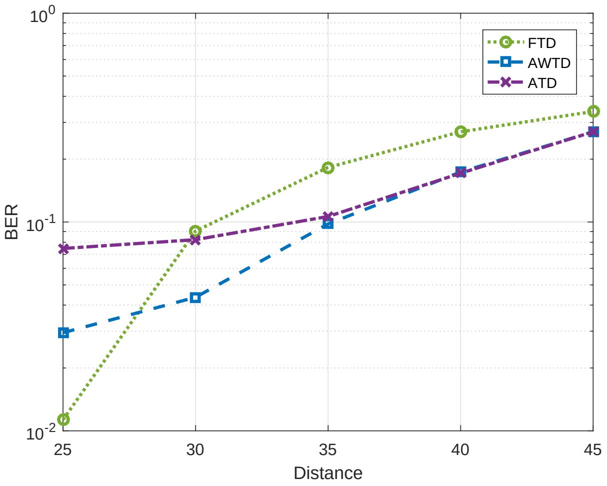


Fig. 6. Impact of distance on BER. N_{mol} and t_s are fixed at 1000 and $0.75s$, respectively.

References

[1] M. Pierobon and I. F. Akyildiz, "A physical end-to-end model for molecular communication in nanonetworks," *IEEE Journal on Selected Areas in Communications*, vol. 28, pp. 602–611, May 2010.

[2] G. Ardeshiri, A. Jamshidi, and A. Keshavarz-Haddad, "Performance analysis of decode and forward relay network in diffusion based molecular communication," in *Electrical Engineering (ICEE), 2017 Iranian Conference on*, pp. 1992–1997, IEEE, 2017.

[3] P. Akhkandi, A. Keshavarz-Haddad, and A. Jamshidi, "A new channel code for decreasing inter-symbol-interference in diffusion based molecular communications," in *Telecommunications (IST), 2016 8th International Symposium on*, pp. 277–281, IEEE, 2016.

[4] T. Nakano, A. W. Eckford, and T. Haraguchi, *Molecular Communication*. Cambridge University Press, 2013.

[5] R. Bigharaz, A. Jamshidi, and A. Keshavarz-Haddad, "A realistic receiver model for neuro-spike communication," in *Telecommunications (IST), 2016 8th International Symposium on*, pp. 239–244, IEEE, 2016.

[6] A. Noel, K. C. Cheung, and R. Schober, "Optimal receiver design for diffusive molecular communication with flow and additive noise," *IEEE Transactions on NanoBioscience*, vol. 13, pp. 350–362, Sept 2014.

[7] B. Atakan and O. B. Akan, "An information theoretical approach for molecular communication," in *2007 2nd Bio-Inspired Models of Network, Information and Computing Systems*, pp. 33–40, Dec 2007.

[8] I. Llatser, A. Cabellos-Aparicio, M. Pierobon, and E. Alarcon, "Detection techniques for diffusion-based molecular communication," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 726–734, December 2013.

[9] R. Mosayebi, H. Arjmandi, A. Gohari, M. Nasiri-Kenari, and U. Mitra, "Receivers for diffusion-based molecular communication: Exploiting memory and sampling rate," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 2368–2380, Dec 2014.

[10] D. Kilinc and O. B. Akan, "Receiver design for molecular communication," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 705–714, December 2013.

[11] M. Damrath and P. A. Hoeher, "Low-complexity adaptive threshold detection for molecular communication," *IEEE Transactions on NanoBioscience*, vol. 15, pp. 200–208, April 2016.

Simultaneous Information and Energy Transmission in Gaussian Interference Channels

Nizar Khalfet and Samir M. Perlaza
INRIA, Université de Lyon and INSA de Lyon
Laboratoire CITI, F-69621 Villeurbanne, France
Email: {nizar.khalfet, samir.perlaza}@inria.fr

***Abstract*—In this paper, the fundamental limits of simultaneous information and energy transmission in the two-user Gaussian interference channel (G-IC) are fully characterized. More specifically, an achievable and converse region in terms of information and energy transmission rates (in bits per channel use and energy-units per channel use, respectively) are presented. The achievable region is obtained using a combination of rate splitting, power-splitting, common randomness and superposition coding. Finally, the converse region is obtained using some of the existing outer bounds on the information transmission rates, as well as a new outer bound on the energy transmission rate.**

I. Introduction

Over the last decade, energy harvesting has been considered as a promising technology with great potential for green technologies, low voltage wearable electronics, and wireless devices. This aligns with the fact that receivers can simultaneously extract information and energy from radio-frequency signals [7], [9]. However, recent research has shown that energy and information transmission are often conflicting tasks. That is, there exists a trade-off between the information transmission rate and the energy transmission rate. This trade-off is observed in point-to-point channels [5] and multi-user channels such as the multiple access channel [1], [4] and the interference channel [8]. However, very little is known about this trade-off in other multi-user channels with energy harvesting. More importantly, very little is known about the fundamental limits of multi-user SIET.

This paper focuses on the Gaussian interference channel with an external energy harvester (EH). This channel models two point-to-point links subject to mutual interference, where both transmitters are engaged with transmitting information to their intended receiver while jointly providing a minimum energy rate at the EH. The fundamental limits of this channel are thoroughly studied. An achievable and converse region in terms of information and energy transmission rates are identified. The achievable region is obtained using a combination of rate splitting, power-splitting, common randomness, and superposition coding. The converse region is obtained using some of the existing outer bounds on the information transmission rates, as well as a new outer bound on the energy transmission rate.

II. Gaussian Interference Channel with Energy Harvester

Consider the Gaussian interference channel with a non-colocated energy harvester depicted in Fig. 1. Transmitter i, with $i \in \{1, 2\}$, aims to execute two tasks: (a) an information transmission task and (b) an energy transmission task.

A. Information Transmission Task

From the information transmission standpoint, the goal of transmitter i is to convey message index $W_i \in \mathcal{W}_i$ to receiver i using N channel input symbols $X_{i,1}, X_{i,2}, \ldots, X_{i,N}$. The channel coefficient from transmitter k to receiver i, with $k \in \{1, 2\}$, is denoted by $h_{ik} \in \mathbb{R}_+$. For channel use n, input symbol $X_{i,n}$ is observed at receiver i in addition to the interference produced by the symbol $X_{j,n}$ sent by transmitter j, with $j \in \{1, 2\}\backslash\{i\}$, and a real additive Gaussian noise $Z_{i,n}$ with zero mean and variance σ_i^2. Hence, the channel output at receiver i during channel use n, denoted by $Y_{i,n}$, is

$$Y_{i,n} = h_{ii}X_{i,n} + h_{ij}X_{j,n} + Z_{i,n}. \tag{1}$$

At each channel use n, the symbol $X_{i,n}$ sent by transmitter i depends on the message index W_i and a randomly generated index $\Omega \in \mathbb{N}$. The random index Ω is assumed to be independent of both W_1 and W_2 and known by all transmitters and receivers. Let $f_{i,n}^{(N)} : \mathcal{W}_i \times \mathbb{N} \rightarrow \mathbb{R}$ be the encoding function at channel use n, such that for all $n \in \{1, 2, \ldots, N\}$:

$$X_{i,n}{=}f_{i,n}^{(N)}(W_i, \Omega). \tag{2}$$

Channel input symbols $X_{i,1}, X_{i,2}, \ldots, X_{i,N}$ are subject to an average power constraint of the form

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{E}[X_{i,n}^2] \leq P_i, \tag{3}$$

where the expectation is taken with respect to W_i and Ω, which follow uniform probability distributions over their corresponding supports. Receiver i observes the channel outputs $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,N}$ and uses a decoding function

$$\phi_i^{(N)}\colon \mathbb{N} \times \mathbb{R}^N \rightarrow \{1, 2, \ldots, 2^{R_i}\}, \tag{4}$$

to get an estimate $\widehat{W_i}^{(N)} = \phi_i^{(N)}(\Omega, Y_{i,1}, Y_{i,2}, \ldots, Y_{i,N})$ of the transmitted message W_i. The information rate at receiver i is denoted by R_i and it is defined by:

$$R_i = \frac{\log_2 |\mathcal{W}_i|}{N}, \tag{5}$$

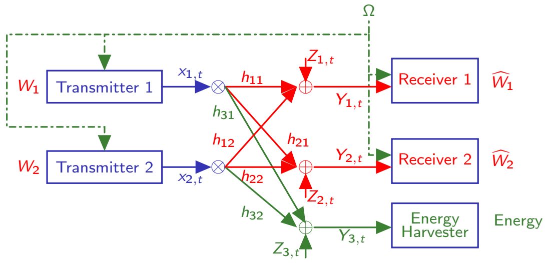


Fig. 1: Two-user Gaussian interference channel with a non-colocated energy harvester.

in bits per channel use. The decoding error probability is given by

$$P_{\mathrm{e}}^{(N)}(R_1, R_2) \\ = \max\left(\Pr\left(\widehat{W_1}^{(N)} \neq W_1\right), \Pr\left(\widehat{W_2}^{(N)} \neq W_2\right)\right). \tag{6}$$

The signal to noise ratio (SNR) at receiver i is denoted by

$$\mathrm{SNR}_i = \frac{h_{ii}^2 P_i}{\sigma_i^2}. \tag{7}$$

Similarly, the interference to noise ratio (INR) at receiver i is denoted by

$$\mathrm{INR}_i = \frac{h_{ij}^2 P_j}{\sigma_i^2}, \quad \text{for } j \neq i. \tag{8}$$

B. Energy Transmission Task

Let $h_{3i} \in \mathbb{R}_+$ be the channel coefficient from transmitter i to the energy harvester (EH). The symbols sent by transmitter 1 and 2 to the EH are subject to an additive Gaussian noise $Z_{3,n}$ with zero mean and variance σ_3^2 . More specifically, the channel output at the EH, denoted by $Y_{3,n}$, is

$$Y_{3,n} = h_{31} X_{1,n} + h_{32} X_{2,n} + Z_{3,n}. \tag{9}$$

The SNR of transmitter i at the EH is denoted by

$$\mathrm{SNR}_{3i} = \frac{h_{3i}^2 P_i}{\sigma_3^2}. \tag{10}$$

Let $b \geqslant 0$ denote the minimum average energy rate that must be guaranteed at the input of the EH. Let $B^{(N)}$ be the average energy transmission rate (in energy-units per channel use) at the end of N channel uses. That is,

$$B^{(N)} \triangleq \frac{1}{N} \sum_{n=1}^{N} Y_{3,n}^2. \tag{11}$$

Note that the maximum average energy rate $B_{\max}$ is

$$B_{\max} = \sigma_3^2 \left(1 + \mathrm{SNR}_{31} + \mathrm{SNR}_{32} + 2\sqrt{\mathrm{SNR}_{31}\mathrm{SNR}_{32}}\right). \tag{12}$$

From the energy transmission standpoint, the goal of both transmitters is to jointly guarantee that the average energy rate $B^{(N)}$ is greater than a given operational energy transmission rate B that must satisfy

$$b \leqslant B \leqslant B_{\max}. \tag{13}$$

The probability of energy outage, given an average energy rate B, is defined as follows:

$$P_{\mathrm{outage}}^{(N,\epsilon)}(B) \triangleq \Pr\left[B^{(N)} < B - \epsilon\right], \tag{14}$$

for some $\epsilon > 0$.

C. Simultaneous Information and Energy Transmission (SIET)

Given a minimum energy rate b to be satisfied at the EH, the system is said to be operating at the information-energy rate triplet $(R_1, R_2, B) \in \mathbb{R}_+^3$ when both transmitter-receiver pairs use a transmit-receive configuration such that: (i) reliable communication at information rates R_1 and R_2 is ensured; and (ii) reliable energy transmission at energy rate B is ensured. A formal definition is given below.

