International Zurich Seminar on Information and Communication (IZS 2022)

Proceedings

International Zurich Seminar on Information and Communication

March 2 – 4, 2022

Sorell Hotel Zürichberg, Zurich, Switzerland

Proceedings

Acknowledgment of Support

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

Conference Organization

General Co-Chairs

Amos Lapidoth and Stefan M. Moser

Technical Program Committee

Yair Be′ery
Stephan ten Brink
Shraga Bross
Yuval Cassuto
Terence H. Chan
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Albert Guillén i Fàbregas
Martin Hänggi
Franz Hlawatsch
Ashish Khisti
Tobias Koch
Gerhard Kramer
Frank Kschischang
Hsuan-Yin Lin
Hans-Andrea Loeliger
Haim Permuter
Ron Roth
Igal Sason
Robert Schober
Yanina Shkel
Anelia Somekh-Baruch
Yossef Steinberg
Christoph Studer
Ido Tal
Giorgio Taricco
Emre Telatar
Pascal Vontobel
Ligong Wang
Michèle Wigger

Organizers of Invited Sessions

Ziv Goldfeld
Cheuk Ting Li
Alfonso Martinez
Ligong Wang

Local Organization

Olivia Bärtsch Popov (Secretary)
Michael Lerjen (Web and Publications)
Patrick Strebel (Registration)

Table of Contents

Keynote Lectures

Session 1 Wed 10:00 – 11:40
Information Measures and Statistical Distances

Invited session organizer: Ziv Goldfeld (Cornell University)

*Invited papers are marked by an asterisk.

Session 2 Wed 13:30 – 14:50
Shannon Theory I

Session 3 Wed 15:20 – 16:40
Shannon Theory II

Session 4 Thu 10:00 – 11:40

Rate-Distortion Theory

Invited session organizer: Ligong Wang (ETIS–ENSEA, Université de Cergy-Pontoise)

*Zero-delay Source Coding in Feedback Systems (retracted)
Jan Østergaard

*The Rate-Distortion-Perception Tradeoff: Deterministic Codes (retracted)
Aaron B. Wagner

*A Rate-Distortion-Perception Theory for Binary Sources 34
Jingjing Qian, George Zhang, Jun Chen, and Ashish Khisti

*Strategic Information Compression (retracted)
Maël Le Treust

*Source Coding with Information Obfuscation (retracted)
Ligong Wang and Gregory Wornell

Session 5 Thu 13:30 – 14:50

Advances in Mismatched Decoding: Theory and Applications

Invited session organizer: Alfonso Martinez (Universitat Pompeu Fabra, Barcelona)

*Mismatched Decoding and Rejection Sampling to Lower Bound the Capacity of the Optical Fiber Channel (retracted)
Marco Secondini

*A New Analysis of Mismatched Decoding (retracted)
Anelia Somekh-Baruch

*A Framework for Deriving Upper Bounds to the Mismatch Capacity (retracted)
Ehsan Asadi and Albert Guillén i Fàbregas

*Achievability Bounds in Multiuser Massive MIMO Systems via Mismatched Decoding (retracted)
Giuseppe Durisi

Session 6 Thu 15:20 – 16:40
Coding

Session 7 Fri 10:20 – 11:40
Classical and Quantum Coding

Session 8 Fri 13:30 – 14:30
Learning and Estimation

Session 9 Fri 15:00 – 15:40
Information Inequalities

Invited session organizer: Cheuk Ting Li (Chinese University Hong Kong)

A Class of Nonbinary Symmetric Information Bottleneck Problems

Michael Dikshtein and Shlomo Shamai (Shitz)
Technion–Israel Institute of Technology
Department of Electrical and Computer Engineering, Haifa 3200003, Israel
Email: {michaeldic@campus, sshlomo@ee}.technion.ac.il

***Abstract*—We study two dual settings of information processing. Let $\mathsf{Y} \to \mathsf{X} \to \mathsf{W}$ be a Markov chain with fixed joint probability mass function P_{XY} and a mutual information constraint on the pair (W, X). For the first problem, known as Information Bottleneck, we aim to maximize the mutual information between the random variables Y and W, while for the second problem, termed as Privacy Funnel, our goal is to minimize it. In particular, we analyze the scenario for which X is the input, and Y is the output of modulo-additive noise channel. We provide analytical characterization of the optimal information rates and the achieving distributions.**

I. Introduction

Let (X, Y) be a pair of random variables specified by a fixed bivariate distribution P_{XY}, of cardinality $|\mathcal{X}| = n$, and respectively $|\mathcal{Y}| = m$. Consider all random variables W satisfying the Markov chain $\mathsf{Y} \to \mathsf{X} \to \mathsf{W}$ subject to a constraint on the mutual information of the pair (X, W). We consider here two extremes of the information processing problem, the Information Bottleneck (IB) function and the Privacy Funnel (PF).

The IB optimization problem, introduced by Tishby et al. [1], is defined as

$$R^{\text{IB}}_{\mathsf{P}_{\mathsf{XY}}}(C) \triangleq \begin{array}{ll} \underset{\mathsf{P}_{\mathsf{W}|\mathsf{X}}}{\text{maximize}} & I(\mathsf{Y};\mathsf{W}) \\ \text{subject to} & I(\mathsf{X};\mathsf{W}) \leq C. \end{array} \tag{1}$$

This problem is illustrated in Figure 1. In our study we aim to determine the maximum value and characterize the achieving conditional distribution $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ (test channels) of (1) for a class of symmetric channels $\mathsf{P}_{\mathsf{Y}|\mathsf{X}}$, and constraints C.

The motivation to study such a model is as follows. Consider a latent random variable Y, which constitutes the Markov chain $\mathsf{Y} \to \mathsf{X} \to \mathsf{W}$ and represents a source of information. The user observes a noisy version of Y, i.e., X, and then tries to compress the observed noisy data such that its reconstructed version, W, will be comparable under the maximum mutual information metric to the original data Y. Thus, (1) is essentially a remote source coding problem [2], choosing the distortion measure as the logarithmic-loss. Here W represents the noisy version (X) of the source (Y) with a constrained number of bits ($I(\mathsf{X};\mathsf{W}) \leq C$), and the goal is to maximize the relevant information in W regarding Y (measured by the mutual information between Y and W). In the standard IB terminology, $I(\mathsf{X};\mathsf{W})$ is referred to as the complexity of W, and $I(\mathsf{Y};\mathsf{W})$ is referred to as the relevance of W.

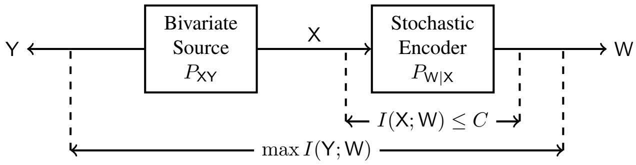

Y
Bivariate
Source
P_{XY}
X
Stochastic
Encoder
$P_{\mathsf{W}|\mathsf{X}}$
W
$I(\mathsf{X};\mathsf{W}) \leq C$
$\max I(\mathsf{Y};\mathsf{W})$

Fig. 1: Block diagram of the Information Bottleneck function.

For the particular case where $(\mathsf{Y}, \mathsf{X}, \mathsf{W})$ are discrete random variables, an optimal $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ can be found by iteratively solving a set of self–consistent equations [1]. A generalized Blahuto-Arimoto algorithm [3] was proposed to solve those equations. The optimal test-channel $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ was characterized using a variation principle in [1]. A particular case of deterministic mappings from X to W was considered in [4], and algorithms that find those mappings were described. Unfortunately, since the underlying optimization problem in (1) is not convex, there are no theoretical guarantees for convergence of the proposed iterative algorithms.

There are two cases for which the solution of (1) is thoroughly characterized. The first one, considered in [5], is where the pair (X, Y) is a Doubly Symmetric Binary Source (DSBS) with transition probability p. It was shown that the optimal test channel $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ is a BSC with transition probability $h_2^{-1}(1-C)$ where $h_2(\cdot)$ is binary entropy function and $h_2^{-1}(\cdot)$ its inverse. The converse can be established by applying Mrs. Gerber's Lemma [6]. This setting was also solved as an example in [7, Section IV.A]. The optimality of BSC test-channel extends also to a Binary Memoryless Symmetric (BMS) channel [8, Ch. 4] from X to Y, as [9, Theorem 2] implies.

The second case, first considered in [10], is where (X, Y) are jointly Gaussian. It was shown that the optimal distribution of $(\mathsf{Y}, \mathsf{X}, \mathsf{W})$ is also jointly Gaussian. The optimality of the Gaussian test-channel can be proved using conditional Entropy Power Inequality [11, Ch. 2]. It can also be established using I-MMSE and Single Crossing Property [12]. Moreover, under the I-MMSE framework, the proof can be easily extended to Jointly Gaussian Random Vectors $(\mathbf{X}, \mathbf{Y})$ [13].

The IB method can also be seen as a variation on some closely related problems in the Information Theory literature. A bound on the conditional entropy for a pair of discrete random variables subject to entropy constraint has been considered in [7] as a method to characterize common information [14]. A method based on convex analysis was proposed to find the achieving distributions and several important examples were given. We will show that the problem addressed in [7] is equivalent to (1).

The problem of Common Reconstruction (CR) [15] is a different type of source coding with side-information, a.k.a. Wyner-Ziv coding [6]. In [15] the distortion was measured with a log-loss merit, and the encoder is required to perfectly reconstruct decoder's sequence. It can be shown that for the CR, the resulting single-letter rate-distortion region is equivalent to IB.

The problem of Information Combining [16] was analyzed in the context of check nodes in LDPC decoding. Two extremes were considered in form of maximization and minimization of mutual information for the binary X setting [9]. It can be shown that the first extreme is equivalent to PF, while the second recovers the IB setting. A recent comprehensive tutorial on the IB method and related problems is given in [5].

Applications of IB methods in Machine Learning are detailed in [17]. Furthermore, the IB methodology connects to many timely aspects, such as Capital Investment [18], Distributed Learning [19], Deep Learning [20], and Convolutional Neural Networks [21].

The PF, which was first introduced in [22], is a dual problem to the IB method. In contrast to IB problem, the goal in PF is to minimize $I(\mathsf{Y};\mathsf{W})$ over all test-channels $P_{\mathsf{W}|\mathsf{X}}$ subject to $I(\mathsf{X};\mathsf{W}) \geq C$. To be more formal, the PF function, $R^{\mathsf{PF}} : [0, H(\mathsf{X})] \to \mathbb{R}_+$ is defined as

$$R^{\mathsf{PF}}_{\mathsf{P}_{\mathsf{XY}}}(C) \triangleq \begin{array}{ll} \underset{\mathsf{P}_{\mathsf{W}|\mathsf{X}}}{\text{minimize}} & I(\mathsf{W};\mathsf{Y}) \\ \text{subject to} & I(\mathsf{X};\mathsf{W}) \geq C. \end{array} \tag{2}$$

Note that since the objective function is a convex function of $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$, taking the constraint here with reverse inequality, i.e. $I(\mathsf{X};\mathsf{W}) \leq C$, will induce a trivial solution, i.e. taking X and W independent.

PF is directly connected to Information Combining [9], [16]. For example, if the channel from X to Y is a BMS, then by [9], $P_{\mathsf{W}|\mathsf{X}}$ is a Binary Erasure Channel (BEC). A rather intriguing example is the setting where the pair (X,Y) are jointly Gaussian, where the result of the minimization is zero, since one can use the channel from X to W to describe the less significant bits of X [23]. Furthermore, the additive noise Helper problem studied in [24], is directly linked to the PF. By reformulating the former as an information combining problem, the solution follows directly as was shown in [23].

In this work we address the input symmetric nonbinary setting for the IB and PF functions. We will find conditions on the bivariate source (X,Y) for which the stochastic encoder from X to W can be completely characterized, thus extending the binary examples from [7], [9] and [5]. Omitted proofs are at the arXiv version of this paper [25].

II. Notations and Basic Properties

We denote by Δ_n the n dimensional probability simplex, $\mathbf{q} \in \Delta_n$ the marginal probability vector of X, and T the transition matrix from X to Y, i.e.,

$$T_{ij} \triangleq \mathsf{P}\left(\mathsf{Y}=i|\mathsf{X}=j\right), \qquad 1 \leq i \leq m, 1 \leq j \leq n. \tag{3}$$

We further rewrite (1) with explicit dependence on $\mathbf{q}$ and T as $R_T(\mathbf{q}, C) = R(C) = R_{\mathsf{P}_{\mathsf{XY}}}(C)$. The entropy of an n-ary probability vector $\mathbf{p} \in \Delta_n$ is denoted by $h_n(\mathbf{p})$.

The following tight cardinality bound was established in [26]. It was actually already proved for the corresponding dual problem, namely the IB Lagrangian, in [27]. But since $R_T(\mathbf{q}, C)$ is generally not a strictly convex function of C, the result in [27] cannot be directly applied for our problem (1).

Lemma 1 ([26, Th. 9]): The optimization over W in (1) can be restricted to $|\mathcal{W}| \leq n$.

As we have already mentioned, the IB function defined in (1) is closely related to the Conditional Entropy Bound (CEB) problem studied in [7], which is given by

$$F_T(\mathbf{q}, x) \triangleq \begin{array}{ll} \underset{\mathsf{W}\to\mathsf{X}\to\mathsf{Y}}{\text{minimize}} & H(\mathsf{Y}|\mathsf{W}) \\ \text{subject to} & H(\mathsf{X}|\mathsf{W}) \geq x. \end{array} \tag{4}$$

Remark 1: Note that originally in [7] the conditional entropy constraint was given with equality, and equivalence to the inequality setting was established in [7, Theorem 2.5].

It turns out that the aforementioned problem is closely connected to the IB function.

Proposition 2.1: The IB function defined in (1) is equivalent to the CEB function defined in (4).

The latter result implies that we can utilize the properties of $F_T(\mathbf{q}, x)$ developed in [7] for our problem in a straightforward manner, an aspect that we will heavily rely on in Section III.

In a very similar manner to Proposition 2.1, we can redefine the Privacy Funnel problem defined in (2) as follows.

$$F^{\mathsf{PF}}_T(\mathbf{q}, x) \triangleq \begin{array}{ll} \underset{\mathsf{P}_{\mathsf{W}|\mathsf{X}}}{\text{maximize}} & H(\mathsf{Y}|\mathsf{W}) \\ \text{subject to} & H(\mathsf{X}|\mathsf{W}) \leq x. \end{array} \tag{5}$$

We have the following characterization of $F^{\mathsf{PF}}_T(\mathbf{q}, x)$.

Theorem 1: The function $F^{\mathsf{PF}}_T(\mathbf{q}, \cdot)$ is concave on the compact convex domain $\{x : 0 \leq x \leq h_n(\mathbf{q})\}$ and for each $(\mathbf{q}, x)$, the maximum is attained with W taking at most $n+1$ values. The proof of this theorem is similar to [7, Theorem 2.3] and is omitted here due to space limitations.

III. The Symmetric Information Bottleneck

In this section we will give a characterization of the achieving conditional distributions and the value of the problem defined in (1) for specific class of input symmetric channels. We begin with the definitions of symmetric group of permutation, symmetry group of stochastic matrix and input symmetric channel [7].

Definition 1: Let $\mathscr{S}_n$ denote the representation of the symmetric group of permutation of n objects by the $n \times n$ permutation matrices. Let $\mathscr{S}_n \times \mathscr{S}_m$ be the representation of the direct product group by the pairs (G, Π), $G \in \mathscr{S}_n$; $\Pi \in \mathscr{S}_m$ with the composition $(G_1, \Pi_1)(G_2, \Pi_2) = (G_1G_2, \Pi_1\Pi_2)$.

For an $m \times n$ stochastic matrix T, (an n input, m output channel), let $\mathscr{G}$ be the set $\{(G, \Pi) \in \mathscr{S}_n \times \mathscr{S}_m | TG = \Pi T\}$, and let $\mathscr{G}_i$ ($\mathscr{G}_o$) be the projections of $\mathscr{G}$ on the first (second) factor. If $TG_1 = \Pi_1 T$, $TG_2 = \Pi_2 T$, then $TG_1 G_2 = \Pi_1 \Pi_2 T$ which shows that $\mathscr{G}$, $\mathscr{G}_i$,$\mathscr{G}_o$ are subgroups of the finite groups $\mathscr{S}_n \times \mathscr{S}_m$, $\mathscr{S}_n$, $\mathscr{S}_m$ respectively. $\mathscr{G}$ is the symmetry group of T, $\mathscr{G}_i$ ($\mathscr{G}_0$) is the input (output) symmetry group.

The channel defined by T will be called input (output) symmetric if $\mathscr{G}_i$ ($\mathscr{G}_o$) is transitive (a subgroup of $\mathscr{S}_n$ is transitive if each element of $\{1,\dots,n\}$ can be mapped to every other element of $\{1,\dots,n\}$ by some member of the subgroup). T is said to be symmetric if both $\mathscr{G}_i$ and $\mathscr{G}_o$ are transitive.

We also define the set of $(\mathbf{q}, C)$ for which we will have a complete characterization of the achieving distributions.

Definition 2: Assume that $\{G_\alpha\}_{\alpha=1}^{n} \in \mathscr{S}_n$ is a set of n distinct elements. Let $\phi(\mathbf{p},\lambda) \triangleq h_m(T\mathbf{p}) - \lambda h_n(\mathbf{p})$ and $\mathbf{p}^* = \operatorname{argmin}_{\mathbf{p}\in\Delta_n} \phi(\mathbf{p},\lambda)$. We define the following set for any $\lambda \in [0,1]$:

$$\mathcal{Q} \triangleq \left\{(\mathbf{q},C) : \mathbf{q}{=}\sum_{\alpha=1}^{n} w_a G_\alpha \mathbf{p}^*, \mathbf{w}{\in}\Delta_n, C{=}1{-}h_n(\mathbf{p}^\star)\right\}. \tag{6}$$

Equipped with this definition we are ready to state our main theorem here.

Theorem 2: Assume that T is input symmetric stochastic matrix with input symmetry group $\mathscr{G}_i$ of order n. Then for every $(\mathbf{q}, C) \in \mathcal{Q}$ defined in (6), the optimal test-channel from W to X is a modulo-additive channel.

Note if $\mathbf{q}$ is uniform over n, then it always in $\mathcal{Q}$, as taking $\mathbf{w}$ to be uniform over n, we obtain

$$\mathbf{q} = \sum_{\alpha=1}^{n} w_a G_\alpha \mathbf{p}^* = \frac{1}{n}\sum_{\alpha=1}^{n} G_\alpha \mathbf{p}^* = \mathbf{u}_n, \tag{7}$$

where $\mathbf{u}_n$ is an n-ary uniform probability vector. This fact induces the following corollary.

Corollary 3.1: Assume that T is input symmetric stochastic matrix with input symmetry group $\mathscr{G}_i$ of order n and X is uniformly distributed over n. Then for every $C \in [0, \log n]$, the test-channel from W to X is a modulo-additive noise channel and W is uniform over n.

A particular case for which T is input symmetric, is when the channel from X to Y is a modulo-additive noise channel, i.e., there exist a random variable Z, with probability vector $\mathbf{z}$ such that $\mathsf{Y} = \mathsf{X} \oplus \mathsf{Z}$, where $\oplus$ is modulo n addition. An equivalent representation of the modulo-additive noise channel is using circulant matrix. A circulant matrix $A \in M_n(\mathbb{F})$ [28, p. 33] has the form

$$A = \begin{pmatrix} a_1 & a_2 & & \cdots & a_n \\ a_n & a_1 & a_2 & \cdots & a_{n-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_2 & a_3 & \cdots & a_n & a_1 \end{pmatrix}, \tag{8}$$

i.e, the entries in each row are cyclic permutations of those in the first. In this case we have the following corollary.

Corollary 3.2: If $T = A$ as defined in (8), then some modulo additive test channel from W to X achieves $R_A(\mathbf{q}, C)$. In particular, there exists an n-ary random variable V, with $H(\mathsf{V}) = \log n - C$, such that $\mathsf{X} = \mathsf{W} \oplus \mathsf{V}$ achieves $R_A(\mathbf{q}, C)$.

Although this result greatly simplifies the optimization space, it does not give a precise analytical solution to the problem. In the following subsection, we provide an example, for which the achieving distribution and the objective function value can be fully characterized.

A. Hamming Channels

Let $T = T_\alpha = \alpha I_n + (1-\alpha)n^{-1}E_n$, where I_n is the $n \times n$ identity matrix, E_n the all ones matrix, and $0 \leq \alpha \leq 1$. The channel with transition matrix T_α is called a Hamming channel with parameter α. Note that T_α is in particular a circulant matrix, therefore by Corollary 3.2 the optimal channel from W to X is a modulo-additive channel. Thus, (4) can be reformulated as follows.

$$F_T(\mathbf{q}, x) \triangleq \begin{array}{ll} \underset{\mathbf{v}\in\Delta_n}{\text{minimize}} & h_n(T_\alpha \mathbf{v}) \\ \text{subject to} & h_n(\mathbf{v}) \geq x. \end{array} \tag{9}$$

The optimization problem defined in (9) is identical to the problem considered in [29]. Furthermore, it was solved for the Hamming channel and the achieving distribution was found.

Lemma 2 ([29, Lemma 7]): For $n \times n$ Hamming channel T_α the solution to (9) is attained for

$$\mathbf{v} = \beta\mathbf{e} + (1-\beta)\mathbf{u}_n. \tag{10}$$

where $\mathbf{e}$ is any standard basis vector of Δ_n.

Since $\mathbf{v}$ is determined by a single parameter β and satisfies $h_n(\mathbf{v}) = \log n - C$, we can find β explicitly as follows:

$$\begin{aligned} &C = \log n - h_n(\mathbf{v}) \\ &= \frac{n-1}{n}(1-\beta)\log(1-\beta) + \frac{\beta n + 1 - \beta}{n}\log(\beta n + 1 - \beta) \\ &\triangleq g_n(\beta). \end{aligned}$$

Thus, β can be recovered from C as $\beta = g_n^{-1}(C)$. In summary, we have the following theorem.

Theorem 3: Assume that T is a Hamming channel with parameter α, then $R_T(\mathbf{u}_n, C)$ is attained with a Hamming channel with parameter $\beta = g_n^{-1}(C)$ and is given by

$$R_T(\mathbf{u}_n{,}C){=}\frac{1{+}(n{-}1)\alpha\beta}{n}\log(1{+}(n{-}1)\alpha\beta){+}\frac{1{-}\alpha\beta}{n}\log(1{-}\alpha\beta). \tag{11}$$

B. Examples

Now let us consider two special cases.

1) BMS: Assume that the channel from X to Y is a BMS channel. Let $\mathbf{z}$ be an m-ary probability vector and G_m be the $m \times m$ anti-diagonal matrix with unit entries. The respective transition matrix in this case is $T = [\mathbf{z}, G_m\mathbf{z}]$. Note that

$$G_m T = [G_m\mathbf{z}, G_m G_m \mathbf{z}] = [\mathbf{z} G_2 \mathbf{z}] = TG_2. \tag{12}$$

Therefore, T is input symmetric stochastic matrix with input symmetry group $\mathscr{G}_i$ of order 2. Thus, since the only binary-input binary-output symmetric channel is a BSC, combining with Theorem 2, we recover the following result from [9].

Corollary 3.3 ([9, Theorem 2]): Given that the channel from X to Y is a BMS, then BSC channel from X to W maximizes $I(\mathsf{W};\mathsf{Y})$.

The latter result can also be deduced from [30].

2) Ternary-Input Ternary-Output (TITO) Circulant Matrix: The general TITO Circulant Matrix is defined as follows:

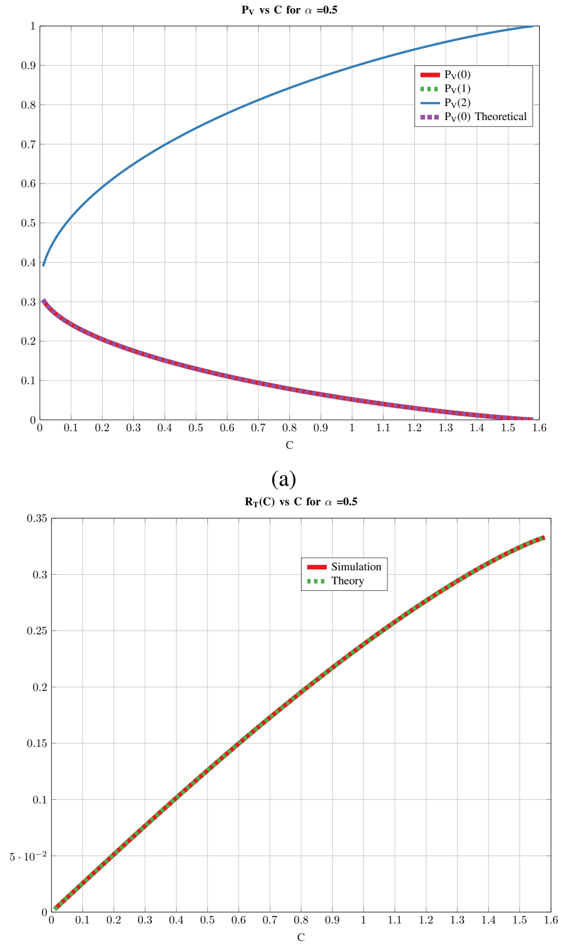
$$T = \begin{pmatrix} 1-\alpha-\beta & \alpha & \beta \\ \beta & 1-\alpha-\beta & \alpha \\ \alpha & \beta & 1-\alpha-\beta \end{pmatrix}. \tag{13}$$

We can further ask if there are values of C such that $R(C)$ can be achieved with W taking at most two points. The following corollary states the opposite.

Corollary 3.4: The minimum cardinality of W that achieves $R(C)$ is exactly 3 for $C \neq 0$.

C. Numerical Simulation

We proceed to verify Theorem 3 via numerical optimization for $n = 3$. Since V is independent of the choice of α, we fix $\alpha = 0.5$ and compare it with respect to the value of C. Figure 2 shows the probability vector V and $R_{T_\alpha}(C)$ for various values of α. We observe that the numerical optimization agrees with theoretical arguments of Theorem 3.



(a)

(b)

Fig. 2: (a) Optimal $\mathbf{v}$ for $\alpha = 0.5$ vs C. (b) $R_{T_\alpha}(C)$ vs C for $\alpha = 0.5$.

IV. The Symmetric Privacy Funnel

In this section we consider a special symmetric setting for the PF problem (5) for which the transition matrix from X to Y is an input symmetric stochastic matrix as defined in Definition 1.

Theorem 4: Let T be an input symmetric stochastic matrix with input symmetry group $\mathscr{G}_i$ of order n, and X be a uniformly distributed random variable. Let $(G_1 = I, G_2, \ldots, G_n) \in \mathscr{G}_i$. Furthermore, denote by $(\mathbf{p}^*, \lambda^*)$ a pair for which

$$\phi(\mathbf{u}, \lambda^*) = \phi(\mathbf{p}^*, \lambda^*) \geq \phi(\mathbf{p}, \lambda^*) \quad \forall \mathbf{p} \in \Delta_n. \tag{14}$$

Then, for every $C \leq C^* \triangleq \log n - h_n(\mathbf{p}^*)$, the transition matrix from W to X, given by

$$B = \begin{pmatrix} \mathbf{p}^* & G_2\mathbf{p}^* & \cdots & G_n\mathbf{p}^* & \mathbf{u} \end{pmatrix}, \tag{15}$$

achieves (2). Moreover,

$$R^{\mathrm{PF}}_{\mathsf{P}_{\mathsf{XY}}}(C) = C \cdot \frac{\log n - h_n(T\mathbf{p}^*)}{\log n - h_n(\mathbf{p}^*)}. \tag{16}$$

Also, (15) implies that the transition matrix from X to W is a class of noisy n-ary symmetric erasure channel.

Note that the optimization procedure in (14) is performed once for every $C \in [0, \log_2 n - h_n(\mathbf{p}^*)]$. Moreover, for $C \in [0, \log_2 n - h_n(\mathbf{p}^*)]$, the optimal test-channel from X to W is no longer symmetric as we show using a numerical example. We now provide some examples that illustrate Theorem 4.

A. Examples

We begin with the simplest scenario where X is a binary random variable. Plugging this choice in Theorem 4 and noting that $\mathbf{p}^* = \mathbf{e}$ in this case, results in the following corollary.

Corollary 4.1: Assume that the channel from X to Y is a BMS, then, BEC test-channel $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ with parameter $\epsilon = 1 - C$ minimizes $I(\mathsf{Y};\mathsf{W})$ subject to $I(\mathsf{X};\mathsf{W}) = C$.

Note that this result recovers [9, Theorem 1], but here with only one-sided symmetry restriction.

We further illustrate Theorem 4 using numerical optimization for a particular choice of the channel from X to Y being a symmetric TITO with parameters $(\alpha, \beta) = (0.1, 0.05)$, as defined in (13). For this choice of channel parameters, $C^* = 0.59$. In Figure 3 we compare the results of global optimization solution of (2) versus choosing $\mathsf{P}_{\mathsf{W}|\mathsf{X}}$ be the respective optimal input-symmetric channel as described in Theorem 4 for various values of C. We observe that our results from Theorem 4 agree with the brute-force numerical optimization for all values of $C \in [0, C^*]$. For values greater than C^*, we observe that the curve which is restricted to input symmetric transition matrices from X to W, is sub-optimal. In this region of link capacity, the numerical optimization achieves lower rates. By carefully observing the numerical solution, one can notice that the optimal test-channel in this region is no longer input symmetric.

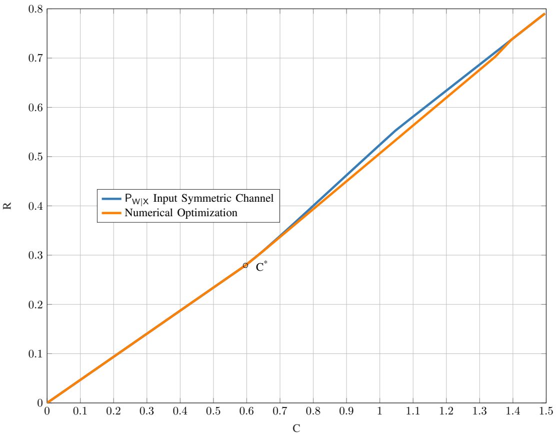


Fig. 3: Optimal R for $\alpha = 0.1$, $\beta = 0.05$ vs C

V. Outlook

As said, the Information Bottleneck and Privacy Funnel are two dual optimization problems which have been applied in a variety of emerging applications such as Deep Neural Networks, Privacy Algorithms, and design of Polar Codes [17]. It also interesting to consider rather more classical use-cases, i.e, multi-user channel capacity and Noisy Source Coding problems. A comprehensive summary of the different relations between the IB and Privacy Funnel problems has been presented in [26].

Acknowledgment

This work has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630 and by the WIN consortium via the Israel minister of economy and science.

References

[1] N. Tishby, F. C. N. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 1999, p. 368–377.

[2] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. 16, pp. 406–411, Jul. 1970.

[3] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, pp. 14–20, Jan. 1972.

[4] N. Slonim, "The information bottleneck: Theory and applications," Ph.D. dissertation, Hebrew University of Jerusalem, Jerusalem, Israel, 2002.