Definition 1 (Achievable Rates): The triplet $(R_1, R_2, B) \in \mathbb{R}_+^3$ is achievable if for all $i \in \{1, 2\}$, there exists a sequence of encoding functions $f_{i,1}^{(N)}, f_{i,2}^{(N)}, \ldots, f_{i,N}^{(N)}$, and the decoding functions $\phi_1^{(N)}$ and $\phi_2^{(N)}$, such that both the average error probability $P_{\mathrm{e}}^{(N)}(R_1, R_2)$ and the energy-outage probability $P_{\mathrm{outage}}^{(N,\epsilon)}(B)$ tend to zero as the block-length N tends to infinity. That is,

$$\limsup_{N\to\infty} P_{\mathrm{e}}^{(N)} = 0, \text{ and} \tag{15}$$

$$\limsup_{N\to\infty} P_{\mathrm{outage}}^{(N,\epsilon)} = 0. \tag{16}$$

Using Definition 1, the fundamental limits of simultaneous information and energy transmission in the Gaussian interference channel can be described by the information-energy capacity region, defined as follows.

Definition 2 (Information-Energy Capacity Region): The information-energy capacity region given a minimum energy rate b, denoted by $\mathcal{E}_b$, corresponds to the closure of all achievable information-energy rate triplets (R_1, R_2, B).

III. Main Results

The main result consists of a description of the information-energy capacity region $\mathcal{E}_b$, for a given $b \geqslant 0$. The following sections show that the information-energy capacity region $\mathcal{E}_b$, with b any positive real number, is approximated by the regions $\underline{\mathcal{E}}_b$ (Theorem 1), which represents an information-energy achievable region, and $\overline{\mathcal{E}}_b$ (Theorem 2), which represents an information-energy converse region.

A. An Achievable Region

The following theorem describes a set of rate-tuples that are achievable (Definition 1).

Theorem 1: Let b be a fixed positive real. Then, the information-energy capacity region $\mathcal{E}_b$ contains all the rate *tuples* (R_1, R_2, B) *that satisfy for all* $i \in \{1,2\}$ *and* $j \in \{1,2\}\backslash\{i\}$*:*

$$R_i \leq \frac{1}{2}\log\left(1 + \frac{(1-\lambda_{ie})\mathrm{SNR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right), \tag{17a}$$

$$\begin{aligned} R_1 + R_2 \leq & \frac{1}{2}\log\left(\frac{1+(1-\lambda_{ie})\mathrm{SNR}_1 + (1-\lambda_{je})\mathrm{INR}_1}{1+\lambda_{jp}\mathrm{INR}_1}\right), \\ & + \frac{1}{2}\log\left(1 + \frac{\lambda_{jp}\mathrm{SNR}_j}{1+\lambda_{ip}\mathrm{INR}_j}\right), \end{aligned} \tag{17b}$$

$$\begin{aligned} R_1 + R_2 \leq & \frac{1}{2}\log\left(\frac{1+\lambda_{1p}\mathrm{SNR}_1 + (1-\lambda_{2e})\mathrm{INR}_1}{1+\lambda_{2p}\mathrm{INR}_1}\right) \\ & + \frac{1}{2}\log\left(\frac{1+\lambda_{2p}\mathrm{SNR}_2 + (1-\lambda_{1e})\mathrm{INR}_2}{1+\lambda_{1p}\mathrm{INR}_2}\right), \end{aligned} \tag{17c}$$

$$\begin{aligned} 2R_i + R_j \leq & \frac{1}{2}\log\left(\frac{1+(1-\lambda_{ie})\mathrm{SNR}_i + (1-\lambda_{je})\mathrm{INR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right) \\ & + \frac{1}{2}\log\left(\frac{1+\lambda_{jp}\mathrm{SNR}_j + (1-\lambda_{ie})\mathrm{INR}_j}{1+\lambda_{ip}\mathrm{INR}_j}\right) \\ & + \frac{1}{2}\log\left(1 + \frac{\lambda_{ip}\mathrm{SNR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right), \end{aligned} \tag{17d}$$

$$\begin{aligned} b \leq B \leq \sigma_3^2 \Big(& 1 + \mathrm{SNR}_{31} + \mathrm{SNR}_{32} \\ & + 2\sqrt{\mathrm{SNR}_{31}\mathrm{SNR}_{32}}\sqrt{\lambda_{1e}\lambda_{2e}}\Big), \end{aligned} \tag{17e}$$

with $(\lambda_{ip}, \lambda_{ie}) \in [0,1]^2$ *such that* $\lambda_{ip} + \lambda_{ie} \leq 1$.

Proof: The sketch of proof of Theorem 1 is presented in the following section. ■

B. Sketch of Proof of Achievability

The achievability scheme used to obtain Theorem 1 is built upon random coding arguments using rate-splitting [6], superposition coding [2], common randomness and power-spliting [1]. Let $W_i \in \{1, 2\ldots, 2^{NR_i}\}$ and $\Omega \in \{1, 2\ldots, 2^{NR_E}\}$ be respectively the message index and the common random index at transmitter i. Following a rate-splitting argument, the index W_i is divided into two sub-indices $W_{i,P} \in \{1, 2\ldots, 2^{NR_{i,P}}\}$ and $W_{i,C} \in \{1, 2\ldots, 2^{NR_{i,C}}\}$, where $R_{i,C} + R_{i,P} = R_i$. The message index $W_{i,C}$ must be decoded at both receivers, whereas the index $W_{i,P}$ must be decoded only at the intended receiver. This rate-splitting is reminiscent of the Han-Kobayashi scheme in [6].

Lemma 1: An achievable information rate pair (R_1, R_2) *satisfies the following inequalities, for all* $i \in \{1,2\}$ *and* $j \in \{1,2\}\backslash\{i\}$*:*

$$R_i \leq I(X_i; Y_i | U_j, V) \tag{18a}$$

$$R_1 + R_2 \leq I(X_i, U_j; Y_i | V) + I(X_j; Y_j | U_i, U_j, V) \tag{18b}$$

$$R_1 + R_2 \leq I(X_1, U_2; Y_1 | U_1, V) + I(X_2, U_1; Y_2 | U_2, V) \tag{18c}$$

$$\begin{aligned} 2R_i + R_j \leq & I(X_i, U_j; Y_i | V) + I(X_i; Y_i | U_i, U_j, V) \\ & + (X_j, U_i; Y_j | U_j, V), \end{aligned} \tag{18d}$$

for a given joint distribution $P_{VU_1U_2S_1S_2}(v, u_1, u_2, s_1, s_2)$ *that factorizes as* $P_V(v)$ $P_{U_1|V}(u_1|v)$ $P_{U_2|V}(u_2|v)$ $P_{S_1|U_1V}(s_1|u_1v)$ $P_{S_2|U_2V}(s_2|u_2v)$ *and* $X_i = \theta_i(V, U_i, S_i)$, *with* θ_1 *and* θ_2 *injective functions.*

Proof: The proof of Lemma 1 uses standard arguments of weak typicality and is omitted in this paper. ■

For all $k \in \{1,2\}$ and a fixed triplet $(\lambda_{kc}, \lambda_{kp}, \lambda_{ke}) \in [0,1]^3$ such that $\lambda_{kc} + \lambda_{kp} + \lambda_{ke} = 1$, consider the following random variables: $V \sim \mathcal{N}(0,1)$; $U_k \sim \mathcal{N}(0, \lambda_{kc})$; and $S_k \sim \mathcal{N}(0, \lambda_{kp})$, which are independent of each other. Let the channel input of transmitter k be

$$X_k = \sqrt{P_k}S_k + \sqrt{P_k}U_k + \sqrt{\lambda_{ke}P_k}V. \tag{19}$$

The choice of this input distribution yields

$$I(X_i; Y_i | U_j, V) = \frac{1}{2}\log\left(1 + \frac{(1-\lambda_{ie})\mathrm{SNR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right) \tag{20a}$$

$$I(X_i, U_j; Y_i | V) = \frac{1}{2}\log\left(\frac{1+(1-\lambda_{ie})\mathrm{SNR}_i + (1-\lambda_{je})\mathrm{INR}_i}{1+\lambda_{jp}\mathrm{INR}_1}\right), \tag{20b}$$

$$I(X_j; Y_j | U_i, U_j, V) = \frac{1}{2}\log\left(1 + \frac{\lambda_{jp}\mathrm{SNR}_j}{1+\lambda_{ip}\mathrm{INR}_j}\right) \tag{20c}$$

$$\begin{aligned} I(X_i, U_j; Y_i | U_i, V) &= \frac{1}{2}\log\left(\frac{1+\lambda_{ip}\mathrm{SNR}_i + (1-\lambda_{je})\mathrm{INR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right), \\ I(X_i, U_j; Y_i | U_i, V) &= \frac{1}{2}\log\left(\frac{1+\lambda_{ip}\mathrm{SNR}_i + (1-\lambda_{je})\mathrm{INR}_i}{1+\lambda_{jp}\mathrm{INR}_i}\right), \end{aligned} \tag{20d}$$

$$I(X_j, U_i; Y_j | U_j, V) = \frac{1}{2}\log\left(\frac{1+\lambda_{jp}\mathrm{SNR}_j + (1-\lambda_{ie})\mathrm{INR}_j}{1+\lambda_{ip}\mathrm{INR}_2}\right). \tag{20e}$$

Finally, plugging (20) into (18) completes the proof of (17a) - (17d).

The average received energy rate $\bar{B}$ achieved by using the input signals in (19) is given by

$$\begin{aligned} \bar{B} =& E[Y_{3,n}^2] \\ =& h_{31}^2 E[X_{1,n}^2] + h_{32}^2 E[X_{2,n}^2] + 2h_{31}h_{32}E[X_{1,n}X_{2,n}] + \sigma_3^2 \\ \leq& h_{31}^2 P_1 + h_{32}^2 P_2 + 2h_{31}h_{32}\sqrt{\lambda_{1e}P_1\lambda_{2e}P_2} + \sigma_3^2 \\ =& \sigma_3^2\Big(1 + \mathrm{SNR}_{31} + \mathrm{SNR}_{32} + 2\sqrt{\mathrm{SNR}_{31}\mathrm{SNR}_{32}}\sqrt{\lambda_{1e}\lambda_{2e}}\Big). \end{aligned}$$

From the weak law of large numbers, it holds that $\forall \epsilon > 0$,

$$\lim_{N\to\infty} \Pr\left(B^{(N)} < \bar{B} - \epsilon\right) = 0. \tag{21}$$

From (21), it holds that for any energy rate B that satisfies $0 < B \leq \bar{B}$, it holds that

$$\lim_{N\to\infty} \Pr\left(B^{(N)} < B - \epsilon\right) = 0, \tag{22}$$

which proves (17e) and completes the sketch of proof.

C. Converse

The following theorem describes a converse region denoted by $\bar{\mathcal{E}}_b$.