[5] A. Zaidi, I. E. Aguerri, and S. S. (Shitz), "On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views," *Entropy*, vol. 22, no. 2, 2020.

[6] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications I," *IEEE Trans. Inf. Theory*, vol. 19, pp. 769–772, 1973.

[7] 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.

[8] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[9] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1313–1325, Apr. 2005.

[10] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information Bottleneck for Gaussian Variables," *J. Mach. Learn. Res.*, vol. 6, pp. 165–188, Dec. 2005.

[11] A. E. Gamal and Y. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[12] D. Guo, S. Shamai (Shitz), and S. Verdú, "The Interplay Between Information and Estimation Measures," *Found. Trends Signal Process.*, vol. 6, no. 4, pp. 243–429, 2012.

[13] R. Bustin, M. Payaro, D. P. Palomar, and S. Shamai (Shitz), "On MMSE crossing properties and implications in parallel vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 818–844, Feb. 2013.

[14] P. Gács and J. Körner, "Common information is far less than mutual information," *Probl. Contr. Inform. Theory*, vol. 2, no. 2, pp. 149–162, 1973.

[15] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4995–5010, Nov. 2009.

[16] I. Land and J. Huber, "Information Combining," *Found. Trends Commun. Inf. Theory*, vol. 3, no. 3, pp. 227–330, Nov. 2006.

[17] Z. Goldfeld and Y. Polyanskiy, "The Information Bottleneck Problem and its Applications in Machine Learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[18] E. Erkip and T. M. Cover, "The Efficiency of Investment Information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[19] P. Farajiparvar, A. Beirami, and M. Nokleby, "Information Bottleneck Methods for Distributed Learning," in *Proc. 56th Annu. Allerton Conf. Commun., Control Comput.*, 2018, pp. 24–31.

[20] R. A. Amjad and B. C. Geiger, "Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle," *IEEE Trans. Pattern Anal.*, vol. 42, no. 9, pp. 2225–2239, Sep. 2020.

[21] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 435–442, Jan. 2021.

[22] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5011–5038, Aug. 2017.

[23] S. Shamai, "The information bottleneck: A unified information theoretic view," National Conference on Communications (NCC2021), Jul. 2021, plenary Address.

[24] S. I. Bross and A. Lapidoth, "The additive noise channel with a helper," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019, pp. 1–5.

[25] M. Dikshtein and S. Shamai, "A class of nonbinary symmetric information bottleneck problems," *CoRR*, vol. abs/2110.00985, 2021. [Online]. Available: https://arxiv.org/abs/2110.00985

[26] S. Asoodeh and F. P. Calmon, "Bottleneck Problems: An Information and Estimation-Theoretic View," *Entropy*, vol. 22, no. 11, p. 1325, 2020.

[27] P. Harremoes and N. Tishby, "The Information Bottleneck Revisited or How to Choose a Good Distortion Measure," in *Proc. 2007 IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 566–570.

[28] R. A. Horn, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 2012.

[29] H. S. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.

[30] N. Chayat and S. Shamai, "Extension of an entropy property for binary input memoryless symmetric channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1077–1079, 1989.

Zero-error source coding when side information may be present

Nicolas Charpenay*, Maël Le Treust †, Aline Roumy ‡
*IRISA, 263 avenue du Général Leclerc, 35 042 Rennes cedex, France, nicolas.charpenay@irisa.fr
†ETIS UMR 8051, CY Université, ENSEA, CNRS, 6 avenue du Ponceau, 95014 Cergy cedex, France, mael.le-treust@ensea.fr
‡INRIA Rennes, Campus de Beaulieu, 35042 Rennes cedex, France, aline.roumy@inria.fr

***Abstract*—Zero-error source coding when side-information (SI) may be present is a fundamental building block of interactive real-world compression systems. In such a scenario, the side information may represent an image that could have been requested previously by the user. We aim at designing a two layer zero-error coding scheme that adapts to the presence or absence of the side information at the decoder. The scenario we consider involves two decoders and two noiseless channels, the first channel to both decoder and the second channel of additional information to decoder 2 only. The side information is available at the encoder and decoder 1, but not at decoder 2. By using a random coding argument we characterize the zero-error achievable rate region. The code construction relies on coset partitioning obtained from a linear code. The encoder sends the coset of the source sequence on the first channel to all decoders, and sends the index of the source sequence in its coset on the second channel to decoder 2.**

I. Introduction

We consider the scenario described in Fig. 1 in which the information source X is correlated to the side information (SI) Y observed by the encoder and decoder 1 only. The information is sent through a noiseless channel at rate R_1 to both decoders and an additional noiseless channel at rate R_2 to decoder 2, which does not observe the SI. All decoders must recover the source X with zero-error, i.e. with a probability of error equal to zero, which is a more restrictive assumption than a vanishing probability of error.

This scenario arises in interactive compression, where the user can randomly access part of the data directly in the compressed domain. A source sequence X^n models the smallest entity that can be requested, for instance a file of a database, a frame of a video, or a block of an omnidirectional image in [1]. Upon request of X^n, and if no request has been previously made (case of decoder 2 in Fig. 1), the encoder sends the complete representation of the data $(f_1(X^n, Y^n), f_2(X^n, Y^n))$ at rate $R_1 + R_2$. If, instead, the block Y^n has already been requested (case of decoder 1), the encoder sends only a part of the compressed representation namely $f_1(X^n, Y^n)$ to complete Y^n. Moreover, we consider the zero-error version of this problem, as zero-error source coding is a fundamental building block of practical video coding schemes. We therefore seek for the set of rates (R_1, R_2), which can be achieved in this scenario.

A way to achieve zero-error coding is to use conditional coding, and send the source X to decoder 1 at rate $R_1 = H(X|Y)$, since both encoder and decoder 1 observe the SI Y. Then, to recover the source X, decoder 2 needs to obtain the SI Y, which requires a rate of $R_2 = H(Y) \geq I(X;Y)$.

In order to be exploitable by both decoders, part of the information sent through the common channel must be independent from Y. For this reason our setting is closely related to the Slepian and Wolf (SW) problem in [2], seen as lossless source coding with side information at the decoder only. In [3], Csiszar proved in that linear codes achieve the optimal SW rate region. Several works in [4]–[6] investigate the duality between SW setting and channel coding using linear codes, as the side-information Y can be seen as the input of a virtual channel with input X. However these tools cannot be straightforwardly adapted to the zero-error setting, as the linear codes proposed also present a vanishing probability of error.

Our setting can be seen as a zero-error variant with side-informations known at the encoder of the successive refinement problem proposed by Kaspi in [7]; later generalized by Timo et al. in [8] for more than two decoders. Even if the lossy reconstruction of the source makes it fundamentally different from the zero-error setting, there are notable examples that present the same tools as in SW. The side-information scalable source coding (i.e. the decoder 2 has a SI Y' s.t. $X \to Y \to Y'$) in [9] for instance uses nested random binning. This random binning approach was further developed in [10] to give a unified coding scheme that works for both scalable source coding and Wyner-Ziv successive refinement in [11] (i.e. the decoder 2 has a SI Y' s.t. $X \to Y' \to Y$).

In the open problem, the zero-error SW scheme requires to send at rate $H(X)$ to the decoder with side information, see [12]–[15]. In [16], Ma and Cheng use linear codes in a zero-error SW restriction, under symmetry assumptions on the source. However, a zero-error SW coding scheme in our setting does not use at all the side information knowledge at the encoder. Therefore, we study the role of the side information at the encoder with a zero-error constraint when side information may be present at the decoder.

In this paper, we characterize the set of rate pairs that are achievable with zero-error source codes, as depicted in Fig. 1. More precisely, we show that the pair of rates $(R_1, R_2) = (H(X|Y), I(X;Y))$ is achievable and moreover, it is the corner-point of the set of achievable pair of rates. Our achievability result relies on a random coding argument. We

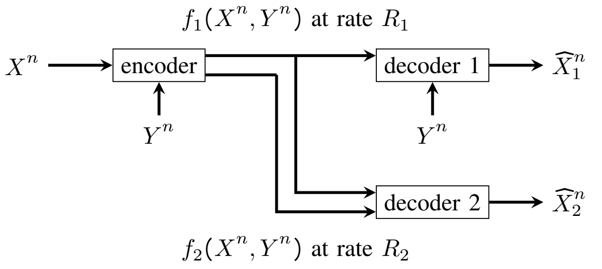


Fig. 1: Source coding when side-information may be present.

use Csiszar and Körner's method of types [17, Chapter 2] in order to calibrate a linear code which is used to partition the set of source sequences. The encoder sends the coset of the source sequence to all decoders and the index of the source sequence in its coset to decoder 2. We show that the zero-error property is satisfied and the corresponding rates converge to the pair of target rates $(H(X|Y), I(X;Y))$.

A. Notations

Random variables and their realizations are represented by uppercase letters (e.g., X) and lowercase letters (e.g., x), respectively; and their set of possible values with the corresponding calligraphic letters (e.g., $\mathcal{X}$). We denote by $|\cdot|$ the cardinality of a set. We denote a sequence of symbols by $x^n = (x_1, ..., x_n)$. The set of probability distributions over a finite set $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The distribution of a random variable X is denoted by $P_X \in \mathcal{P}(\mathcal{X})$. When computing entropies with other distributions than P_X, we specify it in subscript (e.g. $H_Q(X)$ is computed with the distribution Q). The conditional distribution of a random variable X knowing Y is denoted by $P_{X|Y}$, and the joint distribution is denoted by $P_{X,Y}$. We denote by $\{0,1\}^*$ the set of binary words. Throughout the paper the logarithms are in base two.

II. Problem Statement and Main Result

The setting of Fig. 1 is described by:

- Two finite sets $\mathcal{X}$, $\mathcal{Y}$ and a pair of random variables $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ drawn with the distribution $P_{X,Y}$.
- An encoder that observes the realizations of (X,Y).
- Two decoders, where only decoder 1 observes the realizations of the side-information Y.
- The encoder transmits over a first channel to both decoders, and a second channel to decoder 2 only.
- We denote by $n \in \mathbb{N}^\star = \mathbb{N} \setminus \{0\}$ the block size of the coding scheme. For n iterated source uses, we denote by (X^n, Y^n) the sequences of independent copies of (X,Y).

Definition II.1 *Given $n \in \mathbb{N}^\star = \mathbb{N} \setminus \{0\}$, $(R_1^{(n)}, R_2^{(n)}) \in [0, +\infty)^2$, a $(n, R_1^{(n)}, R_2^{(n)})$-zero-error source code consists of encoding functions (f_1, f_2) that assigns variable-length binary sequences and decoding functions (g_1, g_2) defined by:*

$$f_1 : \mathcal{X}^n \times \mathcal{Y}^n \to \{0,1\}^*, \quad f_2 : \mathcal{X}^n \times \mathcal{Y}^n \to \{0,1\}^*, \quad (1)$$

$$g_1 : \{0,1\}^* \times \mathcal{Y}^n \to \mathcal{X}^n, \quad g_2 : (\{0,1\}^*)^2 \to \mathcal{X}^n, \quad (2)$$

that satisfy

$$R_1^{(n)} = \frac{1}{n}\mathbb{E}\Big[l\big(f_1(X^n, Y^n)\big)\Big], \quad R_2^{(n)} = \frac{1}{n}\mathbb{E}\Big[l\big(f_2(X^n, Y^n)\big)\Big],$$

where $l(\cdot)$ denotes the length of a binary word, and that satisfy the zero-error property, i.e. $X^n = g_1\big(f_1(X^n, Y^n), Y^n\big) = g_2\big(f_1(X^n, Y^n), f_2(X^n, Y^n)\big)$ with probability 1.

Definition II.2 *A rate pair $(R_1, R_2) \in [0, +\infty)^2$ is achievable if there exists a sequence of $(n, R_1^{(n)}, R_2^{(n)})$-zero-error source codes such that*

$$\lim_n R_1^{(n)} = R_1, \quad \lim_n R_2^{(n)} = R_2. \quad (3)$$

We denote by $\mathcal{R}$ the zero-error achievable rate region.

Theorem II.3

$$\mathcal{R} = \Big\{(R_1, R_2),\ R_1 \geq H(X|Y),\ R_1 + R_2 \geq H(X)\Big\}. \quad (4)$$

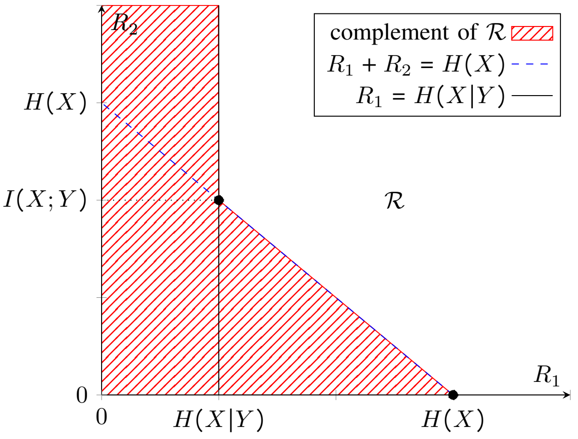


Fig. 2: Zero-error achievable rate region $\mathcal{R}$.

Proof. [Converse of Theorem II.3] In this setting, each decoder must retrieve X with zero-error. Using Shannon lossless source coding result [18, Theorem 5.3.1] and Slepian-Wolf Theorem [2, Theorem 2] on each decoder, we have $R_1 \geq H(X|Y)$ and $R_1 + R_2 \geq H(X)$, as the zero-error source codes are a subclass of lossless codes considered for these converses. □

III. Achievability Proof of Theorem II.3

In order to prove Theorem II.3, we show that

$$\big(H(X|Y), I(X;Y)\big) \in \mathcal{R}. \quad (5)$$

In order to complete the achievability result we use a time sharing with the point $\big(H(X), 0\big)$, which is known to be achievable by compressing X using a Huffman code and sending the resulting binary sequence via f_1.

A. Preliminaries

Definition III.1 (Type) *For all pair of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, the joint type is the distribution from $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denoted Q_{x^n,y^n} that satisfies for all $(x', y') \in \mathcal{X} \times \mathcal{Y}$*

$$Q_{x^n,y^n}(x', y') = \frac{1}{n} \left| \left\{ i \le n \;\middle|\; (x_i, y_i) = (x', y') \right\} \right| . \tag{6}$$

We denote the marginal types by Q_{x^n} and Q_{y^n}, respectively. We denote the conditional type of x^n knowing y^n by $Q_{x^n|y^n}$.

The n-discretized probability simplex $\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ is the set of types that are achievable using sequences of length n.

We denote by Q_{X^n,Y^n} the random variable of the joint type of the random sequences (X^n, Y^n). We denote the random variables of their conditional and marginal types by $Q_{X^n|Y^n}$, Q_{X^n} and Q_{Y^n}, respectively.

Definition III.2 (Type class, V-shell) *For all type $\pi \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$, we denote the type class by $\mathcal{T}_\pi$*

$$\mathcal{T}_\pi = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;\middle|\; Q_{x^n,y^n} = \pi \right\} . \tag{7}$$

Given a conditional type $V \in \mathcal{P}(\mathcal{X})^{|\mathcal{Y}|}$, the V-shell of a sequence y^n is the set $\mathcal{T}_V(y^n) = \left\{ x^n \in \mathcal{X} \;\middle|\; Q_{x^n|y^n} = V \right\}$.

Definition III.3 (Generator/parity matrix, syndrome, coset) *Let $\mathcal{A}$ be a finite set such that $|\mathcal{A}|$ is prime, so we can give $\mathcal{A} \simeq \mathbb{Z}/|\mathcal{A}|\mathbb{Z}$ a field structure. For all $n, k \in \mathbb{N}^\star$, we denote by $\mathcal{M}_{n,k}(\mathcal{A})$ the set of $n \times k$ matrices over the finite field $\mathcal{A}$.*

Let $k \in \mathbb{N}^\star$, a generator matrix is a matrix $\boldsymbol{G} \in \mathcal{M}_{n,k}(\mathcal{A})$. An associated parity matrix is a matrix $\boldsymbol{H} \in \mathcal{M}_{n-k,n}(\mathcal{A})$ such that Im $\boldsymbol{G}$ = Ker $\boldsymbol{H}$, where Im and Ker denote the image and the kernel, respectively.

The syndrome of a sequence $a^n \in \mathcal{A}^n$ is $\boldsymbol{H}x^n$. The coset associated to the syndrome $\boldsymbol{H}a^n$ is the set Im $\boldsymbol{G} + a^n = \{\tilde{a}^n \in \mathcal{A}^n \mid \boldsymbol{H}\tilde{a}^n = \boldsymbol{H}a^n\}$.

B. Coding scheme

For all $n \in \mathbb{N}^\star$, we show the existence of a sequence of $(n, R_1^{(n)}, R_2^{(n)})$-zero-error source codes that achieves the corner-point $\big(H(X|Y), I(X;Y)\big)$ of the zero-error rate region $\mathcal{R}$. Our proof is based on a linear code adjusted depending on Q_{X^n,Y^n}, and coset partitioning of the Hamming space.

We assume w.l.o.g. that $P_{X,Y} \neq P_X P_Y$. We also assume w.l.o.g. that $|\mathcal{X}|$ is prime number by padding (i.e. extending with zeros) $P_{X,Y}$ if necessary. We fix the block-length n and a constant parameter $\delta \in (0; \log|\mathcal{X}| - H(X|Y))$ that will represent a rate penalty.

- Random code generation: For each pair of sequences (x^n, y^n), we define the parameter

  $$k \doteq \left\lceil n - n \frac{H_{Q_{x^n,y^n}}(X|Y) + \delta}{\log|\mathcal{X}|} \right\rceil^+ . \tag{8}$$

  where $\lceil \cdot \rceil$ denotes the ceiling function and $(\cdot)^+$ denotes $\max(\cdot, 0)$. We denote by K the random variable induced by k defined in (8), for the random sequences (X^n, Y^n). A generator matrix $\mathbf{G} \in \mathcal{M}_{n,n}(\mathcal{X})$ is randomly drawn, with i.i.d. entries drawn according to the uniform distribution on $\mathcal{X}$. If $K \neq 0$, let $\mathbf{G}_K$ be the matrix obtained by extracting the K first lines of $\mathbf{G}$, and $\mathbf{H}_K$ a parity matrix associated to $\mathbf{G}_K$.

  The random code $\mathcal{C}$ consists of the set of random matrices $\mathcal{C} = \{(\mathbf{G}_k, \mathbf{H}_k), 1 \le k \le n\}$. Before the transmission starts, a code realization is chosen and revealed to the encoder and both decoders.
- Encoding function f_1: Let $E \in \{0, 1\}$ be such that $E = 0$ if $K \neq 0$ and $\big(\text{Im } \mathbf{G}_K + X^n\big) \cap \mathcal{T}_{Q_{X^n|Y^n}}(Y^n) = \{X^n\}$; $E = 1$ otherwise. Then we define

  $$f_1(X^n, Y^n) = \begin{cases} b(Q_{X^n,Y^n}, E, \mathbf{H}_K X^n) & \text{if } E = 0, \\ b(Q_{X^n,Y^n}, E, X^n) & \text{if } E = 1, \end{cases} \tag{9}$$

  where $b(\cdot)$ denotes the binary expansion.
- Encoding function f_2: If $E = 0$, the index of X^n in its coset Im $\mathbf{G}_K + X^n$ is compressed using a Huffman code with the distribution P_{X^n}. Let $B(\mathbf{G}_K, X^n, Y^n)$ be the resulting binary sequence, then we set

  $$f_2(X^n, Y^n) = B(\mathbf{G}_K, X^n, Y^n). \tag{10}$$

  Otherwise, $f_2(X^n, Y^n) = 0$.
- Decoding function g_1: It observes $f_1(X^n, Y^n)$ and extracts E and Q_{X^n,Y^n}. If $E = 1$,

  $$g_1(f_1(X^n, Y^n), Y^n) = X^n. \tag{11}$$

  Otherwise $E = 0$, it extracts $\mathbf{H}_K X^n$ and determines the coset Im $\mathbf{G}_K + X^n$. Moreover, by using Q_{X^n,Y^n} and Y^n it determines the $Q_{X^n|Y^n}$-shell $\mathcal{T}_{Q_{X^n|Y^n}}(Y^n)$, and therefore returns an element

  $$g_1(f_1(X^n, Y^n), Y^n) \in \big(\text{Im } \mathbf{G}_K + X^n\big) \cap \mathcal{T}_{Q_{X^n|Y^n}}(Y^n).$$
- Decoding function g_2: It observes $f_1(X^n, Y^n)$ and extracts E and Q_{X^n,Y^n}. If $E = 0$, it extracts $\mathbf{H}_K X^n$ and determines the coset Im $\mathbf{G}_K + X^n$, and it returns $g_2\big(f_1(X^n, Y^n), f_2(X^n, Y^n)\big)$, the element of Im $\mathbf{G}_K + X^n$ with index $f_2(X^n, Y^n)$. If $E = 1$, it returns

  $$g_2\big(f_1(X^n, Y^n), f_2(X^n, Y^n)\big) = X^n.$$

Remark III.4 *The parameter K is selected so that when $K > 0$, the number of parity bits of the linear code asymptotically matches the conditional entropy: $\frac{(n-K)\log|\mathcal{X}|}{n} = H_{Q_{X^n,Y^n}}(X|Y) + \delta + O\left(\frac{1}{n}\right)$.*

C. Zero-error property

We now prove that the code built in Section III-B satisfies the zero-error property. It is clear that both decoders retrieve X^n with zero-error when $E = 1$.

If $E = 0$, then by definition of E we have $(\text{Im } \mathbf{G}_K + X^n) \cap \mathcal{T}_{Q_{X^n|Y^n}}(Y^n) = \{X^n\}$, hence $g_1\big(f_1(X^n, Y^n), Y^n\big) = X^n$ with probability 1. On the other hand, $f_2(X^n, Y^n) = B(\mathbf{G}_K, X^n, Y^n)$, so the element of Im $\mathbf{G}_K + X^n$ with index $f_2(X^n, Y^n)$ is X^n. Thus, $g_2\big(f_1(X^n, Y^n), f_2(X^n, Y^n)\big) = X^n$ with probability 1.

D. Rate analysis

Now we prove that for all parameter $\delta > 0$, the sequence of rates of the codes built in Section III-B satisfy

$$R_1^{(n)} \underset{n\to\infty}{\to} H(X|Y) + \delta, \qquad R_2^{(n)} \underset{n\to\infty}{\to} I(X;Y). \tag{12}$$

Lemma 1 (Large deviations) *Let X' be a random variable such that $P_{X'}$ is the uniform distribution over $\mathcal{X}$. Then for each pair of sequences (x^n, y^n), we have:*

$$Pr\left(Q_{X'^n,y^n} = Q_{x^n,y^n}\right) = 2^{nH_{Q_{x^n,y^n}}(X|Y)-n\log|\mathcal{X}|+o(n)} \tag{13}$$

Proof. Since $P_{X'}$ is uniform:

$$\begin{aligned}\Pr\left(Q_{X'^n,y^n} = Q_{x^n,y^n}\right) &= |\mathcal{X}|^{-n}\left|\mathcal{T}_{Q_{x^n|y^n}}(y^n)\right| \qquad (14)\\ &= 2^{-n\log|\mathcal{X}|}2^{nH_{Q_{x^n,y^n}}(X|Y)+o(n)},\end{aligned}$$

as [17, Lemma 2.5] gives the asymptotic size of the $Q_{x^n|y^n}$-shell $\mathcal{T}_{Q_{x^n|y^n}}(y^n)$.
□

Probability of decoding ambiguity. We need to estimate $\Pr(E = 1)$. We have $E = 1$ *iff* $K = 0$ or there exists $(\alpha_1, ..., \alpha_K) \in \mathcal{X}^K \setminus \{0, ..., 0\}$ such that $Q_{\left(X^n+\sum_{i\le K}\alpha_i\mathbf{G}_K^{(i)}\right),Y^n} = Q_{X^n,Y^n}$, where $\mathbf{G}_K^{(i)}$ denotes the i-th column of $\mathbf{G}_K$. Thus

$$\begin{aligned}&\Pr(E = 1) \le \Pr(K = 0) \qquad (15)\\ &+ \Pr\Bigg(\bigcup_{\substack{\alpha\in\mathcal{X}^K\\ \alpha\neq 0}}\left[Q_{\left(X^n+\sum_{i\le K}\alpha_i\mathbf{G}_K^{(i)}\right),Y^n} = Q_{X^n,Y^n}\right]\Bigg| K \neq 0\Bigg).\end{aligned}$$

We provide an upper bound on the second term in (15). For all (x^n, y^n) such that $k \neq 0$, we have:

$$\begin{aligned}&\Pr\Bigg(\bigcup_{\substack{\alpha\in\mathcal{X}^k\\ \alpha\neq 0}}\left[Q_{\left(x^n+\sum_{i\le k}\alpha_i\mathbf{G}_k^{(i)}\right),y^n} = Q_{x^n,y^n}\right]\Bigg)\\ &\le \sum_{\substack{\alpha\in\mathcal{X}^k\\ \alpha\neq 0}}\Pr\Bigg(Q_{\left(x^n+\sum_{i\le k}\alpha_i\mathbf{G}_k^{(i)}\right),y^n} = Q_{x^n,y^n}\Bigg) \qquad (16)\\ &\le|\mathcal{X}|^k 2^{nH_{Q_{x^n,y^n}}(X|Y)-n\log|\mathcal{X}|+o(n)} \qquad (17)\\ &\le 2^{n\log|\mathcal{X}|-nH_{Q_{x^n,y^n}}(X|Y)-\delta n+o(n)}\\ &\quad\times 2^{nH_{Q_{x^n,y^n}}(X|Y)-n\log|\mathcal{X}|+o(n)} \le 2^{-\delta n+o(n)}, \qquad (18)\end{aligned}$$

where (17) comes from Lemma 1 and (18) comes from (8). Therefore,

$$\begin{aligned}&\Pr\Bigg(\bigcup_{\substack{\alpha\in\mathcal{X}^K\\ \alpha\neq 0}}\left[Q_{\left(X^n+\sum_{i\le K}\alpha_i\mathbf{G}_K^{(i)}\right),Y^n} = Q_{X^n,Y^n}\right]\Bigg| K \neq 0\Bigg)\\ &= \sum_{x^n,y^n}\Pr\Big((X^n,Y^n) = (x^n,y^n)\Big| K \neq 0\Big)\\ &\quad\times\Pr\Bigg(\bigcup_{\substack{\alpha\in\mathcal{X}^K\\ \alpha\neq 0}}\left[Q_{\left(X^n+\sum_{i\le K}\alpha_i\mathbf{G}_K^{(i)}\right),Y^n} = Q_{X^n,Y^n}\right]\\ &\qquad\qquad\Big| K \neq 0, (X^n,Y^n) = (x^n,y^n)\Bigg) \qquad (19)\\ &\le \sum_{x^n,y^n}\Pr\Big((X^n,Y^n) = (x^n,y^n)\Big| K \neq 0\Big)2^{-\delta n+o(n)} \qquad (20)\\ &\le 2^{-\delta n+o(n)}, \qquad (21)\end{aligned}$$

where (20) comes from (18) and the fact that $\mathbf{G}$ is independent of (X, Y).

We now provide an upper bound on the first term in (15).

$$\mathcal{S} \doteq \left\{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}), \quad 1 - \frac{H_\pi(X|Y) + \delta}{\log|\mathcal{X}|} \le 0\right\}. \tag{22}$$

Then we have:

$$\begin{aligned}&\Pr(K = 0) \qquad (23)\\ &= \Pr\left(\left\lceil n - n\frac{H_{Q_{X^n,Y^n}}(X|Y) + \delta}{\log|\mathcal{X}|}\right\rceil^+ = 0\right) \qquad (24)\\ &= \Pr\left(n - n\frac{H_{Q_{X^n,Y^n}}(X|Y) + \delta}{\log|\mathcal{X}|} \le 0\right) \qquad (25)\\ &= \Pr(Q_{X^n,Y^n} \in \mathcal{S}) \qquad (26)\\ &= \sum_{\pi\in\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})}\Pr(Q_{X^n,Y^n} = \pi) \qquad (27)\\ &\le |\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})| \sup_{\pi\in\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})}\Pr(Q_{X^n,Y^n} = \pi) \qquad (28)\\ &\le |\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})| \sup_{\pi\in\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})} 2^{-nD(\pi\|P_{X,Y})} \qquad (29)\\ &\le |\mathcal{S}\cap\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})| \sup_{\pi\in\mathcal{S}} 2^{-nD(\pi\|P_{X,Y})} \qquad (30)\\ &\le 2^{-n\inf_{\pi\in\mathcal{S}}D(\pi\|P_{X,Y})+o(n)}, \qquad (31)\end{aligned}$$

where (29) comes from [17, Lemma 2.6]. Since $P_{X,Y} \notin \mathcal{S}$ by definition of δ, we have $\inf_{\pi\in\mathcal{S}} D(P_{X,Y}\|\pi) > 0$. Thus there exists a positive constant $\beta > 0$ such that

$$\Pr(K = 0) \le 2^{-\beta n+o(n)}. \tag{32}$$

Thus by combining (15), (21), (32), we have:

$$\Pr(E = 1) \le 2^{-\delta n+o(n)} + 2^{-\beta n+o(n)}. \tag{33}$$

Rate on the common channel. The encoding function f_1 defined in (9) returns Q_{X^n,Y^n} and E. When $E = 0$, it sends the syndrome $\mathbf{H}_K X^n$ at rate $\frac{n-K}{n}\log|\mathcal{X}|$, otherwise, it sends X^n. Therefore,

$$\begin{aligned}nR_1^{(n)} =&1 + |\mathcal{X}||\mathcal{Y}|\log_2(n+1) + \Pr(E = 1)n\log|\mathcal{X}|\\ &+ \Pr(E = 0)\sum_{x^n,y^n}\Pr\big((X^n,Y^n) = (x^n,y^n)|E = 0\big)\\ &\times(n-k)\log|\mathcal{X}| \qquad (34)\\ \le&1 + |\mathcal{X}||\mathcal{Y}|\log_2(n+1) + \Pr(E = 1)n\log|\mathcal{X}|\end{aligned}$$

$$+ (n - \mathbb{E}[K]) \log|\mathcal{X}| \tag{35}$$
$$\le 1 + |\mathcal{X}||\mathcal{Y}| \log_2(n+1) + \Pr(E=1) n \log|\mathcal{X}|$$
$$+ n\mathbb{E}\Big[H_{Q_{X^n,Y^n}}(X|Y)\Big] + n\delta + 1, \tag{36}$$

where (35) comes from $n - k \ge 0$ for all (x^n, y^n), and (36) comes from (8).