Theorem 2: Let b *be a fixed positive real. Then, the information-energy capacity region* $\mathcal{E}_b$ *is contained into the set* *of all the rate tuples* (R_1, R_2, B) *that satisfy for all* $i \in \{1,2\}$ *and* $j \in \{1,2\}\backslash\{i\}$*:*

$$R_i \leq \frac{1}{2}\log(1+\beta_i \mathrm{SNR}_i), \tag{23a}$$

$$\begin{aligned} R_1+R_2 \leq & \frac{1}{2}\log(1+\beta_i\mathrm{SNR}_i+\beta_j\mathrm{INR}_i) \\ & +\frac{1}{2}\log(1+\frac{\beta_j\mathrm{SNR}_j}{1+\beta_j\mathrm{INR}_i}), \end{aligned} \tag{23b}$$

$$\begin{aligned} R_1+R_2 \leq & \frac{1}{2}\log\left(1+\frac{\beta_1\mathrm{SNR}_1+\beta_2\mathrm{INR}_1+\beta_1\beta_2\mathrm{INR}_1\mathrm{INR}_2}{1+\beta_1\mathrm{INR}_2}\right) \\ & +\frac{1}{2}\log\left(1+\frac{\beta_2\mathrm{SNR}_2+\beta_1\mathrm{INR}_2+\beta_1\beta_2\mathrm{INR}_1\mathrm{INR}_2}{1+\beta_2\mathrm{INR}_1}\right), \end{aligned} \tag{23c}$$

$$\begin{aligned} 2R_i+R_j \leq & \frac{1}{2}\log(1+\frac{\beta_i\mathrm{SNR}_i}{1+\beta_i\mathrm{INR}_j}) \\ & +\frac{1}{2}\log(1+\beta_i\mathrm{SNR}_i+\beta_j\mathrm{INR}_i) \\ & +\frac{1}{2}\log\left(1+\frac{\beta_j\mathrm{SNR}_j+\beta_i\mathrm{INR}_j+\beta_i\beta_j\mathrm{INR}_i\mathrm{INR}_j}{1+\beta_j\mathrm{INR}_i}\right), \end{aligned} \tag{23d}$$

$$\begin{aligned} b \leq B \leq \sigma_3^2 \Big(& 1+\mathrm{SNR}_{31}+\mathrm{SNR}_{32} \\ & +2\sqrt{(1-\beta_1)\mathrm{SNR}_{31}(1-\beta_2)\mathrm{SNR}_{32}}\Big), \end{aligned} \tag{23e}$$

with $(\beta_1, \beta_2) \in [0,1]^2$.

Proof: The sketch of the proof of Theorem 2 is presented in the following section. ■

D. Sketch of Proof of the Converse

Fix an information-energy rate triplet (R_1, R_2, B) achievable with a given coding scheme (Definition 1). Denote by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ the channel inputs resulting from transmitting the independent message W_1 and W_2 using such coding scheme. Denote by $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ the corresponding channel outputs. Define the following random variables:

$$\begin{aligned} U_1 &= h_{21}X_1 + Z_2^{'}, \text{ and} \\ U_2 &= h_{12}X_2 + Z_1^{'}, \end{aligned}$$

where, $Z_1^{'}$ and $Z_2^{'}$ are real Gaussian random variables independent of each other with zero means and variances σ_1^2 and σ_2^2, respectively. Using assumption (15) and Fano's inequality, it follows that the information rates R_1 and R_2 must satisfy the following inequalities:

$$NR_i \leqslant \sum_{n=1}^{N} h(Y_{i,n}|X_{j,n}) - Nh(Z_i) + \mathrm{o}(N) \tag{24a}$$

$$\begin{aligned} N(R_1+R_2) \leqslant & \sum_{n=1}^{N} h(Y_{i,n}) + \sum_{n=1}^{N} h(Y_{j,n}|U_{j,n}, X_{i,n}) \\ & -Nh(Z_j) - Nh(Z_i^{'}) + \mathrm{o}(N) \end{aligned} \tag{24b}$$

$$\begin{aligned} N(R_1+R_2) \leqslant & \sum_{n=1}^{N} h(Y_{1,n}|U_{1,n}) + \sum_{n=1}^{N} h(Y_{2,n}|U_{2,n}) \\ & -Nh(Z_1^{'}) - Nh(Z_2^{'}) + \mathrm{o}(N) \end{aligned} \tag{24c}$$

$$\begin{aligned} N(2R_i+R_j) \leqslant & \sum_{n=1}^{N} h(Y_{i,n}) + \sum_{n=1}^{N} h(Y_{i,n}|U_{i,n}, X_{j,n}) \\ & + \sum_{n=1}^{N} h(Y_{j,n}|U_{j,n}) - N(h(Z_1)+h(Z_2) \\ & +h(Z_1^{'})+h(Z_2^{'})) + \mathrm{o}(N), \end{aligned} \tag{24d}$$

where $\frac{\mathrm{o}(N)}{N}$ tends to zero as N tends to infinity. Using assumption (16), for a given $\epsilon_N > 0$ and an $\eta > 0$, there exist $N_0(\eta)$ such that for any $N \geq N_0(\eta)$ it holds that

$$\Pr\left(B^{(N)} < B - \epsilon_N\right) < \eta. \tag{25}$$

Equivalently,

$$\Pr\left(B^{(N)} \geq B - \epsilon_N\right) \geq 1 - \eta. \tag{26}$$

From Markov's inequality, the following holds:

$$(B-\epsilon_N)\Pr\left(B^{(N)} \geq B - \epsilon_N\right) \leq E[B^{(N)}]. \tag{27}$$

Combining (26) and (27) yields

$$(B-\epsilon_N)(1-\eta) \leq E[B^{(N)}], \tag{28}$$

which can be written as

$$(B-\delta_N) \leq E[B^{(N)}], \tag{29}$$

for some $\delta_N > \epsilon_N$ and a sufficiently large N. In the following, for all $n \in \mathbb{N}$, the bounds in (24) and (29) are evaluated assuming that the channel inputs $X_{1,n}$ and $X_{2,n}$ are independent random variables with mean and variance:

$$\mu_{i,n} \triangleq \mathrm{E}[X_{i,n}], \tag{30}$$

$$\gamma_{i,n}^2 \triangleq \mathrm{Var}[X_{i,n}]. \tag{31}$$

The input sequences must satisfy the input power constraint (3) which can be written for $i \in \{1,2\}$, as

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}[X_{i,n}^2] = \left(\frac{1}{N}\sum_{n=1}^{N}\gamma_{i,n}^2\right) + \left(\frac{1}{N}\sum_{n=1}^{N}\mu_{i,n}^2\right) \leqslant P_i. \tag{32}$$

Using these elements, the terms in the right-hand side of (24) can be upper-bounded as follows:

$$h(Y_{i,n}|X_{j,n}) \leqslant \frac{1}{2}\log\left(2\pi e(\sigma_i^2 + h_{ii}^2\gamma_{i,n}^2)\right), \tag{33a}$$

$$h(Y_{i,n}) \leqslant \frac{1}{2}\log\left(2\pi e(\sigma_i^2 + h_{ii}^2\gamma_{i,n}^2 + h_{ij}^2\gamma_{j,n}^2)\right), \tag{33b}$$

$$\begin{aligned} & h(Y_{i,n}|U_{i,n}, X_{j,n}) \leqslant \frac{1}{2}\log\left(1+\frac{\frac{h_{ii}^2\gamma_{i,n}^2}{\sigma_i^2}}{1+\frac{h_{ji}^2\gamma_{i,n}^2}{\sigma_j^2}}\right) \\ & +\frac{1}{2}\log(2\pi e\sigma_i^2\sigma_j^2), \text{ and} \end{aligned} \tag{33c}$$

$$\begin{aligned} & h(Y_{i,n}|U_{i,n}) \leqslant \frac{1}{2}\log(2\pi e\sigma_i^2\sigma_j^2) \\ & +\frac{1}{2}\log\left(1+\frac{\frac{h_{ii}^2\gamma_{i,n}^2}{\sigma_i^2}+\frac{h_{ij}^2\gamma_{j,n}^2}{\sigma_i^2}+\frac{\gamma_{i,n}^2\gamma_{j,n}^2h_{ij}^2h_{ji}^2}{\sigma_i^2\sigma_j^2}}{1+\frac{\gamma_{i,n}^2h_{ji}^2}{\sigma_j^2}}\right). \end{aligned} \tag{33d}$$

The expectation of the average received energy rate is given by

$$\mathrm{E}\left[B^{(N)}\right] = \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} Y_{3,n}^2\right] =$$
$$h_{31}^2\left(\frac{1}{N}\sum_{n=1}^{N}(\gamma_{1,n}^2+\mu_{1,n}^2)\right) + h_{32}^2\left(\frac{1}{N}\sum_{n=1}^{N}(\gamma_{2,n}^2+\mu_{2,n}^2)\right)$$
$$+2h_{31}h_{32}\frac{1}{N}\sum_{n=1}^{N}\mu_{1,n}\mu_{2,n} + \sigma_3^2. \tag{34}$$

Using Cauchy-Schwarz inequality, combining (29) and (34) yields the following upper-bound on the energy rate B:

$$B \leqslant \sigma_3^2 + \frac{h_{31}^2}{N}\sum_{n=1}^{N}(\gamma_{1,n}^2+\mu_{1,n}^2) + \frac{h_{32}^2}{N}\sum_{n=1}^{N}(\gamma_{2,n}^2+\mu_{2,n}^2)$$
$$+2h_{31}h_{32}\left(\frac{1}{N}\sum_{n=1}^{N}\mu_{1,n}^2\right)^{1/2}\left(\frac{1}{N}\sum_{n=1}^{N}\mu_{2,n}^2\right)^{1/2} + \delta_N. \tag{35}$$

Consider the following definitions, for all $i \in \{1,2\}$:

$$\mu_i^2 \triangleq \frac{1}{N}\sum_{n=1}^{N}\mu_{i,t}^2, \tag{36a}$$
$$\gamma_i^2 \triangleq \frac{1}{N}\sum_{n=1}^{N}\gamma_{i,n}^2, \text{ and} \tag{36b}$$
$$\beta_i \triangleq \frac{\gamma_i^2}{P_i}. \tag{36c}$$

Plugging (33) in (24) and after some manipulations using the definitions in (36) and using Jensen's inequality complete the sketch of the proof.

E. Approximation of the Information-Energy Capacity Region

Using the inner region $\underline{\mathcal{E}}_b$ and the outer region $\overline{\mathcal{E}}_b$, described respectively by Theorem 1 and Theorem 2, the information-energy capacity region $\mathcal{E}_b$ can be approximated according to the following theorem.

Theorem 3 (Approximation of $\mathcal{E}_b$): Let $\underline{\mathcal{E}}_b \subset \mathbb{R}_+^3$ and $\overline{\mathcal{E}}_b \subset \mathbb{R}_+^3$ be the sets of tuples (R_1, R_2, B) described by Theorem 1 and Theorem 2, respectively. Then,

$$\underline{\mathcal{E}}_b \subset \mathcal{E}_b \subset \overline{\mathcal{E}}_b, \tag{37}$$

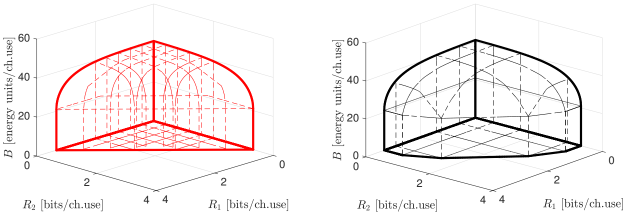
and for all $(R_1, R_2, B) \in \overline{\mathcal{E}}_b$ it follows that $((R_1 - 1/2)^+, (R_2 - 1/2)^+, (B - \frac{B_{\max}}{2})^+) \in \underline{\mathcal{E}}_b$.

Proof: Following similar steps as in [3], it can be shown that for all $(R_1, R_2, 0) \in \overline{\mathcal{E}}_b$ it follows that $((R_1 - 1/2)^+, (R_2 - 1/2)^+, 0) \in \underline{\mathcal{E}}_b$. Note also that for all $(R_1, R_2, B) \in \overline{\mathcal{E}}_b$ and for all $(R_1, R_2, B') \in \underline{\mathcal{E}}_b$, there always exists a tuple $(\beta_1, \beta_2, \lambda_{1e}, \lambda_{2e})$ such that: $\frac{B-B'}{B_{\max}}$

$$= \frac{2|h_{31}||h_{32}|\sqrt{P_1P_2}\left(\sqrt{(1-\beta_1)(1-\beta_2)} - \sqrt{\lambda_{1e}\lambda_{2e}}\right)}{\sigma_3^2 + h_{31}^2P_1 + h_{32}^2P_2 + 2h_{31}h_{32}\sqrt{P_1P_2}}$$
$$\leq \frac{2\sqrt{\mathrm{SNR}_{31}\mathrm{SNR}_{32}}}{1+4\sqrt{\mathrm{SNR}_{31}\mathrm{SNR}_{32}}}$$
$$\leq \frac{1}{2},$$

which completes the proof. ■

IV. Example



(a) 3-D representation of $\underline{\mathcal{E}}_b$.