By the law of large numbers [18, Theorem 11.2.1] $\mathbb{E}\Big[H_{Q_{X^n,Y^n}}(X|Y)\Big] \underset{n\to\infty}{\to} H(X|Y)$, and by using (33), we obtain

$$\lim_{n\to\infty} R_1^{(n)} \le H(X|Y) + \delta. \tag{37}$$

Rate on the secondary channel. The encoding function f_2 is defined in (10). If $E = 0$, then $K \neq 0$ and the encoder transmits the index of X^n in its coset. The Huffman algorithm has an average output length $R_2^{(n)}$ that satisfies

$$R_2^{(n)} \le \frac{1}{n}\Big(1 + \sum_{k\neq 0} \Pr(K = k|E = 0)$$
$$\times H(X^n|\mathbf{H}_k X^n, K = k, \mathcal{C}, E = 0)\Big) \tag{38}$$
$$= \frac{1}{n} + \frac{1}{n} H(X^n|K, \mathcal{C}, E = 0)$$
$$- \frac{1}{n} H(\mathbf{H}_K X^n|K, \mathcal{C}, E = 0), \tag{39}$$

where (39) follows from the fact that $\mathbf{H}_K X^n$ is a deterministic function of X^n, given a random code $\mathcal{C}$.

We now provide an upper bound to the last term $-\frac{1}{n}H(\mathbf{H}_K X^n|K, \mathcal{C}, E = 0)$ in (39). To do so, we introduce a new encoding scheme that first encodes the sequences X^n and Y^n with the encoding function f_1, and then encode the output by using an entropy coder. The rate of this code r is upperbounded by $H(f_1(X^n, Y^n)|\mathcal{C}) + 1$.

Moreover, the decoder 1 retrieves X^n with zero error (see Sec. III-C), and the entropy coder is also lossless. Thus r is greater than the rate achieved by a conditional entropy coder that compresses X^n knowing the side information Y^n, whose rate is lower bounded by $nH(X|Y)$.

Therefore, we have

$$nH(X|Y) \le r < H(f_1(X^n, Y^n)|\mathcal{C}) + 1 \tag{40}$$
$$= 1 + H(Q_{X^n,Y^n}, E|\mathcal{C})$$
$$+ \Pr(E = 0) H(\mathbf{H}_K X^n|Q_{X^n,Y^n}, \mathcal{C}, E = 0)$$
$$+ \Pr(E = 1) H(X^n|Q_{X^n,Y^n}, \mathcal{C}, E = 1) \tag{41}$$
$$\le H(\mathbf{H}_K X^n|Q_{X^n,Y^n}, \mathcal{C}, E = 0) + o(n) \tag{42}$$
$$= H(\mathbf{H}_K X^n|Q_{X^n,Y^n}, K, \mathcal{C}, E = 0) + o(n) \tag{43}$$
$$\le H(\mathbf{H}_K X^n|K, \mathcal{C}, E = 0) + o(n) \tag{44}$$

where $o(n)$ in (42) corresponds to the term $1 + H(Q_{X^n,Y^n}, E|\mathcal{C}) + \Pr(E = 1)H(X^n|Q_{X^n,Y^n}, \mathcal{C}, E = 1)$, and (43) follows from the fact that K is a deterministic function of Q_{X^n,Y^n}.

We now provide an upper bound on the second term of (39).

$$\frac{1}{n} H(X^n|K, \mathcal{C}, E = 0) \le \frac{1}{n \Pr(E = 0)}\Big(H(X^n|K, \mathcal{C}, E)$$
$$- \Pr(E = 1) H(X^n|K, \mathcal{C}, E = 1)\Big)$$
$$\le \frac{1}{n} H(X^n|K, \mathcal{C}, E) + o(1) \tag{45}$$
$$\le H(X) + o(1). \tag{46}$$

By combining (39), (44) and (46), we obtain

$$\lim_{n\to\infty} R_2^{(n)} \le I(X;Y). \tag{47}$$

Conclusion. The rates in (37) and (47) are evaluated on average over the random code $\mathcal{C}$ with a parameter $\delta > 0$ arbitrarily small. This shows that there exists a sequence of $\Big(n, R_1^{(n)}, R_2^{(n)}\Big)$-zero-error source codes, such that

$$\Big(R_1^{(n)}, R_2^{(n)}\Big) \underset{n\to\infty}{\to} \big(H(X|Y), I(X;Y)\big). \tag{48}$$

References

[1] N. M. Bidgoli, T. Maugey, and A. Roumy, "Fine granularity access in interactive compression of 360-degree images based on rate-adaptive channel codes," *IEEE Transactions on Multimedia*, 2020.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[3] I. Csiszar, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Transactions on Information Theory*, vol. 28, no. 4, pp. 585–592, 1982.

[4] J. Chen, D.-K. He, and A. Jagmohan, "The equivalence between slepian-wolf coding and channel coding under density evolution," *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2534–2540, 2009.

[5] J. Chen, D.-K. He, A. Jagmohan, L. A. Lastras-Montaño, and E.-H. Yang, "On the linear codebook-level duality between slepian–wolf coding and channel coding," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5575–5590, 2009.

[6] L. Wang and Y.-H. Kim, "Linear code duality between channel coding and slepian-wolf coding," in *2015 53rd IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.

[7] A. H. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2031–2034, 1994.

[8] R. Timo, T. Chan, and A. Grant, "Rate distortion with side-information at many decoders," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5240–5257, 2011.

[9] C. Tian and S. N. Diggavi, "Side-information scalable source coding," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5591–5608, 2008.

[10] E. Akyol, U. Mitra, E. Tuncel, and K. Rose, "On scalable coding in the presence of decoder side information," in *2014 IEEE International Symposium on Information Theory (ISIT)*.

[11] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1636–1654, 2004.

[12] A. El Gamal and A. Orlitsky, "Interactive Data Compression," in *Annual Symposium on Foundations of Computer Science*, 1984, p. 9.

[13] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 99–111, Jan. 2003.

[14] E. Tuncel, "Kraft inequality and zero-error source coding with decoder side information," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4810–4816, 2007.

[15] E. Tuncel, J. Nayak, P. Koulgi, and K. Rose, "Zero-error distributed source coding," in *Distributed source coding: theory, algorithms, and applications*. Elsevier, 2009.

[16] R. Ma and S. Cheng, "Zero-error Slepian–Wolf coding of confined-correlated sources with deviation symmetry," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8195–8209, 2013.

[17] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[18] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

An Alon-Boppana theorem for powered graphs, generalized Ramanujan graphs and robust community detection

Emmanuel Abbe
EPFL
Email: emmanuel.abbe@epfl.ch

Peter Ralli
Weizmann Institute of Science
Email: peter.ralli@weizmann.ac.il

***Abstract*—Motivated by the problem of robust community detection, we study the r-th power of a graph, i.e., the new graph obtained by connecting every vertex pair in the original graph within distance r. This paper gives a generalization of the Alon-Boppana Theorem for the r-th power of graphs, including irregular graphs. This leads to a generalized notion of Ramanujan graphs, those for which the powered graph has a spectral gap matching the derived Alon-Boppana bound. In particular, we show that certain graphs that are not good expanders due to local irregularities, such as Erdős-Rényi random graphs, become almost Ramanujan once powered. A different generalization of Ramanujan graphs can also be obtained from the nonbacktracking operator. We next argue that the powering operator gives a more robust notion than the latter: Sparse Erdős-Rényi random graphs with an adversary modifying a subgraph of $\log(n)^\varepsilon$ vertices are still almost Ramanujan in the powered sense, but not in the nonbacktracking sense. As an application, this gives robust community testing for different block models.**

I. Introduction

The Alon-Boppana Theorem implies that a family of d-regular n-graphs with adjacency matrix A_n satisfies

$$\lambda_2(A_n) \geq 2\sqrt{d-1} - o_n(1). \tag{1}$$

A family of d-regular n-graphs with adjacency matrix A_n is *Ramanujan*, denoted here *A-Ramanujan*, if

$$\lambda_2(A_n) \leq 2\sqrt{d-1}. \tag{2}$$

Explicit constructions of such families were obtained in [8], [10], and it was shown in [5] that random d-regular graphs R_n are almost Ramanujan, in that

$$\lambda_2(A_n) = 2\sqrt{d-1} + o_{n,\mathbb{P}}(1), \tag{3}$$

where we use the notation $A_n = B_n + o_{n,\mathbb{P}}(1)$ when $A_n - B_n$ tends to 0 in probability when n tends to infinity. Obviously the above definitions are not directly relevant for irregular graphs. More specifically, Erdős-Rényi (ER) random graphs with an expected degree d will have their top two eigenvalues of order $\sqrt{\log(n)/\log\log(n)}$, due to eigenvectors localized on high-degree nodes, and therefore afford no spectral gap. Nonetheless, ER random graphs are similar to random d-regular graphs in various respects, e.g., their local neighborhoods for typical vertices are trees of either fixed or expected degree d. In particular, Lubotzky [9] gives a definition of Ramanujan that generalizes to irregular graphs, where G is Ramanujan if for every non-trivial eigenvalue λ of $A(G)$, $|\lambda| \leq \rho(\hat{G})$ where $\rho(\hat{G})$ is the spectral radius of the universal cover $\hat{G}$ of G. We observe that this definition does not fix the previously mentioned issue, as the spectral gap of G may be in some sense maximally large given the universal cover, but this property does not overcome the problem of the corresponding eigenvectors simply isolating on high-degree nodes. Thus one may wonder whether ER random graphs could also be good expanders, or even almost Ramanujan, if their local irregularity could be smoothed out. We will next discuss how to formalize and quantify such statements, and give motivating applications.

Generalized Ramanujan: beyond the adjacency operator. We start with a concrete example of a generalization of the Ramanujan property that can be obtained using the nonbacktracking operator of the graph. Given a graph G, the nonbacktracking matrix B_G is defined by the matrix on the set of directed edges of the graph (i.e., its dimension is twice the number of edges), and for two directed edges $e = (e_1, e_2)$, $f = (f_1, f_2)$, $B_{e,f} = 1_{e_2=f_1} 1_{e_1 \neq f_2}$. It was shown in [7] that for regular graphs, the Ramanujan property can be equivalently defined using the nonbacktracking spectral gap:

Definition I.1. A family of d-regular n-graphs with nonbacktracking matrix B_n is *B-Ramanujan*, if

$$|\lambda_2(B_G)| \leq \sqrt{d}, \tag{4}$$

where a graph satisfying (4) is also said to satisfy the *graph Riemann hypothesis [7],* since the eigenvalues of the nonbacktracking operator are the reciprocal of the poles of the Ihara zeta function of the graph [6], [7].

This definition is indeed equivalent to the former definition for regular graphs.

Lemma I.1. *[7] For regular graphs, A-Ramanujan is equivalent to B-Ramanujan.*

Further, it was shown in [4] that the B-Ramanujan definition extends more naturally to some irregular graphs than the A-Ramanujan definition, with the ER random graph being almost B-Ramanujan.

Theorem I.2. *[4] For a random ER graph,*

$$|\lambda_2(B_n)| = \sqrt{\lambda_1(B_n)} + o_{n,\mathbb{P}}(1), \tag{5}$$

with $\sqrt{\lambda_1(B_n)} = d + o_{n,\mathbb{P}}(1)$ *in this case.*

Therefore, the nonbacktracking operator meets our objective to turn ER random graphs into almost Ramanujan graphs *in the nonbacktracking domain*. We next argue that this approach can be further improved.

Symmetry and robustness. The B-Ramanujan definition suffers from two drawbacks: (1) Complex spectrum: as opposed to A, the matrix B is no longer symmetrical and thus has a complex spectrum. This makes some of the spectral intuition more delicate, where expansions are in terms of directed walks (that do not backtrack) and where the Courant-Fisher theorem (connecting cuts to eigenvalues) requires the use of oriented path symmetry. In particular, a tight Alon-Boppana theorem in the nonbacktracking domain is not obtained in [4]. (2) Robustness: the nonbacktracking operator meets the objective of making ER random graphs almost Ramanujan, but this property is lost once one slightly deviates from such models. For instance, perturbing the ER graph by adding a clique of size $c = \Omega(\sqrt{d})$ edges already makes the perturbed graph far from B-Ramanujan.

Some solutions have been proposed for these issues. First, the Bethe-Hessian operator [12] has been shown to essentially act as a symmetrized version of the nonbacktracking operator, however it does not fix the robustness issue. In [3], the generalized notion of r-nonbacktracking operator is used to gain generality in the proofs, but this is still nonsymmetrical and the complexity of the eigenpair computation increases significantly with r.

In [2], graph powering was proposed to address issues (1) and (2), testing robustness on a geometric block model, with parallel results in [13] using a related operator based on graph distances. However, these papers no longer investigate the connection to Ramanujan graphs, which is explicit in the case of the nonbacktracking operator [4] (cf. previous paragraphs).

This paper. In this paper we consider the symmetric operator of graph powering, and investigate its robustness and extremal spectral gap properties. The r-th graph power $G^{(r)}$ of a graph G modifies the graph by adding edges between any vertex pair at distance less or equal to r [2]. Equivalently, the adjacency matrix of $G^{(r)}$ is given by $A^{(r)} = \mathbb{1}((I+A)^r \geq 1)$, where the indicator function is applied point-wise to the matrix. We are typically interested in r large but significantly less than the graph diameter (otherwise powering turns the graph into a complete graph). In general, a regular graph may no longer be regular once powered, so even for regular graphs, we cannot bound the spectral gap for powered graph simply by using the Alon-Boppana result with degree d^r. Nonetheless, if we take a regular graph of girth larger than $2r$, then the r-th power is regular and the Alon-Boppana Theorem gives the bound $\lambda_2 \geq 2\sqrt{d^r - 1}$, so approximately $2d^{r/2}$ for large r or d. In fact, we shall see that random d-regular graphs have a second eigenvalue around $rd^{r/2}$ instead of $2d^{r/2}$ for large r [2]. For this reason one might conclude that in the powered domain, random d-regular graphs are not almost Ramanujan, creating contrast to the classical definition, and suggesting that powering may be misleading for generalizing Ramanujan graphs. The main result of this paper shows that this argument is false. Instead we will observe that applying the general Alon-Boppana bound to powered graphs is suboptimal, since powered graphs are not arbitrary graphs - instead, they are *powers* of arbitrary graphs.

We show an Alon-Boppana bound that applies to powers of (possibly irregular) graphs and which matches the scaling $rd^{r/2}$ for random d-regular graphs. Further, it is shown that both ER and random regular graphs have a comparably large and 'optimal' spectral gap in the powered domain, i.e., they are almost r-Ramanujan, just as they are almost B-Ramanujan. However, we show that this r-Ramanujan definition is more robust to local density variations (e.g., degrees) and adversaries than the B-Ramanujan definition: an adversary modifying a subgraph containing $\log(n)^{\varepsilon}$ vertices in ER or the random regular model cannot disrupt the P-Ramanujan property, while the B-Ramanujan property is lost after such a perturbation. We finish the introduction by motivating why such robust extensions are useful for spectral algorithms.

Community detection from the spectrum We will consider the problem of detecting the presence of a hidden structure in a graph, such as communities in the Stochastic Block Model (SBM). To give a formal statement of the problems we are interested in:

Definition I.2. Consider a graph G drawn from the symmetric 2-community SBM ensemble $SBM(n, a/n, b/n)$. An algorithm solves the problem of **weak recovery** if, with high probability in the choice of G, the algorithm identifies more that $\frac{1}{2} + \varepsilon$ of vertices to the correct community (up to a relabelling of the communities).

A related problem is to identify that there is community behavior in a random graph:

Definition I.3. Consider a graph G which is chosen with probability $\frac{1}{2}$ according to the random graph model $ER(n, d/n)$ and with 12 from $SBM(n, a/n, b/n)$. An algorithm solves the problem of **distinguishability** if, with high probability in the choice of G, the algorithm correctly identifies which of the two ensembles G was chosen from.

This means that we want to distinguish between two cases, either the graph is drawn from ER$(n, d/n)$, or, on the other hand, it is the assembly of two independent ER$(n/2, a/n)$ subgraphs with a random bipartite graph connecting each pair of vertices across the groups independently with probability b/n. One can view the SBM adjacency matrix as an ER matrix A perturbed as $A+Z$ where Z adds/subtracts edges within/across clusters with the specified probabilities. If $(a+b)/2 \neq d$, the average degree or edge density of the graph allows to distinguish the two models, so we consider the case where $(a+b)/2 = d$. As we are concerned with a regime where the second eigenvector does not give information about the communities in the SBM case, similarly spectral analysis will fail to distinguish the SBM from the ER random graph.

Rather than using the spectrum of A, we may attempt to use the cycle counts, as originally proposed in [11]. This allows us to distinguish the models down to the optimal Kesten-Stigum (KS) threshold, i.e., $\lambda_2(SBM) > \sqrt{\lambda_1(SBM)}$, which reads $(a-b)/2 > \sqrt{(a+b)/2}$. One can also use spectral methods, not based on the adjacency matrix but on the nonbacktracking matrix [4], which does not suffer from ER irregularities due to the weak Ramanujan property: its second eigenvalue is $\sqrt{(a-b)/2} + o_{n,\mathbb{P}}(1)$ in the ER case, and $(a-b)/2 + o_{n,\mathbb{P}}(1)$ in the SBM case due to the community eigenvector. Thus the second eigenvalue of the nonbacktracking matrix allows us to solve the distinguishability problem down to the optimal KS threshold.

The relevance of the almost-Ramanujan property in the B-domain is now clear: the idea of using the spectral method to analyze the SBM is that v_2 will approximate the community vector while $v_{\geq 3}$ will be noise from randomness. For this analysis to work, the community signal λ_2 needs to be separated from the noisy values λ_3,; if $\lambda_2 \approx \lambda_3$ that means that λ_2 (and v_2) might be controlled by the noise rather than the community vector. And so a large spectral gap for the null model (ER) leaves more room for the community signal to be visible in the SBM, and thus gives a broader range of parameters for which testing is solvable.

We want to consider the question of whether algorithms for solving distinguishability are robust to an adversarial modification of the graph:

Definition I.4. Suppose a random graph G is chosen according to the distinguishability problem; i.e., from $SBM(n, a/n, b/n)$ with probability $1/2$ and from $ER(n, d/n)$ with $1/2$. Then we allow an adversary to create G' by modifying a subgraph of c vertices (adding and removing any number of edges) . The **robust testing problem** (for distinguishability) is solved by an algorithm that takes input G' (i.e., without seeing G) and can, with high probability in the choice of G, decide whether G' is the result of adversarial modification of a graph from SBM or ER.

It is not hard to check that a budget of $c = \Omega(a+b)$ suffices to disrupt the two previous methods based on cycle counts and the nonbacktracking operator. However, for graph powering and for $c = o\left(\frac{((a-b)/2)^r}{\log(n)\sqrt{(a+b)/2}^{\,r-1}}\right)$, we will prove that that the ER model perturbed by such an adversary affords still a maximal spectral gap in the powered domain. This allows one to distinguish the models down to the KS threshold despite such adversaries — See Corollary II.7. For this case, a similar result has recently been obtained in parallel work [13] for the SBM using a slightly different operator based on the distance matrix of the graph. [13] further covers the case of weak recovery.

II. Results

Remark. The proofs of all theorems in this section are found in the long version of this paper: https://arxiv.org/pdf/2006.11248.pdf

A. Notations

We will start by recalling some standard notations. In a graph G, $\text{dist}_G(v, w)$ is the graph distance metric, measuring the length (in edges) of the shortest $v - w$ walk in G. If G is a finite connected graph, $\text{diam}(G)$ is the maximum graph distance between any pair of vertices. If G has $|V| = n$ the adjacency matrix $A(G)$ is an $n \times n$ matrix indexed by V in which $A_{ij} = 1$ if $i \sim_G j$ and 0 otherwise. The eigenvalues of A are $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$.

Let G be a graph, G may have self-loops but we do not allow repeated edges. If $r \geq 1$, the r-th power of G is $G^{(r)}$, the graph with vertex set $V(G)$ and an edge $\{x, y\}$ iff $\text{dist}_G(x, y) \leq r$. This definition was introduced in [2]. $A^{(r)}$ is the adjacency operator of $G^{(r)}$. The graph power $G^{(r)}$ will by definition contain a self-loop at every vertex, it does not contain repeated edges.

In order to model community behavior, we sample random graphs from the (balanced 2-community) Stochastic Block Model: a graph G sampled from $SBM(n, a/n, b/n)$ is a random graph on n vertices generated by the following two steps. First, each vertex is put in community X_1 or X_2 uniformly and independently. Second, for every pair of vertices v, w, $\{v, w\}$ is taken to be an edge with probability a/n if v and w are in the same community and b/n if not. The SBM is a generalization of the well-known Erdős-Rényi graph $ER(n, d/n) := SBM(n, d/n, d/n)$.

In an SBM, the community vector is the $\{\pm 1\}$-vector on the vertices with $v_i = 1$ if $i \in X_1$ and $v_i = -1$ if $i \in X_2$.

The *Kesten-Stigum (KS) threshhold* $\frac{a-b}{2} = \sqrt{\frac{a+b}{2}}$ is a limit on weak recovery in the SBM: when $\frac{a-b}{2} \leq \sqrt{\frac{a+b}{2}}$ weak recovery is not possible [11].

B. Alon-Boppana for powered graphs

We investigate the maximum size of the spectral gap following graph powering. We modify well-known methods of finding a lower bound on the second-largest eigenvalue in a graph in order to derive a version of the Alon-Boppana theorem for graph powering.

Recall that the Alon-Boppana result for (non-powered) d-regular graphs is

$$\lambda_2(A) \geq (1 - o_{\text{diam}(G)}(1)) 2\sqrt{d-1}.$$

A Ramanujan graph is one for which the lower bound is tight, as first investigated in [8].

Friedman [5] argued that d-regular random graphs are almost Ramanujan with high probability. We will replicate this result under powering, arguing that with high probability, a d-regular random graph under powering is almost r-Ramanujan.

The Alon-Boppana like bound for powered graphs is:

Theorem II.1. *Let $\{G_n\}_{n\geq 1}$ be a sequence of graphs such that diam$(G_n) = \omega(1)$, and $\{r_n\}_{n\geq 1}$ a sequence of positive integers such that $r_n = \varepsilon \cdot$ diam(G_n). Then,*

$$\lambda_2(G_n^{(r_n)}) \geq (1-o_\varepsilon(1))(r_n+1)\hat{d}_{r_n}^{r_n/2}(G_n), \tag{6}$$

where

$$\hat{d_r}(G) = \left(\frac{1}{r+1}\sum_{i=0}^{r}\sqrt{\delta^{(i)}(G)\delta^{(r-i)}(G)}\right)^{2/r}, \tag{7}$$

$$\delta^{(i)}(G) = \min_{(x,y)\in E(G)} |\{v : d_G(x,v) = i, d_G(y,v) \geq i\}|. \tag{8}$$

In [2], the authors (jointly with Boix and Sandon) prove that the quantity $\hat{d}$ is $d \pm o(d)$ with high probability in a d-regular random graph, for such graphs we get the following version of Theorem II.1

Theorem II.2. *Let G be a random d-regular graph and $r = \varepsilon\log(n)$, where $\varepsilon\log d < 1/4$. Then, with high probability,*

$$\lambda_2(G^{(r)}) \geq (1-o(1))(r+1)\sqrt{d}^{\,r}. \tag{9}$$

We say that a graph G is r-Ramanujan if the bound of Theorem II.1 is tight, that is, if $\lambda_2(G^{(r)}) \leq (1+o(1))(r+1)\hat{d}^{r/2}$. We demonstrate that a class of Ramanujan graphs are also r-Ramanujan, and therefore that r-Ramanujan graphs exist.

Lemma II.3. *Let G be a d-regular Ramanujan graph with girth g, and let $2r < \mathrm{girth}(G)$. Then $G^{(r)}$ is r-Ramanujan, so that $\lambda_2(G^{(r)}) = (1+o_d(1))\,(r+1)d^{r/2}$.*

As first seen in the original construction of [8], there are families of Ramanujan graphs may have girth $\Theta(\log(n)/\log(d)$; it is straightforward to observe that this is an upper bound on girth. Using such graphs, we can now construct r-Ramanujan powered graphs.

C. Robustness of graph powering

If a graph contains large cliques but otherwise appears to be randomly generated (such as by the stochastic block model), analysis of the leading eigenvectors will reliably identify those cliques rather than any communities that may exist. We will investigate what happens to such graphs under powering. Observe that a shortest path of length r may have, at most, 1 edge from any clique. Intuitively, powering the graph means gives a graph whose edges are paths in G that take edges mostly from the "expander part" of the graph with at most one edge from any large clique, and so we can observe the community behaviour. In fact the results will be more general; we will consider any perturbation of an expander graph and not just the addition of a clique.

In order to examine specific examples of random graphs, we will first briefly give a general discussion on how adding edges to a graph impacts the eigenvalues of $G^{(r)}$.

Definition II.1. Let G be a graph on vertex set $[n]$ and let H be a graph whose vertex set is a subset of $[n]$. Then $G+H$ is the graph with vertex set $[n]$ and satisfying the equation $E(G+H)\Delta E(G) = E(H)$, in other words, $G+H$ is obtained from G by adding or removing all edges of H as applicable.

We will use the following simple theorem for bounding eigenvalues of graphs of the form $G+H$.

Theorem II.4. *Let $k \geq 1$. Then $|\lambda_k(A_{G+H}) - \lambda_k(A_G)| \leq \|A_H\|$.*

An application of Theorem II.4 is the following result, which we will use to prove the main theorems of this section.

Theorem II.5. *Let G be a graph with $c < |V(G)|$, and let H be a graph whose vertex set consists of at most c elements of $V(G)$. Define $D^{(i)}(G+H)$ to be the maximum degree in $G^{(i)}$ over all the vertices of H. Then*

$$\left|\lambda_k((G+H)^{(r)}) - \lambda_k(G^{(r)})\right| \leq \sum_{q=0}^{r-1} c\max_i \sqrt{D^{(i)}D^{(q-i)}}.$$

We apply Theorem II.5 to the case of an Erdös-Rényi random graph or a sparse SBM.

Theorem II.6. *Let $G = SBM(n, a/n, b/n) + H$ where $|V(H)| \leq c$. There is a universal constant α so that, with high probability in the random choice of graph from the SBM ensemble, the following statements hold:*

1) *If $\sqrt{(a+b)/2} \leq (a-b)/2$ (the KS threshold), then, independently of the choice of H,*

$$(1-o(1))\left(\tfrac{a-b}{2}\right)^r - c\log(n)^\alpha\sqrt{\tfrac{a+b}{2}}^{\,r-1} \leq \lambda_2(G^{(r)})$$
$$\leq c\log(n)^\alpha\sqrt{\tfrac{a+b}{2}}^{\,r-1} + (1+o(1))\left(\tfrac{a-b}{2}\right)^r.$$

2) *On the other hand, if $\sqrt{(a+b)/2} \leq (a-b)/2$, then independently of the choice of H,*

$$\left(\sqrt{\tfrac{a+b}{2}} - c\right)\log(n)^\alpha\sqrt{\tfrac{a+b}{2}}^{\,r-1} \leq \lambda_2(G^{(r)})$$
$$\leq \left(c + \sqrt{\tfrac{a+b}{2}}\right)\log(n)^\alpha\sqrt{\tfrac{a+b}{2}}^{\,r-1}.$$

In particular, if $G = ER(n, d/n) + H$ (i.e., the SBM with values $a = b = d$), then

$$\left(\sqrt{d} - c\right)\log(n)^\alpha\sqrt{d}^{\,r-1} \leq \lambda_2(G^{(r)})$$
$$\leq \left(c + \sqrt{d}\right)\log(n)^\alpha\sqrt{d}^{\,r-1}.$$

Remark. Note that we are thinking of $r = \log(n)^\gamma$ where γ is a constant. If $c = \log(n)^\varepsilon$, then in this result, we obtain the bound $\lambda_2(ER(n, d/n)^{(r)}) \leq \log(n)^\alpha d^{r/2}$ for the original ER graph and $\lambda_2(G^{(r)}) \leq \log(n)^{\alpha+\varepsilon}d^{r/2}$ for the perturbed graph. Our Alon-Boppana result for powering states that (if the unpowered graph is d-regular) then $\lambda_2 \geq (r+1)d^{r/2} = \log(n)^\gamma d^{r/2}$. Because the upper bounds for $\lambda_2(ER(n, d/n)^{(r)})$ and $\lambda_2(G^{(r)})$ are tight up to a power of $\log(n)$ we say that those graphs are both almost Ramanujan.

We will use the following result to solve the distinguishability problem. That is, suppose with equal probability either the $ER(n, \frac{a+b}{2n})$ or $SBM(n, a/n, b/n)$ random graph model is chosen, and then a graph G is drawn from that model at random and perturbed to $G + H$ by an adversarial choice of H with the constraint $|V(H)| \leq c$. Then for what values of a, b, c is it possible to guess with high probability which of the two models G comes from?

Corollary II.7. *Assume* $c = o\left(\frac{((a-b)/2)^r}{\log(n)\sqrt{(a+b)/2}^{\,r-1}}\right)$.

Let $G = SBM(n, a/n, b/n) + H$. *With high probability, independently of the choice of* H,

$$\lambda_2(G^{(r)}) = (1 \pm o(1)) \left(\tfrac{a-b}{2}\right)^r .$$

Let $G = ER(n, \frac{a+b}{2}/n) + H$. *With high probability, independently of the choice of* H,

$$\lambda_2(G^{(r)}) = o(\left(\tfrac{a-b}{2}\right)^r).$$

The proof of each statement is just an application of Theorem II.6.

Remark. A common method of solving the distinguishability problem is by examining the number of m-cycles [1]. In brief, the number of cycles is $\frac{1}{2m}\left(d^m \pm d^{m/2}\right)$ in an ER graph and $\frac{1}{2m}\left(d^m + (\frac{a-b}{2})^m \pm d^{m/2}\right)$ in an SBM, so that the decision is possible up to the KS threshhold. However this method is not robust to adversarial perturbation of the graph. If the perturbation is a c-regular clique where $c >> d$, the number of m-cycles is $\frac{1}{2m}\left(c^m + d^2c^{m-2} \pm \sqrt{d^2c^{m-2}}\right)$ for both the ER and SBM random graph models, this makes the decision impossible. But the method of graph powering lets us solve this decision problem even with the addition of much larger cliques. This result is similar to one found in the work of Stephan and Massoulié [13], working with the distance matrix rather than $A^{(r)}$.

Remark. Implicit in the result of Theorem II.6 is that if $G = SBM(n, a/n, b/n) + H$ under the hypotheses of Corrolary II.7, then the second eigenvector v_2 of $G^{(r)}$ will approximate the second eigenvector of $SBM(n, a/n, b/n)^{(r)}$. Theorem 2.6 of [2] tells us that the second eigenvector of $SBM(n, a/n, b/n)^{(r)}$ is useful for weak recovery of the communities.

III. Open problems

- In Lemma II.3 we show that a Ramanujan graph with girth more than $2r$ must be also r-Ramanujan after powering, taking advantage of the fact that all r-neighborhoods in that graph are trees. Is this true in general - is every Ramanujan graph also r-Ramanujan (with some reasonable bound on r)?
- The converse of the previous problem - is every r-Ramanujan powered graph necessarily the r-th power of a Ramanujan graph?
- Observe that we, along with our concurrent work [2], do not in general investigate the exponent in the factors $\log(n)^\alpha$ which appears in our paper. In particular in the Alon-Boppana result for powering, we see the bound $(r + 1)d^{r/2}$ where $r = \varepsilon \log(n)$, but in the bound for $\lambda_2(ER(n, d/n)^{(r)})$ we have $\log(n)^{alpha}d^r$. Because these bounds are equivalent up to a factor of a power of $\log(n)$ we say that $ER(n, d/n)^{(r)}$ is r-Ramanujan. Is it possible to better characterize the exponents of $\log(n)$ that appear in this work (especially related to the ER graph) and to understand why they exist in view of the Alon-Boppana result for powering?