(b) 3-D representation of $\overline{\mathcal{E}}_b$.

Fig. 2: 3-D representation of $\underline{\mathcal{E}}_b$ and $\overline{\mathcal{E}}_b$.

Consider a Gaussian interference channel with an external EH with parameters $\mathrm{SNR}_1 = \mathrm{SNR}_2 = 20$ dB, $\mathrm{INR}_1 = \mathrm{INR}_2 = \mathrm{SNR}_{31} = \mathrm{SNR}_{32} = 10$ dB and $\sigma_3^2 = 1$.

Figure 2a and Figure 2b show $\underline{\mathcal{E}}_b$ and $\overline{\mathcal{E}}_b$, respectively, with $b = 0$. Note that for all $B \in [0, 1 + \mathrm{SNR}_{31} + \mathrm{SNR}_{32}]$, transmitting information with independent codewords is enough to satisfy the energy rate constraints. This implies that $\beta_1 = \beta_2 = 1$ is optimal in this regime. Alternatively, for all $B \in [1 + \mathrm{SNR}_{31} + \mathrm{SNR}_{32}, B_{\max}]$, transmitters deal with trade-off between the information and energy rate. Increasing B reduces the information region and makes the information-energy capacity region shrink.

Acknowledgment

This research was supported in part by the European Commission under Marie Skłodowska-Curie Individual Fellowship No. 659316 and Euro-Mediterranean Cooperation ERA-NET project COM-MED.

References

[1] S. Belhadj Amor, S. M. Perlaza, I. Krikidis, and H. V. Poor, "Feedback enhances simultaneous wireless information and energy transmission in multiple access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5244–5265, Aug. 2017.

[2] H.-F. Chong, M. Motani, H. K. Garg, and H. El Gamal, "On the Han-Kobayashi region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3188–3195, Jul. 2008.

[3] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

[4] M. Fouladgar and O. Simeone, "On the transfer of information and energy in multi-user systems," *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1733–1736, Nov. 2012.

[5] P. Grover and A. Sahai, "Shannon meets Tesla: Wireless information and power transfer," in *IEEE International Symposium on Information Theory*, Austin, TX, USA, Jun. 2010, pp. 2363–2367.

[6] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, pp. 49–60, Jan. 1981.

[7] S. U. O. Ozel, K. Tutuncuoglu and A. Yener, "Fundamental limits of energy harvesting communications," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 126–132, Apr. 53.

[8] J. Park and B. Clerckx, "Joint wireless information and energy transfer in a two-user MIMO interference channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4210–4221, Aug. 2013.

[9] B. Suzhi, K. H. Chin, and Z. Rui, "Wireless powered communication: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 117–125, Apr. 2015.

A Novel Centralized Strategy for Coded Caching with Non-uniform Demands

Pierre Quinton, Saeid Sahraei and Michael Gastpar
EPFL
IPG (IC)
CH-1015 Lausanne, Switzerland
Email: {pierre.quinton, saeid.sahraei, michael.gastpar}@epfl.ch

Abstract—**Despite significant progress in the caching literature concerning the worst case and uniform average case regimes, the algorithms for caching with nonuniform demands are still at a basic stage and mostly rely on simple grouping and memory-sharing techniques. In this work we introduce a novel centralized caching strategy for caching with nonuniform file popularities. Our scheme allows for assigning more cache to the files which are more likely to be requested, while maintaining the same sub-packetization for all the files. As a result, in the delivery phase it is possible to perform linear codes across files with different popularities without resorting to zero-padding or concatenation techniques. We will describe our placement strategy for arbitrary range of parameters. The delivery phase will be outlined for a small example for which we are able to show a noticeable improvement over the state of the art.**

I. Introduction

Caching is a communication technique for redistributing the traffic in a broadcast network and thereby reducing its variability over time. The idea is to transfer part of the data to the users during low traffic periods. This data is stored at the caches of the users and helps as side information when later the server transfers the remaining data in a second phase. The central question in the caching literature is that for a given cache size, by how much one can reduce the traffic in this second (delivery) phase, assuming that in the first (placement) phase one only had partial or no knowledge at all of the requests of the users. There has been significant progress in answering this question under two paradigms. Firstly, when we look at the worst case delivery rate, meaning that we aim at minimizing the delivery rate for *any* request vector. Secondly, when we consider an average delivery rate under *uniform* distribution of the popularity of the files. For both of these scenarios the exact tradeoff between the size of the cache and the delivery rate has been characterized under uncoded placement [1], [2] , i.e., when in the placement phase users are not permitted to perform coding across several files.

By comparison, the question about minimizing the average delivery rate when the file popularities are non-uniform is still largely open. The main line of work [3]–[6] consists of partitioning the files into two or more groups, where each group contains files with similar popularity. Then one performs memory-sharing between these groups: each user divides his cache into several chunks, and assigns a chunk to each group of files. Naturally, if a group includes the more popular files a larger chunk of the cache (per file) will be allocated to them. Finally in the delivery phase each group is served individually, ignoring coding opportunities between files from different groups.

This simple scheme even when restricted to two groups has been proved to be order-optimal, meaning that it achieves a rate within a constant factor of an information theoretic converse bound. Nevertheless, the fact that coding opportunities between files from different groups are ignored should be viewed as an unfortunate technical obstacle rather than a natural extension of the strategies that exist for uniform caching. The dilemma is clear: assigning unequal amounts of cache to different groups and applying the centralized caching strategy in [1] for each group results in different sub-packetizations for files that belong to different groups. As a result, their sub-files will be of unequal size. It is therefore impossible to apply linear codes between different groups unless we resort to zero padding strategies or we concatenate the subfiles. Problems of the same nature - but perhaps less severe - appear if we resort to decentralized caching strategies [3], [5], [7].

Our main contribution in this paper is to propose a centralized caching strategy that bypasses this seemingly inevitable barrier. Specifically our placement strategy allows us to assign different amount of cache per file to different groups while maintaining equal sub-packetization for all the files. It is then very natural to allow for coding between files even if they do not belong to the same group. To the best of our knowledge this is the first centralized caching strategy that is specifically tailored for nonuniform file popularity. We will demonstrate the potential of this caching strategy by providing explicit delivery schemes for a small choice of the parameters and comparing its performance with the grouping strategies discussed earlier.

The rest of the paper is organized as follows. In Section II we will briefly describe the model. We will then move on to explaining our placement strategy in Section III. Next, in section IV we will describe our delivery strategy for a small choice of the parameters and compare its performance to the literature. Finally, we will conclude our work in Section V.

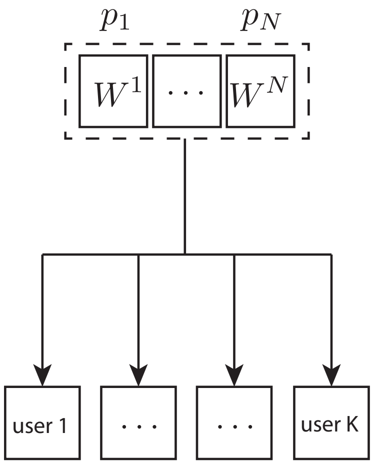


Fig. 1: An illustration of the caching network

II. Model Description

Our model and notation will be almost identical to the one described in [1]. We have a server which is in possession of N independent files $\{W^1, \dots, W^N\}$ of equal size F and K users each equipped with a cache of size MF. The communication is done in two phases. In the placement phase, the server fills in the cache of each user without prior knowledge of their requests but with the knowledge of the popularity of the files. Next, each user requests precisely one file from the server. The request of each user is drawn independently from a distribution $p_{[1:N]}$ where p_i represents the probability of requesting file i. Note that this distribution does not vary across different users. We represent the set of requests by a vector d where $d_i \in [1 : N]$ for all $i \in [1 : K]$. In the delivery phase the server broadcasts a message of rate $R(d)$ to satisfy all the users simultaneously. See Figure 1 for an illustration. We deviate from the model in [1] in that we look at the expected delivery rate instead of the peak delivery rate. We say that a memory-rate pair (M, R) is achievable if and only if there exists a joint caching and delivery strategy with a cache of size MF such that for any request vector d a delivery message of rate $R(d)F$ satisfies all the users simultaneously, and

$$R = \sum_d \mathbb{P}(d)R(d) = \sum_{d\in[1:N]^K} \prod_{i=1}^{K} p_{d_i} R(d).$$

III. The Placement Phase of Strategy β

The placement phase of our strategy, which we refer to as strategy β, starts by partitioning the N files into L groups G_1,...,G_L of respective size N_1,...,N_L, such that $\sum_{i=1}^{L} N_i = N$. How to perform this partitioning is left as a design parameter but in general files within one partition should have similar probabilities of being requested. We represent by $g_i \in [1 : L]$ the group to which the file W^i belongs. Accordingly, each user partitions his cache into L chunks of size M_1,...,M_L such that for any $\ell \in [1 : L]$, we have $M_\ell \in \{0, N_\ell/K, 2N_\ell/K, \dots, N_\ell\}$. It should be clear that this is only possible for discrete values of $M = \sum_{i=1}^{L} M_i$. The overall achievable memory-rate region will be the convex hull of all the discrete pairs (M, R) which can be served by our strategy. We define

$$r_\ell = KM_\ell/N_\ell \tag{1}$$

and assume without loss of generality that $r_1 \geq r_2 \geq \cdots \geq r_L$. Note that $r_{[1:L]}$ are integers.
Naturally, the following two identities hold.

$$\sum_{\ell=1}^{L} N_\ell r_\ell = MK \tag{2}$$

$$0 \leq r_\ell \leq K \qquad \forall \ell \in [1 : L]. \tag{3}$$

Every file in the network regardless of which group they belong to is divided into S subfiles of equal size where

$$S = \binom{K}{K - r_1, r_1 - r_2, \dots, r_{L-1} - r_L, r_L}.$$

The subfiles are indexed as follows

$$\begin{aligned} W^i_{\tau_1,\dots,\tau_L} \qquad & \text{where } \tau_1 \subseteq [1 : K] & \\ & \tau_j \subseteq \tau_{j-1} & \text{for } j \in 2, \dots, L, \\ & |\tau_i| = r_i & \text{for } i \in [1 : L]. \end{aligned}$$

Note that there are precisely S such distinct indices.
For any (i, k) user k stores subfile $W^i_{\tau_1,\dots,\tau_L}$ in his cache if and only if $k \in \tau_{g_i}$.

At this point it may help to illustrate this placement strategy via a simple example. Let us say that we have 3 users and 2 files and 2 groups such that each group contains exactly one file. Let us call the files $A = W^1$ and $B = W^2$ and assume that $r_1 = 2$ and $r_2 = 1$, so $M = \frac{r_1 N_1 + r_2 N_2}{K} = 1$. We must divide file A into 6 subfiles $A = \{A_{12,1}, A_{12,2}, A_{13,1}, A_{13,3}, A_{23,2}, A_{23,3}\}$. Same division applies to file B. The contents of the caches of the two users are illustrated in Table I.

user 1	user 2	user 3
$A_{12,1}$ $A_{12,2}$ $A_{13,1}$ $A_{13,3}$	$A_{12,1}$ $A_{12,2}$ $A_{23,2}$ $A_{23,3}$	$A_{13,1}$ $A_{13,3}$ $A_{23,2}$ $A_{23,3}$
$B_{12,1}$ $B_{13,1}$	$B_{12,2}$ $B_{23,2}$	$B_{13,3}$ $B_{23,3}$

TABLE I: Placement phase of strategy β for parameters $N = 2$, $K = 3$, $M = 1$, $r_1 = 2$, $r_2 = 1$

Let us now go back to the general placement strategy and calculate the amount of cache that user k dedicates to the ℓ'th group. By definition the index of the k'th user must be present in all the sets $\tau_1, \dots, \tau_\ell$ whereas its index may or may not be present in the sets $\tau_{\ell+1}, \dots, \tau_L$. We should divide the number of such indices $\tau_1 \dots \tau_L$ by the total number of subfiles S to find the amount of cache dedicated to each file in group ℓ.

$$\begin{aligned} M_\ell = \frac{N_\ell}{S} \binom{K-1}{r_1-1} \times \prod_{i=1}^{\ell-1} \binom{r_i-1}{r_{i+1}-1} \times \prod_{i=\ell}^{L-1} \binom{r_i}{r_{i+1}} \\ = N_\ell \frac{\binom{K-1}{r_1-1}}{\binom{K}{r_1}} \times \prod_{i=1}^{\ell-1} \frac{\binom{r_i-1}{r_{i+1}-1}}{\binom{r_i}{r_{i+1}}} \\ = N_\ell \frac{r_1}{K} \times \prod_{i=1}^{\ell-1} \frac{r_{i+1}}{r_i} \\ = \frac{r_\ell N_\ell}{K}. \end{aligned}$$

Note that this expression matches with the way we defined the parameter r_ℓ in Equation (1).

IV. Delivery Strategy for $K = 3, N = 2$ and Comparison To the Literature

Let us start by describing our delivery strategy for the same toy example as in the previous section. The explicit delivery messages for all possible request vectors are provided in Table II.

request vector	delivery message	delivery rate
(A, A, A)	$A_{12,1} \oplus A_{13,1} \oplus A_{23,2}$ $A_{12,2} \oplus A_{13,3} \oplus A_{23,3}$	$1/3$
(A, A, B)	$B_{12,1} \oplus A_{23,2}$, $B_{13,1} \oplus A_{23,3}$ $B_{12,2} \oplus A_{13,1}$, $B_{23,2} \oplus A_{13,3}$	$2/3$
(A, B, B)	$B_{12,1} \oplus A_{23,2}$, $B_{13,1} \oplus A_{23,3}$ $B_{12,2} \oplus B_{13,3}$, $B_{23,2} \oplus B_{23,3}$	$2/3$
(B, B, B)	$B_{12,1} \oplus B_{12,2}$, $B_{12,1} \oplus B_{13,3}$ $B_{13,1} \oplus B_{23,2}$, $B_{13,1} \oplus B_{23,3}$	$2/3$

TABLE II: the set of delivery messages for $N = 2, K = 3$ and $r_1 = 2, r_2 = 1$ for all possible request vectors (different permutations are omitted.)

Let us say that file A is requested with probability p and file B with probability $1-p$. We assume without loss of generality that $p \geq 0.5$. The expected delivery rate is

$$R = \frac{1}{3}p^3 + \frac{2}{3}(1-p^3) = \frac{2}{3} - \frac{1}{3}p^3.$$

Alternatively we can set $(r_1, r_2) = (3, 0)$ which results in an expected delivery rate of $1 - p^3$. Therefore,

$$R_\beta = \min\{\frac{2}{3} - \frac{1}{3}p^3, 1 - p^3\}. \tag{4}$$

Therefore, the point $(M, R) = (1, \min\{\frac{2}{3} - \frac{1}{3}p^3, 1 - p^3\})$ is achievable with strategy β. We want to compare this with the achievable rate of grouping strategy in [3]. The strategy in [3] is particularly designed for decentralized caching, which by nature has an inferior performance (in terms of delivery rate) compared to its centralized counterpart. Thus, before we perform the comparison we slightly modify the strategy in [3] without compromising its basic concepts: the files are grouped in L disjoint sets and each user partitions his cache into L segments. Coding opportunities between several groups are ignored in the placement and delivery phase. However, instead of performing decentralized caching within each group we deploy the centralized caching strategy from [1], [2]. We refer to this as strategy α. It is easy to see that strategy α always outperforms the strategy in [3] in terms of expected delivery rate. It is also easy to see that strategy α always performs at least as good as the strategy in [1], [2] since by definition we can have only one partition which includes all the files. Let us now proceed to compare the two strategies α and β.

For the same choice of parameters $K = 3, N = 2, M = 1$, strategy α can be deployed with $L = 1$ or $L = 2$ groups. The former gives an expected rate of

$$\begin{aligned} R_{\alpha,L=1} &= \frac{1}{2}(p^3 + (1-p)^3) + \frac{2}{3}(1 - p^3 - (1-p)^3) \\ &= \frac{2}{3} - \frac{1}{6}(p^3 + (1-p)^3). \end{aligned}$$

If instead we set $L = 2$, we must divide the cache into two segments of sizes M_1 and $M_2 = 1 - M_1$. We will then ignore any coding opportunities between the files A and B, so the delivery rate is given by

$$\begin{aligned} R_{\alpha,L=2} &= [1 - M_1]p^3 + [1 - M_2](1-p)^3 \\ &+ [(1 - M_1) + (1 - M_2)](1 - p^3 - (1-p)^3) \\ &= 1 - M_1 p^3 - (1 - M_1)(1-p)^3. \end{aligned}$$

Assuming $p \geq \frac{1}{2}$ it is then profitable to set $M_1 = 1$ and we get a rate of

$$R_{\alpha,L=2} = 1 - p^3.$$

To summarize, we can write

$$R_\alpha = \min\{\frac{2}{3} - \frac{1}{6}(p^3 + (1-p)^3), 1 - p^3\}. \tag{5}$$

Comparing Equations (4) and (5) we see that strategy β strictly outperforms strategy α as long as $\frac{1}{2} < p < (1/2)^{\frac{1}{3}} \approx 0.794$. Let us summarize this in a table.

probability of file A	Expected Delivery Rate	
	strategy α	Strategy β
$0.5 \leq p \leq 0.739$	$\frac{2}{3} - \frac{1}{6}(p^3 + (1-p)^3)$	$\frac{2}{3} - \frac{1}{3}p^3$
$0.739 < p \leq 0.794$	$1 - p^3$	$\frac{2}{3} - \frac{1}{3}p^3$
$0.794 < p \leq 1$	$1 - p^3$	$1 - p^3$

TABLE III: Comparison of the expected delivery rate of strategies α and β when $K = 3, N = 2$ and $M = 1$. We assume that file A is requested with probability $p \geq 1/2$.

In Figure 2 we compare the delivery rates of the two strategies for $N = 2, K = 3, M = 1$. On the horizontal axis the probability of ordering file A increases from 0.5 to 1 and on the vertical axis we have the expected delivery rate. The maximum gain is offered over strategy α when $p = 0.738$ in which case $R_\beta \approx 0.89 R_\alpha$. A converse bound from [8] is plotted for comparison.

Similar analysis can be done for other cache sizes. In Table IV we summarize the achievable rate of strategy β for

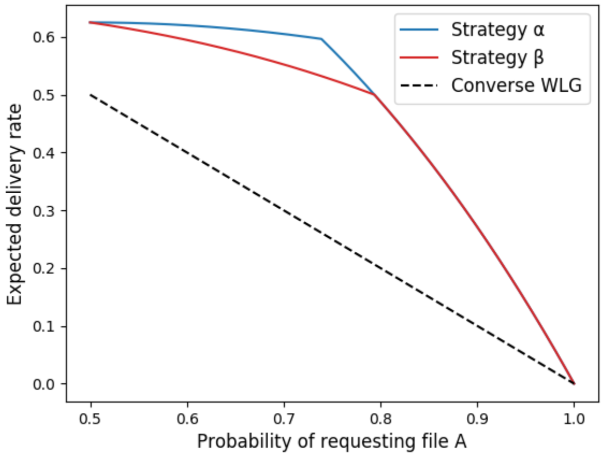


Fig. 2: Comparison of strategies α and β resumed in Table III together with the converse bound from [8].

cache size, M	(r_1, r_2)	Expected Delivery Rate
0	$(0,0)$	$2 - p^3 - (1-p)^3$
$\frac{1}{3}$	$(1,0)$	$\frac{5}{3} - p^3 - \frac{2}{3}(1-p)^3$
$\frac{2}{3}$	$(1,1)$	$1 - \frac{1}{3}p^3 - \frac{1}{3}(1-p)^3$
1	$(2,1)$	$\frac{2}{3} - \frac{1}{3}p^3$
1	$(3,0)$	$1 - p^3$
$\frac{4}{3}$	$(2,2)$	$\frac{1}{3}$
$\frac{4}{3}$	$(3,1)$	$\frac{2}{3} - \frac{2}{3}p^3$
$\frac{5}{3}$	$(3,2)$	$\frac{1}{3} - \frac{1}{3}p^3$
2	$(3,3)$	0

TABLE IV: The expected delivery rate of strategy β when $K = 3, N = 2$ for different values of (r_1, r_2) which results in different cache sizes M. We assume that file A is requested with probability $p \geq 1/2$.

difference choices of the parameters r_1 and r_2 which results in $M = (r_1 + r_2)/K$.

The achievable memory-rate region for $K = 3$, $N = 2$ is the convex hull of all these points. Note that depending on the value of p some of these points may become irrelevant. For instance if $p = 1$, the points achieved by setting $(r_1, r_2) = (2, 2)$ does not lie on the boundary of the convex hull. In Figure 3 we have plotted the achievable memory-rate region for strategies α and β for $N = 2, K = 3$ and for $p = 0.765$, where the improvements offered by strategy β are most visible. Again, the converse bound from [8] has been included for comparison. Note that the plot has been trimmed, since the performance is identical for very small or very large cache sizes. The gains are most visible in the vicinity of $M = 1$.

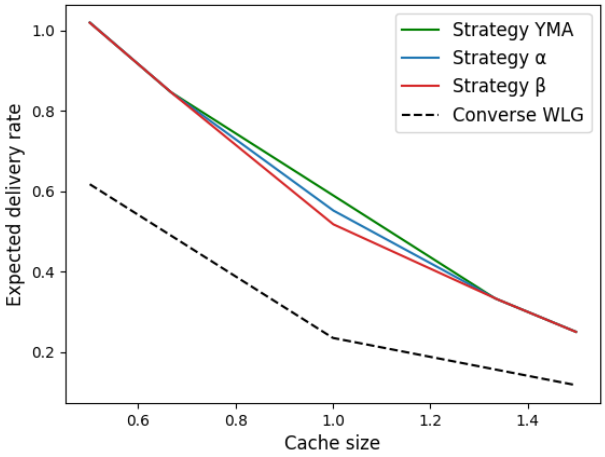


Fig. 3: Comparison of the expected delivery rate of strategies α and β and the strategy described in [2] (YMA) when $K = 3, N = 2$, together with the converse bound from [8]. We assume that file A is requested with probability $p = 0.765$.

V. Conclusion and Future Work

In this paper we presented a novel centralized caching strategy for non-uniform demands and demonstrated that for a small choice of parameters it outperforms the state of the art. For our future work, we intend to generalize our delivery strategy to arbitrary range of parameters. It is noteworthy that our strategy has the potential to be adapted to a user-specific popularity scenario, that is when the probability of requesting different files varies across the users. This can serve as another interesting direction for future research.

References

[1] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[2] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1613–1617.