References

[1] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. *ArXiv e-prints 1512.09080*, December 2015.

[2] Emmanuel Abbe, Enric Boix-Adserà, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *SIAM J. Math. Data Sci.*, 2(1):132–157, 2020.

[3] Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 71(7):1334–1406, 2017.

[4] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 1347–1357, Washington, DC, USA, 2015. IEEE Computer Society.

[5] J. Friedman. A proof of Alon's second eigenvalue conjecture. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC '03, pages 720–724, New York, NY, USA, 2003. ACM.

[6] K.-I. Hashimoto. Zeta functions of finite graphs and representations of p-adic groups. *In Automorphic forms and geometry of arithmetic varieties. Adv. Stud. Pure Math.*, 15:211–280, 1989.

[7] Matthew D. P. Horton, Harold M. Stark, and Audrey A. Terras. What are zeta functions of graphs and what are they good for? In *Contemporary Mathematics, Quantum Graphs and Their Applications*, 2006.

[8] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[9] Alexander Lubotzky. Cayley graphs: eigenvalues, expanders and random walks. *Surveys in combinatorics, 1995 (Stirling)*, 218:155–189, 1995.

[10] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families i: Bipartite Ramanujan graphs of all degrees. *Annals of Mathematics*, 182(1):307–325, 2015.

[11] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

[12] A. Saade, F. Krzakala, and L. Zdeborová. Spectral Clustering of Graphs with the Bethe Hessian. *Advances in Neural Information Processing Systems*, 27.

[13] Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. *Proceedings of Machine Learning Research*, vol. 99:1–30, 2019.

A CSI Compression Scheme Using Context Trees

Henrique K. Miyamoto
University of Campinas (Unicamp)
Institute of Mathematics (IMECC)
Campinas, SP, Brazil
Email: hmiyamoto@ime.unicamp.br

Sheng Yang
CentraleSupélec, Paris-Saclay University
Laboratory of Signals and Systems (L2S)
Gif-sur-Yvette, France
Email: sheng.yang@centralesupelec.fr

***Abstract*—We propose novel compression algorithms for time-varying channel state information (CSI) in wireless communications. The proposed schemes combine (lossy) vector quantisation and (lossless) compression. The vector quantisation technique is based on data-adapted parametrised companders applied on each component of the normalised vector. Then, the sequences of quantisation indices are compressed according to estimated distributions computed with a context-tree approach. The algorithms have low complexity, are linear-time in spatial dimension and time duration, and can be implemented in an online fashion. We run numerical experiments to demonstrate the effectiveness of the proposed algorithms in such scenarios.**

I. Introduction

In wireless communication systems, efficiently representing the channel state information (CSI) is crucial for storage and dissemination. Typically, in the downlink transmission from a base station (BS) with multiple antennas to multiple users, beamforming techniques rely on precise CSI at the transmitter side [1]. For the BS to acquire the CSI, however, it usually requires that each user feeds back the CSI measurements in a timely and accurate fashion. How to reduce the bandwidth cost of such feedback traffic, which is highly non-negligible, is becoming a crucial problem. This is essentially a lossy data compression problem.

The spatial correlation inherent to the antenna structures has been exploited to reduce CSI dimension in recent works using deep learning and compressed sensing techniques (see, e.g., [2], [3] and the references therein), while the temporal correlation of CSI measurements is less exploited for feedback. Indeed, if the sequence of the quantised symbols is stationary, it can be losslessly compressed up to the entropy rate of the underlying process. A possible approach for CSI compression is therefore to apply any universal compression algorithm [4], [5], such as Lempel-Ziv [6], [7] (known as LZ77 and LZ78), to the quantisation indices.

Another universal compressor is the *context-tree weighting* (CTW) algorithm [8], which learns the distribution of a given sequence in an efficient way. This distribution can then be used to compress the sequence in combination with arithmetic coding, achieving the Rissanen lower bound [8]. A modification of CTW yields the *context-tree maximising* (CTM) algorithm [9], which can produce the maximum *a posteriori* (MAP) probability tree model. Connections with Bayesian inference have been explored in [4], [10].

However, directly applying these algorithms to compress quantisation indices in an online fashion presents some difficulties. First, the output bit-stream is of variable length, making the feedback difficult to implement. Second, in Lempel-Ziv methods, the input symbol block is also of variable length, as it depends on parsing the original sequence. Finally, arithmetic coding has to be carefully implemented so as to deal with digital computers finite precision constraints [11]. Trying to avoid such difficulties motivates us to propose new compression algorithms adapted to applications such as the communication scenarios considered here.

In this work, we focus on the problem of online lossy compression of a sequence of CSI vectors and propose a two-step compression procedure. First, a new vector quantisation technique, based on a class of parametrised companders, is applied on the components of the normalised vector. The quantisation is composed of a non-linear transformation, followed by a uniform quantiser. The companders can be designed and updated with available empirical data. In particular, we consider the widely used μ-law compander and a new one, the β-law compander, inspired by the beta distribution. Then, we compress the sequence of quantisation indices using a context-tree-based approach. We propose two solutions: 1) to directly apply CTW with arithmetic coding, or 2) to apply CTM to estimate the conditional distribution of the upcoming symbol at each time instant and use this probability to compress the symbol. In the latter case, we encode each symbol with a fixed number of levels to limit the fluctuation of the encoded bit-flow—a desirable property in communication systems. In addition, the algorithms have low complexity, are linear-time in both the spatial dimension and time duration, and can be implemented in an online fashion.

This paper is organised as follows. In Section II we present the system model and review basic concepts of vector quantisation and context-tree representation. Our CSI compression algorithm is described in Section III. The simulation of CSI acquisition is analysed in Section IV, followed by some conclusions. Due to the space limitation, we omit some important details that can be found in the extended version in [12], where implementation codes are also available.

Notation: Vectors ($\boldsymbol{v}$) are denoted by bold italic lower-case letters. Random variables (X) are in non-italic upper-case. L_2 vector norms are denoted by $\|\boldsymbol{v}\|$. Logarithms are to the base 2. We denote $[n] := \{1, \dots, n\}$.

II. Problem Formulation and Preliminaries

A. Main Problem

We consider a network composed of a transmitter (e.g., base station) and N_r receivers (e.g., mobile users). Assume that the CSI between the transmitter and receiver k at time t can be described by a complex vector $\boldsymbol{h}_k[t] \in \mathbb{C}^{N_\text{t} \times 1}$, for $k \in [N_\text{r}]$. For different purposes (e.g., feedback, storage), each receiver is required to represent its state sequence using as few bits as possible, for a given distortion constraint. This is known as the lossy source coding problem [5]. In most practical scenarios, the norm of the vectors $\boldsymbol{h}_k[t]$ is less important than the direction. Therefore, our goal here is to compress the normalised vector $\boldsymbol{h}_k[t]/\|\boldsymbol{h}_k[t]\|$.

B. Vector Quantisation

A *vector quantiser* [13] of dimension p and size M, is a mapping $q : \mathbb{R}^p \to \mathcal{C}$, with $\mathcal{C} := \{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{M-1}\} \subset \mathbb{R}^p$, that associates each vector $\boldsymbol{x} \in \mathbb{R}^p$ to a codeword $\hat{\boldsymbol{x}} := q(\boldsymbol{x}) = \boldsymbol{y}_k$, for some $k \in \{0, 1, \ldots, M-1\}$. For a sequence of vector symbols $\boldsymbol{x}_1^n := \boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_n$, we can apply vector-by-vector quantisation, generating a sequence of quantised vectors $\hat{\boldsymbol{x}}_1^n := \hat{\boldsymbol{x}}_1 \hat{\boldsymbol{x}}_2 \cdots \hat{\boldsymbol{x}}_n$ and a sequence of quantisation indices $k_1^n := k_1 k_2 \cdots k_n$, where $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_{k_i}$, for each $i \in [n]$.

Two important parameters to assess the performance of a vector quantiser are the quantisation rate and the mean distortion. The *quantisation rate*, defined as $R := (\log M)/p$, is an indicator of the cost to describe the vector, while the *mean distortion* measures the error induced by the quantisation. We use, as distortion measure between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, the mean squared chordal distance (MSCD), defined as

$$\text{MSCD}(\boldsymbol{x}, \hat{\boldsymbol{x}}) := 1 - \mathbb{E}\left[\frac{|\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle|^2}{\|\boldsymbol{x}\|^2 \|\hat{\boldsymbol{x}}\|^2}\right]. \qquad (1)$$

C. Variable-Order Markov Chain and Context Tree

Let $x_i^j := x_i x_{i+1} \cdots x_j$ be a scalar sequence over an alphabet $\mathcal{A} := \{0, 1, \ldots, m-1\}$, generated by a source with probability distribution P. We denote $l(x_i^j) := j - i + 1$ the length of sequence x_i^j. A *variable-order Markov chain* with order or memory D (also called *bounded memory tree source*) is a random process for which $P(x_i | x_{-\infty}^{i-1}) = P(x_i | x_{i-D}^{i-1})$. Our interest in Markov chains comes from the fact that any stationary ergodic source can be approximated by a Markov chain with sufficiently large order [4], [5].

The statistical behaviour of a variable-order Markov chain is described by a *context set* $\mathcal{S}$ (also known as *suffix set* or *model*), which is defined as a subset of $\bigcup_{i=0}^D \mathcal{A}^i$ that is proper (i.e., no element in $\mathcal{S}$ is a proper suffix of any other) and complete (i.e., each $x_{-\infty}^n$ has a suffix in $\mathcal{S}$, which is unique by properness). The *context function* $c : \mathcal{A}^D \to \mathcal{S}$ maps each length-D context x_{i-D}^{i-1} to a suffix $c(x_{-\infty}^{i-1}) = c(x_{i-D}^{i-1}) = x_{i-j}^{i-1}$, $j \leq D$. Furthermore, each suffix $s \in \mathcal{S}$ is associated with a parameter $\boldsymbol{\theta}_s := (\theta_s(0), \theta_s(1), \ldots, \theta_s(m-1))$, where $\theta_s(j) := P(j|s)$. The *parameter vector* $\Theta := \{\boldsymbol{\theta}_s : s \in \mathcal{S}\}$ groups all parameters in the context set $\mathcal{S}$. Therefore, the Markov chain is completely characterised by the couple $(\mathcal{S}, \Theta)$. We use $\mathcal{C}_D$ to denote the class of all context sets of order up to D. Finally, we define $L_D(\mathcal{S}) := |\{s \in \mathcal{S} : \; l(s) = D\}|$ the number of contexts with length D.

Since the context set $\mathcal{S}$ is proper, its elements can be represented as leaf nodes of a tree $\mathcal{T}_D$, called *context tree*, i.e., $\mathcal{S} \subseteq \mathcal{T}_D$. For a given sequence x_1^n, each leaf node $s \in \mathcal{S}$ is associated with a *counter* $\boldsymbol{a}_s := \boldsymbol{a}_s(x_1^n) := (a_s(0), a_s(1), \ldots, a_s(m-1))$, where $a_s(j)$ stores the number of times that symbol $j \in \mathcal{A}$ follows context s in x_1^n. The counter of each inner node of the tree is recursively defined as the sum of the counters of its children nodes, i.e., $\boldsymbol{a}_s := \sum_{j \in \mathcal{A}} \boldsymbol{a}_{js}$, $\forall \, s \in \mathcal{T}_D \setminus \mathcal{S}$. In particular, we use the empty string λ to denote the root of the tree.

With the above definitions and the Markov property for a D-th order Markov chain, if both $\mathcal{S}$ and Θ are known, the probability of a sequence can be written as [10]

$$P(x_1^n | x_{D-1}^0, \mathcal{S}, \Theta) = \prod_{s \in \mathcal{S}} \prod_{j \in \mathcal{A}} \theta_s(j)^{a_s(j)}. \qquad (2)$$

If only the model $\mathcal{S}$ is known, but not its parameters Θ, the *marginal distribution* of a sequence x_1^n, given its past x_{1-D}^0 and model $\mathcal{S}$, is

$$P(x_1^n | x_{1-D}^0, \mathcal{S}) = \int P(x_1^n | x_{1-D}^0, \mathcal{S}, \Theta) \pi(\Theta | \mathcal{S}) \, \mathrm{d}\Theta, \qquad (3)$$

assuming the distribution of the parameters, $\pi(\Theta|\mathcal{S})$, is known. While this distribution is unknown in general, using the so-called *Jeffrey's prior* is asymptotically optimal in the minimax sense [4]. This choice corresponds to setting $\pi(\Theta|\mathcal{S})$ to be the Dirichlet distribution with parameters $\left(\frac{1}{2}, \cdots, \frac{1}{2}\right)$. In this case, the distribution (3) can be simplified to the so-called *Krichevsky–Trofimov (KT) distribution*, which can be easily computed as

$$P(x_1^n | x_{1-D}^0, \mathcal{S}) = \prod_{s \in \mathcal{S}} P_e(\boldsymbol{a}_s), \qquad (4)$$

where

$$P_e(\boldsymbol{a}_s) = \frac{\prod_{j=0}^{m-1} \left(\frac{1}{2}\right) \left(\frac{3}{2}\right) \cdots \left(a_s(j) - \frac{1}{2}\right)}{\left(\frac{m}{2}\right) \left(\frac{m}{2} + 1\right) \cdots \left(\frac{m}{2} + M_s - 1\right)}, \quad s \in \mathcal{T}_D, \quad (5)$$

with $M_s := \sum_{j=0}^{m-1} a_s(j)$.

Finally, if the model $\mathcal{S}$ is also unknown, then we shall marginalise over $\mathcal{S}$ with a given prior distribution π_D on all models $\mathcal{S}$ of maximal depth D. Fixing $\gamma \in \,]0, 1[$ and

$$\pi_D(\mathcal{S}) := (1 - \gamma)^{\frac{|\mathcal{S}| - 1}{m - 1}} \gamma^{|\mathcal{S}| - L_D(\mathcal{S})}, \qquad (6)$$

we obtain a mixture of different distributions (4), corresponding to the coding distribution of CTW [4], [10]:

$$Q_n(x_1^n | x_{1-D}^0) := \sum_{\mathcal{S} \in \mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s \in \mathcal{S}} P_e(\boldsymbol{a}_s). \qquad (7)$$

Not only is this coding distribution universal for the class of stationary ergodic sources (i.e., it asymptotically achieves optimal coding rate irrespective of the source distribution), but also it can be recursively computed so that the complexity is linear in n [8].

The CTM algorithm [9] comes from a modification of the CTW algorithm and can be used to compute the maximum *a posteriori* model $\mathcal{S}$ for a given sequence x_{1-D}^n.

Definition 1. For $\gamma \in]0,1[$, the *maximised probability* P_m^s of each node $s \in \mathcal{T}_D$ with length $d = l(s)$ is

$$P_m^s := \begin{cases} \max\{\gamma P_e(\boldsymbol{a}_s), (1-\gamma)\prod_{j=0}^{m-1} P_m^{js}\}, & 0 \le d < D, \\ P_e(\boldsymbol{a}_s), & d = D, \end{cases} \tag{8}$$

and the *maximising tree* $\mathcal{S}_m^s$ is obtained by pruning the descendants of the nodes s where the maximum is achieved by the first term.