[3] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 1146–1158, 2017.

[4] J. Zhang, X. Lin, and X. Wang, “Coded caching under arbitrary popularity distributions,” in *Information Theory and Applications Workshop (ITA), 2015*. IEEE, 2015, pp. 98–107.

[5] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, “On the average performance of caching and coded multicasting with random demands,” in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*. IEEE, 2014, pp. 922–926.

[6] ——, “Order-optimal rate of caching and coded multicasting with random demands,” *IEEE Transactions on Information Theory*, 2017.

[7] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” *IEEE/ACM Transactions On Networking*, vol. 23, no. 4, pp. 1029–1040, 2015.

[8] C.-Y. Wang, S. H. Lim, and M. Gastpar, “A new converse bound for coded caching,” in *Information Theory and Applications Workshop (ITA), 2016*. IEEE, 2016, pp. 1–6.

On the errors of a predictor which is calibrated to its random environment

(Extended Abstract)

Joel Ratsaby
Department of Electrical and Electronics Engineering
Ariel University
Email: ratsaby@ariel.ac.il

***Abstract*— Let $X^{(m)}$ and $X^{(n)}$ be two binary Markov chains that are randomly drawn according to a stationary Markov environment. We are interested in the following question: given $X^{(m)}$, how non-random can the sequence of errors that are made when predicting $X^{(n)}$ be and how is it influenced by the complexity of a predictor? To answer this, we consider a predictor which is 'calibrated' to the environment based on a sample $X^{(m)}$, that is, it minimizes an upper bound (based on $X^{(m)}$) on the probability of making a prediction error. We define a test of randomness for the sequence of errors made by this predictor in predicting the other sample $X^{(n)}$. The test is based on the difference between the average number of errors made on $X^{(m)}$ and on $X^{(n)}$. We derive a bound on the possible range of this difference and show how the complexity of a predictor influences how atypically random the error sequence can be.**

I. Introduction

Prediction and compression are fundamentally related [4]. The number of bits (codeword length) sufficient to represent the next symbol in a data sequence equals the logarithm of the inverse of the prediction probability of that symbol conditioned on the previous symbols [2, 4]. Typically, a random environment (or source) generates data in a non-i.i.d. manner. Predicting the next symbol x in the data, given a context which consists of available data in proximity to x (either in time or space) means that the probability distribution of x conditioned on this context is less uniform. This means that the number of bits sufficient to encode x is smaller [2] and the data sequence can be compressed more efficiently. This fact is used in many compression algorithms (for instance, in image compression) by encoding the sequence of prediction errors instead of the actual data sequence. This leads to more efficient compression since the errors are often 'less random' and their distribution has a smaller support (or is less uniform) than the actual data itself.

In this paper we are interested in how less random are prediction errors and how this depends on the complexity of the predictor. We consider an environment that is represented as a binary Markov chain of order k^* (which is assumed to be unknown). As a predictor of the environment, we consider any binary function defined over a space of all states that consists of k bits (where k may be different than k^*). Amongst all such binary functions, we choose one which is calibrated to the environment based on a finite Markov chain which is sampled from the environment. By calibrated, roughly speaking, we mean that it has the lowest prediction error and we set this to be our criterion. We define the complexity of the calibrated predictor to be the uncertainty in meeting this criterion.

To asses how non-random the errors are, we apply the predictor on another Markov chain which is sampled from the same environment. We then look at the discrepancy between the average number of errors made on the above two samples, where on the second sample we consider time instants when the prediction is sufficiently confident. From this we define a test of randomness which is based on a bound on the possible range of values of this discrepancy. We then obtain an explicit dependence of the discrepancy on the complexity of the calibrated predictor. We start with describing the setup.

II. setup

Let $\{X_t : t \in \mathbb{Z}\}$ be a sequence of binary random variables possessing the following Markov property,

$$\begin{aligned} & P\left(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots\right) \\ & \quad = \; P\left(X_t = x \mid X_{t-1} = x_{t-1}, \ldots, X_{t-k^*} = x_{t-k^*}\right) \quad (1) \end{aligned}$$

where $x_{t-k^*}, \ldots, x_{t-1}$, x_t take binary values in the set $\{-1, 1\}$. This sequence is known as a Markov *chain* of order k^*. We model the *environment* as a stationary time-homogeneous Markov chain of order k^*. We assume that k^* is unknown.

From the environment, we sample $m + \max\{k, k^*\}$ consecutive values that form a finite Markov chain

$$X^{(m)} \;:=\; \{X_t\}_{t=-\max\{k,k^*\}+1}^{m}. \tag{2}$$

Denote by $\mathbb{S}_{k^*}$ a set of states $s^{(i)}$, $i = 0, 1, \ldots, 2^{k^*} - 1$, where $s^{(0)} := [s^{(0)}_{k^*-1}, \ldots, s^{(0)}_0] = [-1, \ldots, -1, -1]$, $s^{(1)} := [s^{(1)}_{k^*-1}, \ldots, s^{(1)}_0] = [-1, \ldots, -1, 1]$, $\ldots$,$s^{(2^{k^*}-1)} := [1, \ldots, 1]$. Based on $\mathbb{S}_{k^*}$, the chain $X^{(m)}$ can be represented as a sequence

$$S^{*(m)} = \{S^*_t\}_{t=1}^{m} \tag{3}$$

of random states where

$$S^*_t \;:=\; \left(X_{t-(k^*-1)}, X_{t-(k^*-2)}, \ldots, X_t\right) \in \mathbb{S}_{k^*} \tag{4}$$

defines the random state at time t. With respect to $\mathbb{S}_{k^*}$, a state transition occurs from S^*_t to S^*_{t+1} by shifting left the sequence of bits in (4), to obtain $S^*_{t+1} := (X_{t-(k^*-2)}, \ldots, X_t, X_{t+1})$. There are two possible transitions that can occur from S^*_t into S^*_{t+1}: a *negative* transition, where the lower bit X_{t+1} is -1 and *positive* transition where X_{t+1} is 1.

We denote by Q a $2^{k^*} \times 2^{k^*}$ transition probability matrix of the Markov chain $\{X_t : t \in \mathbb{Z}\}$. Its $(ij)^{th}$ entry is denoted by

$$Q[i,j] := p\left(s^{(j)} \middle| s^{(i)}\right). \tag{5}$$

We denote by $p(1|i)$ and $p(-1|i)$, the probability of the two possible transitions from state $s^{(i)}$ and we assume that for all $0 \le i \le 2^{k^*}-1$, $p(1|s^{(i)}) > 0$, thus the environment's Markov chain is irreducible. Let

$$\pi := \left[\pi_0, \ldots, \pi_{2^{k^*}-1}\right] \tag{6}$$

denote the stationary probability distribution where π_i is the probability that $S^*_t = s^{(i)}$. That a stationary probability distribution exists is explained in the full version of the paper [3].

For a positive integer k, which may be different from k^*, we define a state space $\mathbb{S}_k$. Denote by $S^{(m)}$ the sequence of states of $\mathbb{S}_k$ that corresponds to $X^{(m)}$, that is,

$$S^{(m)} = \{S_t\}_{t=1}^m \tag{7}$$

and

$$S_t \;:=\; \left(X_{t-(k-1)}, X_{t-(k-2)}, \ldots, X_t\right) \in \mathbb{S}_k. \tag{8}$$

Consider a De Bruijn graph of dimension k with vertex set that corresponds to $\mathbb{S}_k$. An edge exists between two distinct vertices if the transition probability (5) from at least one of the corresponding states to the other, is positive. Such graph is 2-connected (maximum degree 4). Define the distance $\mathsf{d}(s, s')$ between states $s, s' \in \mathbb{S}_k$ to be the length of the shortest path between the corresponding vertices. Define the diameter of $\mathbb{S}_k$ as $\mathrm{diam}(\mathbb{S}_k) := \max_{s,s' \in \mathbb{S}_k} \mathsf{d}(s, s')$ and it equals k.

A *γ-cover* of $\mathbb{S}_k$ with respect to the metric d is a set $C \subseteq \mathbb{S}_k$ such that for every element $s \in \mathbb{S}_k$ there exists an $s' \in C$ such $\mathsf{d}(s, s') \le \gamma$. The size of the smallest γ-cover of $\mathbb{S}_k$ is defined as the *γ-covering number* of $\mathbb{S}_k$ with respect to d, and is denoted by N_γ.

III. Prediction rules, width and margin

The value of k is chosen by us as a guess to the true unknown order k^* of the environment's Markov chain. As possible predictors of the next symbol of the environment, we use binary functions on $\mathbb{S}_k$. Denote by $\mathcal{H}$ the class of all binary functions $h : \mathbb{S}_k \to \{-1, 1\}$. For a subset $R \subseteq \mathbb{S}_k$ let

$$\mathrm{dist}(s, R) := \min_{s' \in R} \mathsf{d}(s, s').$$

From [1], we define the *width* of h at s as

$$w_h(s) := \mathrm{dist}\left(s, R_{\overline{h}(s)}\right) \tag{9}$$

where R_+, $R_- \subseteq \mathbb{S}_k$ are regions classified as 1 and -1, respectively, by h, and $\overline{h}(s)$ is the complement of $h(s)$. Because $s \notin R_{\overline{h}(s)}$ then $w_h(s) > 0$. Define $f_h : \mathbb{S}_k \to \mathbb{R}$ by

$$f_h(s) := h(s) w_h(s) \tag{10}$$

to be a *margin* function associated with h. We can evaluate the width and margin functions because k is known and thus the edges of the De Bruijn graph on $\mathbb{S}_k$ are known (the De Bruijn graph of the environment's space $\mathbb{S}_{k^*}$ and its corresponding transition matrix Q are not needed). For the purpose of bounding the prediction error it is convenient to express the classification $h(s)$ in terms of the margin as follows, $h(s) = \mathrm{sgn}\left(f_h(s)\right)$ where $\mathrm{sgn}(a)$ equals 1 if $a > 0$ and -1 otherwise. The absolute value of $f_h(s)$ expresses the 'confidence' in the decision $h(s)$. We use this fact to consider errors made at confident predictions.

Given any binary function $h \in \mathcal{H}$, the predictor based on h decides at time t according to the following rule: if $h(S_{t-1}) = 1$ it predicts for X_t the value 1, otherwise it predicts -1. The probability that h makes a prediction error is constant with respect to time t because the environment is stationary. We denote it by

$$L(h) := \mathbb{P}\left(h(S_{t-1}) \ne X_t\right) = \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < 0\right). \tag{11}$$

Denote the l_∞ -norm of f_h by $\|f_h\| := \max_{s \in \mathbb{S}_k} |f_h(s)|$. Denote the class of margin functions by $\mathcal{F} := \{f_h : h \in \mathcal{H}\}$. An α-cover of $\mathcal{F}$ with respect to the l_∞ norm on $\mathbb{S}_k$ is a set $\hat{F}_\alpha := \left\{f_j^{(\alpha)}\right\}_{j=1}^r$ such that for every element $f \in \mathcal{F}$ there exists an $f_j^{(\alpha)} \in \hat{F}_\alpha$ such $\left\|f - f_j^{(\alpha)}\right\|_{l_\infty} \le \alpha$. We denote by $h_j := \mathrm{sgn}\left(f_j^{(\alpha)}\right)$ the binary function that corresponds to $f_j^{(\alpha)}$ (note that $j := j(\alpha)$ and we omit the dependence on α for brevity). The size r of the smallest α-cover of $\mathcal{F}$ is defined as the α-covering number of $\mathcal{F}$ with respect to l_∞ norm on $\mathbb{S}_k$ and is denoted by $\mathcal{N}_\alpha$. From [1] it follows that

$$\mathcal{N}_\alpha \le \left(2\left\lceil \frac{3\mathrm{diam}\left(\mathbb{S}_k\right)}{\alpha} \right\rceil + 1\right)^{N_{\alpha/3}} \tag{12}$$

where N_α is the α-covering number of $\mathbb{S}_k$ with respect to the metric d. It follows that

$$\log_2 \mathcal{N}_{\alpha/2} \le N_{\alpha/6} \log_2\left(\frac{15k}{\alpha}\right). \tag{13}$$

A. Margin error

Denote by $\gamma \in [0, \mathrm{diam}(\mathbb{S}_k)]$ a margin parameter and let $a := a(\gamma)$ and $b := b(\gamma)$ be non-decreasing functions. The *prediction margin error* of h at time t is

$$X_t f_h(S_{t-1}) < b(\gamma). \tag{14}$$

When $b(\gamma) = 0$, (14) is the prediction error (whose probability is (11)). Based on $X^{(m)}$ define the *margin-error sequence* as

$$\Psi^{(m,\gamma)}(h) := \left\{\Psi_t^{(m,\gamma)}(h)\right\}_{t=1}^m = \left\{\mathbb{I}\left\{X_t f_h(S_{t-1}) < b(\gamma)\right\}\right\}_{t=1}^m.$$

The average number of times that a margin-error occurs on $X^{(m)}$ by h is defined as

$$L_m^{(b(\gamma))}(h) := \frac{1}{m}\sum_{t=1}^m \Psi_t^{(m,\gamma)}(h).$$

Its expected value is denoted by

$$L^{(b(\gamma))}(h) \quad := \quad \mathbb{P}\left(X_t f_h(S_{t-1}) < b(\gamma)\right).$$

In case $b(\gamma) = \gamma$ we denote by

$$L_m^{(\gamma)}(h): \quad = \tfrac{1}{m}\textstyle\sum_{t=1}^{m} \mathbb{I}\left\{X_t f_h(S_{t-1}) < \gamma\right\}$$

and

$$L^{(\gamma)}(h) \quad = \quad \mathbb{P}\left(X_t f_h(S_{t-1}) < \gamma\right).$$

After obtaining a sample $X^{(m)}$ from the stationary environment, we sample $n + \max\{k, k^*\}$ consecutive bits to obtain a second sample

$$X^{(n)} \quad := \quad \{X_t\}_{t=-\max\{k,k^*\}+1}^{n}$$

with a corresponding state sequence $S^{(n)}$ as defined in (7). Based on $X^{(n)}$, the prediction *error sequence* is defined as

$$\Psi^{(n)}(h) := \{\Psi_t\}_{t=1}^{n} := \{\mathbb{I}\left\{X_t f_h(S_{t-1}) < 0\right\}\}_{t=1}^{n}.$$

The average number of times that an error occurs when predicting $X^{(n)}$ by h is defined as

$$L_n(h) := \frac{1}{n}\sum_{t=1}^{n} \Psi_t^{(n)}(h).$$

We are interested in the non-randomness of the sequence of errors that correspond to instants of time at which a predictor is confident. Observing $\Psi^{(n)}$ only at such time instants yields the following two error subsequences,

$$\Psi^{(\nu_-^{(\gamma)})}(h) := \{\mathbb{I}\left\{X_{t_l} f_h(S_{t_l-1}) < 0\right\}\}_{l=1}^{\nu_-^{(\gamma)}}$$

and

$$\Psi^{(\nu_+^{(\gamma)})}(h) := \{\mathbb{I}\left\{X_{t_l} f_h(S_{t_l-1}) < 0\right\}\}_{l=1}^{\nu_+^{(\gamma)}}$$

where $\nu_-^{(\gamma)}$, $\nu_+^{(\gamma)}$ are the number of times that the sequence $S^{(n)}$ enters a state $s \in \mathbb{S}_k$ such that $f_h(s) < -a(\gamma)$ and $f_h(s) > a(\gamma)$, respectively. We define the averages of $\Psi^{(\nu_-^{(\gamma)})}$ and $\Psi^{(\nu_+^{(\gamma)})}$ as follows,

$$H_{\nu_-}^{(\gamma,n)}(h) = \frac{1}{\nu_-^{(\gamma)}} \sum_{l: f_h(S_{t_l-1}) < -a(\gamma)} \mathbb{I}\left\{X_{t_l} f_h(S_{t_l-1}) < 0\right\} \tag{15}$$

and

$$H_{\nu_+}^{(\gamma,n)}(h) = \frac{1}{\nu_+^{(\gamma)}} \sum_{l: f_h(S_{t_l-1}) > a(\gamma)} \mathbb{I}\left\{X_{t_l} f_h(S_{t_l-1}) < 0\right\}. \tag{16}$$

B. Discrepancy

We define the following measures of *discrepancy*, which are functions of $X^{(m)}$ and $X^{(n)}$:

$$\Upsilon_-^{(\gamma)}(h) := \Upsilon_{m,n}^{(\nu_-^{(\gamma/2)})}(h) = H_{\nu_-}^{(\gamma/2,n)}(h) - L_m^{(b(2\gamma))}(h)$$

and

$$\Upsilon_+^{(\gamma)}(h) := \Upsilon_{m,n}^{(\nu_+^{(\gamma/2)})}(h) = H_{\nu_+}^{(\gamma/2,n)}(h) - L_m^{(b(2\gamma))}(h).$$

The expected value is,

$$\begin{aligned}\mathbb{E}\left[H_{\nu_-}^{(\gamma/2,n)}\right] &= \mathbb{E}_{\nu_-}\left[\frac{1}{\nu_-}\mathbb{E}\left[\sum_{l=1}^{\nu_-}\mathbb{I}\left\{X_{t_l} f_h\left(S_{t_l}\right) < 0\right\}\middle|\nu_-\right]\right] \\ &= \mathbb{E}_{\nu_-}\left[\frac{1}{\nu_-}\nu_- L(h)\right] = L(h),\end{aligned}$$

and

$$\mathbb{E}L_m^{(b(2\gamma))}(h) = L^{(b(2\gamma))}(h).$$

We have,

$$\begin{aligned}L(h) &= \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < 0\right) \leq \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < b(2\gamma)\right) \\ &= L^{(b(2\gamma))}(h)\end{aligned}$$

therefore

$$\begin{aligned}\mathbb{E}\Upsilon_-^{(\gamma)}(h) &= \mathbb{E}H_{\nu_-}^{(\gamma/2)}(h) - \mathbb{E}L_m^{(b(2\gamma))}(h) \\ &= L(h) - L^{(b(2\gamma))}(h) \\ &\leq 0.\end{aligned} \tag{17}$$

Similarly, we have

$$\mathbb{E}\Upsilon_+^{(\gamma)}(h) \leq 0. \tag{18}$$

IV. Calibrated predictor

In [3] it is shown that there exists a finite integer $\mathfrak{l}_0$, such that for $l \geq \mathfrak{l}_0$, the transition matrix Q in (5) satisfies $Q^l > 0$, that is, every entry of Q^l, denoted by $p^{(l)}(s^{(j)}|s^{(i)})$, is positive. We choose $\mathfrak{l}_0 := \min\{l : Q^l > 0\}$ and in theory, if Q was known then $\mathfrak{l}_0$ can be evaluated by computing Q^l for a sequence $l \geq 1$ until the first l is found such that $Q^l > 0$. Define the minimum entry of $Q^{\mathfrak{l}_0}$ by μ_0. We henceforth make the following assumption:

Assumption 1. *The environment's transition matrix Q satisfies one of the following conditions: (i) the minimum entry of $Q^{\mathfrak{l}_0}$ is $\mu_0 \neq 2^{-k^*}$ or (ii) $\mu_0 = 2^{-k^*}$ and for all $0 \leq i \leq 2^{k^*} - 1$, the transition probabilities $p(1|i) = p(-1|i) = \frac{1}{2}$.*

In both parts (i) and (ii) of the above assumption, Q may have a uniform stationary distribution $\pi^T = \left[2^{-k^*}, \ldots, 2^{-k^*}\right]$, which means Q is doubly stochastic and $\lim_{l\to\infty} Q^l$ is a matrix U, of the same size as Q, with all its entries identical to 2^{-k^*}. Part (ii) treats the special case where this limit U is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$. According to the cases of Assumption 1, define

$$\rho(k^*, \mathfrak{l}_0) := \begin{cases} \frac{1-2^{k^*}\mu_0}{2\mu_0} & \text{if case (i) holds} \\ & \text{and } \mathfrak{l}_0 = 1, \\ \frac{2^{k^*-1}}{\left(1-2^{k^*}\mu_0\right)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-\left(1-2^{k^*}\mu_0\right)^{1/\mathfrak{l}_0}\right)} & \text{if case (i) holds} \\ & \text{and } \mathfrak{l}_0 \geq 2, \\ 2^{k^*-1} & \text{if case (ii) holds.} \end{cases}$$

Define

$$\eta(m,\gamma,\delta) := r(k,k^*)\rho(k^*,\mathfrak{l}_0) \sqrt{\frac{2}{m}\left((\ln 2)\left(1 + \left(N_{\gamma/12}+1\right)\log_2\left(\frac{30k}{\gamma}\right)\right) + \ln\left(\frac{1}{\delta}\right)\right)}$$

where

$$r(k,k^*) := \begin{cases} 1 & \text{if } k^* \geq k+1 \\ k-k^*+2 & \text{if } k^* \leq k. \end{cases}$$

We define the *penalized margin error* of h as

$$\hat{L}_m^{(\gamma)}(h) := L_m^{(\gamma)}(h) + \eta(m,\gamma,\delta)$$

which is a random variable since it depends on $X^{(m)}$ through $L_m^{(\gamma)}(h)$. The following is a concentration bound for a Markov chain which holds uniformly over the class $\mathcal{H}$ and over the range of values for γ.

Lemma 1. *For $\gamma > 0$ let N_γ be the γ-covering number of $\mathbb{S}_k$ with respect to the metric* d. *Let $X^{(m)}$ be a Markov chain sampled from the environment. For any $0 < \delta \leq 1$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$ and for every $0 < \gamma \leq diam(\mathbb{S}_k)$, the following holds*

$$L(h) \leq \hat{L}_m^{(\gamma)}(h). \tag{19}$$

The proof is provided in [3]. Next, we use $\hat{L}_m^{(\gamma)}(h)$ as a *criterion function* for selecting a good predictor. Given a random sequence $X^{(m)}$ let (h',γ') be any pair that satisfies the following:

$$\hat{L}_m^{(\gamma')}(h') = \min_{h\in\mathcal{H},\gamma\in(0,\text{diam}(\mathbb{S}_k)]} \hat{L}_m^{(\gamma)}(h). \tag{20}$$

Let

$$\gamma_m := \max\{\gamma' : (h',\gamma') \text{ satisfies } (20)\}$$

and denote by h_m its corresponding function. Define (h_m,γ_m) as a *calibrated predictor*, that is, a predictor system which is calibrated to its random environment based on a sample $X^{(m)}$. It is shown in [3] that the calibrated predictor (h_m,γ_m) always exists.

Remark 2. The calibrated predictor (h_m,γ_m) minimizes the penalized margin error over $h \in \mathcal{H}$ and over the range of values of γ. By Lemma 1, with probability at least $1-\delta$, h_m has a minimum upper bound $L(h_m) \leq \hat{L}_m^{(\gamma_m)}(h_m)$. If this occurs, we say that the criterion *succeeds*.

Note that while γ_m is not used in the predictor's decision, its choice influences which $h \in \mathcal{H}$ is selected to be h_m. The higher the value of γ_m, the less sensitive the value of $\hat{L}_m^{(\gamma_m)}(h_m)$ is to the particular realization of the random sample $X^{(m)}$, and the more that h_m generalizes to fit the typical behavior of the actual environment (as opposed to fitting just a particular sample of the environment). We therefore call γ_m the *level of adaptivity* of the calibrated predictor to the environment.

V. Complexity of calibrated predictor

For a fixed $m \geq 1$ and $0 < \delta \leq 1$, and for any $h \in \mathcal{H}$, $0 < \gamma \leq \text{diam}(\mathbb{S}_k)$ let us define

$$E_h^{(\gamma)} := \left\{x^{(m)} : L(h) > L_m^{(\gamma)}(h) + \eta(m,\gamma,\delta)\right\}$$

to be the set of bad samples on which the upper bound (19) fails to hold for system (h,γ). Let the class of such sets be defined as

$$\mathcal{E}_{\mathcal{H}} := \left\{E_h^{(\gamma)} : h \in \mathcal{H},\, 0 < \gamma \leq \text{diam}\,(\mathbb{S}_k)\right\}.$$

Next we approximate $\mathcal{E}_{\mathcal{H}}$ by a finite class of sets that are defined in a similar way. Let l be a non-negative integer. Consider a minimal $(1/2)^{l+2}k$-cover $\hat{F}_{k(1/2)^{l+2}}$ of $\mathcal{F}$ (the factor k is the diameter of $\mathbb{S}_k$). For $f_j^{(k(1/2)^{l+2})} \in \hat{F}_{k(1/2)^{l+2}}$ denote by $h_j = \text{sgn}(f_j^{(k(1/2)^{l+2})})$. Define a set of bad samples associated with h_j as

$$B_j^{(\gamma)} := \left\{x^{(m)} : L^{(\gamma)}(h_j) > L_m^{(\gamma)}(h_j) + \eta(m,4\gamma,\delta)\right\}.$$

For $0 \leq l < \infty$ denote by $B_{j,l} := B_j^{((1/2)^{l+2})}$ and define the class

$$C^{(l)} := \{B_{j,l}\}_{j=1}^{\mathcal{N}_{k(1/2)^{l+2}}}$$

where $\mathcal{N}_\gamma$ is the γ-covering number of $\mathcal{F}$ with respect to the l_∞-norm on $\mathbb{S}_k$. The next lemma states that given (h_m,γ_m) we can approximate the set $E_{h_m}^{(\gamma_m)}$ by an element of the class $C^{(l_m)}$ where l_m is directly obtained from γ_m by checking in which interval γ_m is contained.

Lemma 3. *For $m \geq 1$ let (h_m,γ_m) be a predictor calibrated based on $X^{(m)}$. Define l_m as a non-negative integer that satisfies $(1/2)^{l_m+1}k \leq \gamma_m \leq (1/2)^{l_m}k$. Then there exists a $1 \leq j \leq \mathcal{N}_{k(1/2)^{l_m+2}}$, which is denoted j_m, such that $E_{h_m}^{(\gamma_m)} \subseteq B_{j_m,l_m}$ where $B_{j_m,l_m} \in C^{(l_m)}$.*

The proof is in [3]. Next we define a notion of complexity of the calibrated predictor. In the context of [5], we set the functional requirement of the calibrated predictor (selected by the criterion) to be that if the bound (19) fails to hold for (h_m,γ_m) (the criterion fails), then this must be detected. We define the complexity of the system (h_m,γ_m) to be the level of uncertainty in detecting that the criterion fails given that it fails.

The event that represents failure of the criterion is $X^{(m)} \in E_{h_m}^{(\gamma_m)}$. From Lemma 3, if $X^{(m)} \in E_{h_m}^{(\gamma_m)}$ then $X^{(m)} \in B_{j_m,l_m}$ therefore it is possible to detect failure of the criterion by detecting that $X^{(m)}$ falls in at least one element B_{j,l_m} of $C^{(l_m)}$.

Given $X^{(m)} \in E_{h_m}^{(\gamma_m)}$, the index j_m of the set B_{j_m,l_m} that contains $X^{(m)}$ is random because the set $E_{h_m}^{(\gamma_m)}$ is random due to (h_m,γ_m). This index j_m takes values in the set $\left\{1,\ldots,\left|C^{(l_m)}\right|\right\}$ and its conditional entropy is bounded by the entropy of the uniform probability distribution on this set,

$$H\left(j_m \middle| X^{(m)} \in E_{h_m}^{(\gamma_m)}\right) \leq \log_2 \left|C^{(l_m)}\right| \text{ bits.}$$

Therefore, the uncertainty in detecting that the criterion fails, given that it fails, is what we define as the complexity of the system. It is no more than $\log_2\left|C^{(l_m)}\right|$ bits and, from (12), is bounded from above as

$$\log_2\left|C^{(l_m)}\right| \leq N_{k(1/2)^{l_m+1}/6}\left(l_m + \log_2(30)\right). \tag{21}$$

By definition of l_m we have $l_m \leq \log_2\left(\frac{k}{\gamma_m}\right)$ and $\frac{\gamma_m}{2} \leq \left(\frac{1}{2}\right)^{l_m+1} k$ therefore (21) is no larger than $N_{\gamma_m/12}\log_2\left(\frac{30k}{\gamma_m}\right)$. This is defined as the complexity of the calibrated predictor,

$$\mathcal{C}(h_m, \gamma_m) := N_{\gamma_m/12}\log_2\left(\frac{30k}{\gamma_m}\right) \text{ bits.}$$

Note that the *larger* the adaptivity level γ_m of the calibrated predictor, the *lower* its complexity $\mathcal{C}(h_m, \gamma_m)$. Thus, a calibrated predictor which is better adapted to its random environment has a lower complexity.

VI. Testing for randomness

We aim to understand how the complexity of the calibrated predictor (h_m, γ_m) affects the randomness of its output which consists of two subsequences of prediction errors, $\Psi^{(\nu_-^{(\gamma_m)})}(h_m)$ and $\Psi^{(\nu_+^{(\gamma_m)})}(h_m)$ that correspond to instants when the predictor is confident (the margin is of absolute value greater than $a(\gamma_m)$).

Let us define the following *null hypothesis*: the expected value of the discrepancy is non-positive, that is, $\mathbb{E}\Upsilon_-^{(\gamma_m)}(h_m) \leq 0$ and $\mathbb{E}\Upsilon_+^{(\gamma_m)}(h_m) \leq 0$. From (17), (18) it follows that the null hypothesis holds. Draw $X^{(m)}$ and $X^{(n)}$. Denote the realization of these sequences by $x^{(m)}$ and $x^{(n)}$. Use the criterion (20) on $x^{(m)}$ to obtain a calibrated predictor (h_m, γ_m) and measure its margin error $L_m^{(\gamma_m)}(h_m)$, which we denote by α_m. Denote by β_n^- and β_n^+ the negative and positive errors $H_{\nu_-}^{(\gamma_m)}(h_m)$, $H_{\nu_+}^{(\gamma_m)}(h_m)$ based on $x^{(n)}$, respectively.

Theorem 4, presented below, shows that the event $\Upsilon_-^{(\gamma_m)}(h_m) > \epsilon$ or $\Upsilon_+^{(\gamma_m)}(h_m) > \epsilon$ (based on error subsequences of some minimum length) has a probability less than δ. Thus we have the following test of randomness with significance level δ and critical value ϵ: if $\beta_n^- - \alpha_m > \epsilon$ or $\beta_n^+ - \alpha_m > \epsilon$ then reject the null hypothesis.

If the test rejects the null hypothesis then at least one of the discrepancy values, $\beta_n^- - \alpha_m$ or $\beta_n^+ - \alpha_m$, deviates significantly from the expected value, which means that the error subsequences do not pass the test of randomness.

VII. Main result

Let $a(\gamma_m) = 68\gamma_m$, $b(\gamma_m) = 76\gamma_m$ and substitute them in (15) and (16). The next theorem gives the expression for the critical value ϵ of the test of randomness.

Theorem 4. *For any positive integers m, n, k, k^* and $\ell \leq n$, let N_γ denote the γ-covering number of the space $\mathbb{S}_k$ with respect to* d. *Let*

$$\epsilon(m,n,\ell,\delta) = \frac{2r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\left(\frac{2}{n}\left(N_{\gamma_m/6}\ln\left(2\left\lceil\frac{6k}{\gamma_m}\right\rceil+1\right)\right.\right.$$
$$\left.\left.+\ln\left(\frac{8k(3-2(\ell/n))}{\delta\gamma_m(\omega-(\ell/n))}\right)\right)\right)^{1/2} + r(k,k^*)\rho(k^*,\mathfrak{l}_0)$$
$$\left(\frac{2}{m}\left(N_{\gamma_m/3}\ln\left(2\left\lceil\frac{3k}{\gamma_m}\right\rceil+1\right)+\ln\left(\frac{4k(3-2(\ell/n))}{\delta\gamma_m}\right)\right)\right)^{1/2}.$$

Then for any $0 < \delta \leq 1$ and $\omega \in (\ell/n, 1]$, the probability that $\Upsilon_{m,n}^{(\nu_-^{(\gamma_m/2)})}(h_m) > \epsilon$ or $\Upsilon_{m,n}^{(\nu_+^{(\gamma_m/2)})}(h_m) > \epsilon$ with $\min\left\{\nu_-^{(\gamma_m/2)}, \nu_+^{(\gamma_m/2)}\right\} \geq \omega n$ is no more than δ.

The proof of the theorem is in [3].

Remark 5. The value ℓ is the assumed minimum length of the output error sequences $\Psi^{(\nu_-^{(\gamma_m)})}(h_m)$ and $\Psi^{(\nu_+^{(\gamma_m)})}(h_m)$. The value of ω is set, only after the random lengths of these output sequences is known, to a value that satisfies $\min\{\nu_-, \nu_+\} \geq \omega n$. If ω is less than ℓ/n then the theorem cannot be applied.

The expression for ϵ involves two factors that are of the same order as the complexity $\mathcal{C}(h_m, \gamma_m)$ of the calibrated predictor. Thus ϵ increases with the predictor's complexity, which suggests that a simple predictor has a small critical value ϵ therefore it only takes a small discrepancy value $\beta_n^- - \alpha_m$ or $\beta_n^+ - \alpha_m$ to reject the null hypothesis. In other words, the more complex the calibrated system, the higher the critical value ϵ and the higher the discrepancy that is allowed without rejecting the output as being atypically random for that system. This means that a more complex calibrated predictor may produce atypically-random output error sequences.

Also, since the output error sequences consist only of the instants when the predictor has an absolute margin greater than $a(\gamma_m)$ (which is a non-decreasing function of γ_m), and since ϵ decreases as γ_m increases, then, assuming that ω is fixed (and the lengths of the error subsequences are at least ωn long), it follows that the critical value ϵ decreases as the predictor's level of confidence $a(\gamma_m)$ increases. This means that a calibrated predictor system which is based on a *lower* decision confidence value a has a *higher* critical value ϵ and hence may produce more atypically-random output error sequences.

If the adaptivity level γ_m and the error α_m are small, the system (h_m, γ_m) is not well adapted to the environment (it overfits the sample $X^{(m)}$). In this case, the critical value ϵ is large and the system may produce output error sequences with large β_n^- or β_n^+ while not rejecting the null hypothesis. This means that a calibrated predictor with a low level of adaptivity γ_m may produce more atypically-random output error sequences.

References

[1] M. Anthony and J. Ratsaby. Learning bounds via sample width for classifiers on finite metric spaces. *Theoretical Computer Science*, 529:2–10, 2014.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 2006.

[3] J. Ratsaby. Full version of the paper. http://www.ariel.ac.il/sites/ratsaby/Publications/PDF/Medium.B.pdf.

[4] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann, second edition, 2000.

[5] N. P. Suh. Complexity in engineering. *CIRP Annals - Manufacturing Technology*, 54(2):46 – 63, 2005.

Author Index