Lemma 1 (See [10]). *The maximised coding distribution* P_m^λ *of the root node* $\lambda \in \mathcal{T}_D$ *satisfies*

$$P_m^\lambda = \max_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \tag{9}$$

We find then that the maximising tree $\mathcal{S}_m^\lambda$, which is associated to the maximised probability P_m^λ, corresponds to the maximum *a posteriori* model:

$$\mathcal{S}_m^\lambda = \arg\max_{\mathcal{S}\in\mathcal{C}_D} P(\mathcal{S}|x) = \arg\max_{\mathcal{S}\in\mathcal{C}_D} \pi_D(\mathcal{S}) \prod_{s\in\mathcal{S}} P_e(\boldsymbol{a}_s). \tag{10}$$

III. Proposed Scheme

A. Quantisation

The vector quantisation that we propose consists in vector normalisation, decomposition into real components, and individual scalar quantisation based on parametric companders.

1) Vector Normalisation: In this step, the input vector $\boldsymbol{x} = [x(1) \cdots x(N_\mathrm{t})]$ is normalised by the component with the largest absolute value, i.e., $\bar{\boldsymbol{x}} := \boldsymbol{x}/x(i^*)$ where $i^* := \arg\max_i |x(i)|$. Note that $\bar{x}(i^*) = 1$, while the other normalised components are complex in general, with absolute value in $[0,1]$. The i^*-th component can skip the following steps and be directly assigned a special quantisation index indicating it as the strongest component.

2) Decomposition: Before quantisation, each complex component should be decomposed into real values. We consider the polar decomposition into amplitude and phase, since these components are usually less correlated in wireless applications, thus providing a less 'redundant' representation.

3) Quantisation with Parametric Companders: The amplitude and phase are quantised separately with different scalar quantisers of M_abs and M_ang quantisation levels, respectively.

If the input is uniformly distributed, then a uniform quantiser is optimal. In general, however, uniform quantisation can be far from optimal in the rate-distortion sense [5]. Let X be a random variable representing the input, following some distribution P over the support interval $[0,1]$. The idea of using a compander is to apply a non-linear and non-decreasing mapping $g : [0,1] \to [0,1]$ to the signal (*compression*) before quantising it, so that the signal is more 'uniform' in the image space. To recover the signal, the inverse mapping $g^{-1} : [0,1] \to [0,1]$ is used (*expansion*). It is practical to use parametric companders, i.e., (differentiable) maps g that can be described by a few number of parameters. One of the widely used such companders is the μ-law compander, which is parametrised by a value $\mu > 0$. Note that, as compared to the Lloyd quantiser [5], compander-based quantisers have much lower complexity of quantisation and representation.

TABLE I
Two Compander Functions.

Compander	Parameters	pdf $g'(x)$
μ-law	$\mu > 0$	$\dfrac{\mu}{(1+\mu x)\ln(1+\mu)}$
β-law	$\alpha > 0,\ \beta > 0$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

In this work, we propose a data-driven design of a compander parametrised by some θ (which can contain multiple scalar parameters). Assume that we have a set of training data $x_1, \ldots, x_n$. Our design is a two-step procedure: 1) uniformisation of the data, and 2) adjustment of the compander parameter.

We assume that the training data are formed by independent samples from some distribution P. If we knew the cumulative distribution function (cdf) F_P of P, we could apply the mapping F_P so that $F_P(x_1), \ldots, F_P(x_n)$ are samples from a uniform distribution. If, however, we are restricted to a class of companders $\{g_\theta,\ \theta \in \mathcal{Q}\}$ for some set $\mathcal{Q}$, then we have to approximate F_P with g_θ. Since a compander as defined above is non-decreasing from 0 to 1, it is equivalent to a cdf. Thus, a sensible criterion for the approximation is through the Kullback-Leibler divergence:

$$\theta^* = \arg\min_{\theta\in\mathcal{Q}} D(P \,\|\, g_\theta) = \arg\max_{\theta\in\mathcal{Q}} \mathbb{E}_P[\log(g_\theta'(\mathrm{X}))]. \tag{11}$$

Remarkably, this is equivalent to maximising the differential entropy of $g_\theta(\mathrm{X})$. Since the uniform distribution maximises differential entropy among all bounded support distributions [5], the criterion (11) returns indeed the best 'uniformiser'. Note that, since g_θ is a cdf, g_θ' is the corresponding probability density function (pdf).

The true distribution of the data is, nevertheless, unknown in most practical scenarios. But we can adapt the probabilistic criterion (11) into a data-driven one by replacing the expectation with the sample mean:

$$\arg\max_{\theta\in\mathcal{Q}} \frac{1}{n} \sum_{i=1}^n \log\left(g_\theta'(x_i)\right). \tag{12}$$

In this paper, we consider the μ-law compander and another one that we call β-law compander, as shown in Table I. The β-law compander is equivalent to the beta cdf, parametrised by $\alpha > 0$ and $\beta > 0$. An attractive feature of the β-law compander is that its pdf is log-concave in (α, β) [14, Theorem 6], so that the maximisation (12) can be easily solved.

The first step (uniformising the input) is not enough in the sense of rate-distortion. We also need to adjust the parameter to balance the distortion generated in different intervals, which is the role of the second step. While the exact solution is hard to find, we provide a heuristic, yet efficient, way to make the adjustment.

If we assume that the distortion generated in the interval i is proportional to the squared length Δ_i^2 of the interval, then the average distortion is proportional to $\sum_{i=0}^{M-1} N_i \Delta_i^2$ where N_i is the number of samples inside interval i. Starting with the solution given by step 1, all N_i's are comparable (since it is roughly uniform), and the largest interval contributes the most to the average distortion. Similarly, the smallest interval contributes the least. The idea is therefore to reduce the largest interval until

$$N_S \Delta_S^2 \geq N_L \Delta_L^2, \tag{13}$$

where 'S' and 'L' stand for the 'smallest' and 'largest' intervals, respectively.

Although the presented compander design is based on training data, we can also start with a uniform compander and update it regularly when more data are available. A great advantage of the parametric compander design is the negligible communication overhead of the (few) quantisation parameters.

Remark 1. *It is well known that, followed by entropic encoding, a uniform quantiser is asymptotically optimal in the high-rate regime. We emphasise, however, that here we do not operate in the high-rate regime unlike many other applications. More importantly, a large alphabet size would make the following context-tree-based compression highly inefficient. Hence, a carefully designed quantiser is crucial for the overall performance.*

After the quantisation is done, one has to compress the sequence of quantisation indices. One way to do that is to directly apply CTW with arithmetic coding to this sequence. In the following subsections, we describe an alternative solution that limits the fluctuation of the output bit-stream.

B. Tree Estimation

Given a scalar sequence k_1^n, we use the CTM algorithm (cf. Section II-C) to find the maximum *a posteriori* tree model $\hat{\mathcal{S}}$ that describes that sequence. This algorithm consists in building the same tree $\mathcal{T}_D$ as in CTW algorithm, followed by a pruning procedure as described in Definition 1. Both the computational and storage complexity of CTM algorithm are known to be $O(nmD)$, i.e., linear with sequence length n, alphabet size m and maximum tree depth D, cf. [10].

When training data are available, we can apply the CTM algorithm on the training data to estimate the MAP model $\hat{\mathcal{S}}$, and use it to estimate symbol probabilities and encode the incoming sequence. This, however, is not necessary: we could initialise the full tree $\mathcal{T}_D$ with empty counters, keep updating the counters with incoming data, and regularly prune a copy of this tree to have an updated estimate of the MAP model $\hat{\mathcal{S}}$.

C. Coding Distribution and Encoding

Once a tree model $\hat{\mathcal{S}}$ is estimated, we can encode a sequence k_1^n according to the probabilities issued from that model. Note that, given a model $\hat{\mathcal{S}}$ and past symbols k_{1-D}^0, the estimated probability of a sequence k_1^n can be computed via the KT estimator, using (4) and (5). In particular, denoting $s := c(k_{i-D}^{i-1})$, we can compute the probabilities $\hat{P}(\cdot) = P(\cdot|\hat{\mathcal{S}})$ that the next symbol is $k_i = j$, for all $j \in \mathcal{A}$, as

$$\begin{aligned}\hat{P}(j|k_{i-D}^{i-1}) &= \frac{\hat{P}(k_{i-D}^{i})}{\hat{P}(k_{i-D}^{i-1})} = \frac{\prod_{s'\in\hat{\mathcal{S}}} P_e(\boldsymbol{a}_{s'}(k_1^i))}{\prod_{s'\in\hat{\mathcal{S}}} P_e(\boldsymbol{a}_{s'}(k_1^{i-1}))} \\ &= \frac{P_e(\boldsymbol{a}_s(k_1^i))}{P_e(\boldsymbol{a}_s(k_1^{i-1}))} = \frac{a_s(j)+\frac{1}{2}}{\frac{m}{2}+\sum_{j'\in\mathcal{A}} a_s(j')}.\end{aligned} \tag{14}$$

With $\hat{P}$, one may apply arithmetic coding to encode k_i. But the encoded bits would have a variable length depending on both $\hat{P}$ and k_i. Reducing the fluctuation of the coded bit length is important for practical communication systems. Here, we propose an encoding scheme with three possible codeword lengths, as described below.

Fix two integers $q_1, q_2 \leq \log m$ such that $m_1 := 2^{q_1}$, $m_2 := 2^{q_2}$, and $m_1 + m_2 \leq m$. If k_i is among the m_1 most probable symbols according to $\hat{P}$ (tie could be broken with a fixed rule), then the encoded bit string $\mathbf{c}_i$ is 0 followed by q_1 bits indicating the position of k_i in the list of the m_1 most probable symbols. Otherwise, if k_i is among the next m_2 most probable symbols, the encoded bit string $\mathbf{c}_i$ is 10 followed by q_2 bits indicating the position of k_i in the second list. Finally, if k_i is not among the $m_1 + m_2$ most probable symbols, the encoded bit string $\mathbf{c}_i$ is 11 followed by q_2 bits corresponding to the index $\tilde{k}_i$ from a lower resolution quantiser with size m_3. Hence, in our scheme, we also need to keep a lower resolution quantiser to apply on least probable symbols. It follows that the codeword length is either $1 + q_1$, $2 + q_2$ or $2 + \lceil \log m_3 \rceil$.

D. More Implementation Details

Some more implementation details are omitted and can be found in the long version in [12].

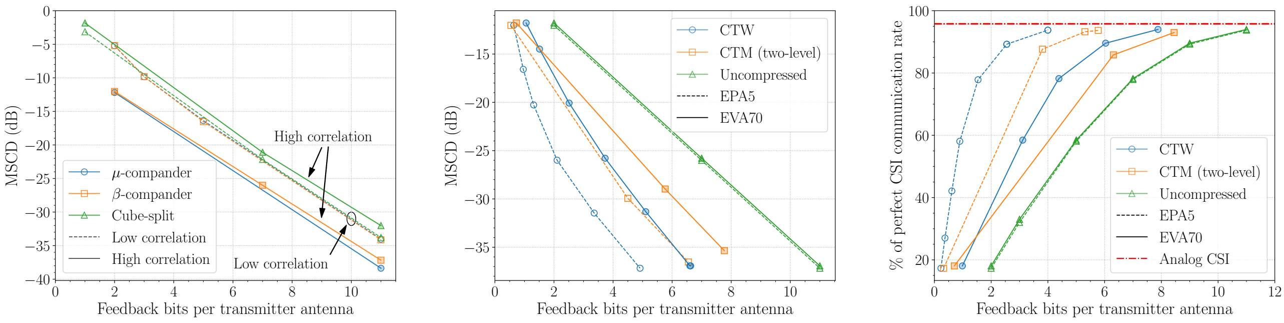
First, the bit allocation between the amplitude and phase quantisations can be optimised to minimise the overall distortion on the complex symbol. We can show that a rule of thumb is to use two more bits on the phase than on the amplitude.

Then, for practical uses, we have multiple trees, each one corresponding to a quantised sequence (amplitude or phase) of a given user and antenna. While each tree provides the marginal distribution of the given sequence, all the marginal distributions can be jointly used to encode the parallel streams together, in order to improve the coding rate.

IV. Simulation Results and Conclusions

We use the MATLAB LTE Toolbox to simulate an LTE MIMO downlink channel, with $N_{\mathrm{t}} = N_{\mathrm{r}} = 4$. We consider both the low mobility (EPA5, Doppler 5 Hz) and high mobility (EVA70, Doppler 70 Hz) scenarios, with either low or high correlation between antennas at the base station. In our implementations, we use $D = 2$, $\gamma = 0.5$ and $q_1 = 0$.

We consider three quantisation schemes: the μ-law compander, the β-law compander, and the cube-split quantiser [15]. Interestingly, the cube-split quantiser can be regarded as a complex compander adapted to the distribution of normalised complex Gaussian vectors. For each quantisation scheme, we



(a) MSCD distortion, EPA5.

(b) MSCD distortion, high antenna correlation.

(c) Communication rate, low antenna correlation.

Fig. 1. Simulation results.

consider three scenarios: no compression, compression with ideal CTW using arithmetic coding [8], and compression with the two-level resolution CTM scheme. The ideal CTW case is simply evaluated with $\frac{1}{n}\left(\left\lceil -\log Q_n(x_1^n|x_{1-D}^0)\right\rceil + 1\right)$.

In all cases, we assess the MSCD versus the average number of CSI bits per antenna, and, for low antenna correlation, we also assess the downlink communication sum rate with zero-forcing beamforming, evaluated approximately using the formula provided in [1, Eq. (20)], at 30 dB. The results are obtained with the best quantisation parameters (sizes of different codebooks) over those that we have tried.

Fig. 1a compares the performance of the different quantisers, with no compression, for low mobility scenario (EPA5). For low antenna correlation, the cube-split and the proposed quantisers achieve almost the same results. On the other hand, when antenna correlation is high, both proposed quantisers have similar performances and are noticeably better than the cube-split (which assumes uniformity of the distribution by design).

In Fig. 1b, we fix the β-law compander and study the performance of different compression methods, under high antenna correlation. The compression gains are significant and can reduce the CSI bits by up to half in the lower rate regime. For EPA5, in the higher rate regime, the two-level CTM scheme can reduce the feedback bits in 4.5 bits and is 1.5 bits away from the CTW performance, approximately. For EVA70, the gains are smaller, due to the lower time correlation. Nevertheless, in the higher rate regime, the two-level CTM can save more than 2 bits, and CTW, more than 4 bits. Furthermore, for EVA70 in the extreme low rate regime, the the two-level CTM outperforms CTW, thanks to the low-resolution quantiser.

Finally, Fig. 1c presents the communication rates for different compression schemes, using the β-law compander. The results are normalised by the rate achieved when perfect (i.e., noiseless) CSI knowledge is available. For both EPA5 and EVA70, we see that the communication rate converges much faster to the analog CSI rate (i.e., with no quantisation) when some of the proposed compression schemes are employed.

More importantly, the proposed schemes have low complexity, can be implemented in an online fashion, and are modular. In particular, the context-tree-based compression scheme can be applied on any other quantisers, including those recently designed with neural networks, e.g., [3]. Similarly, the proposed quantiser can be combined with any other lossless compression schemes.

References

[1] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, “Multiuser MIMO achievable rates with downlink training and channel state feedback,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.

[2] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, “Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.

[3] M. B. Mashhadi, Q. Yang, and D. Gündüz, “Distributed deep convolutional compression for massive MIMO CSI feedback,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2621–2633, 2021.

[4] E. Gassiat, *Universal Coding and Order Identification by Model Selection Methods*. Cham, Switzerland: Springer, 2018.

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

[6] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[7] ——, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[8] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context-tree weighting method: basic properties,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[9] F. Willems, T. Tjalkens, and Y. Shtarkov, “Context-tree maximizing,” in *Proc. 34th Annu. Conf. Inf. Sciences and Syst.*, Princeton, New Jersey, 2000, pp. TP6–7–TP6–12.

[10] I. Kontoyiannis, L. Mertzanis, A. Panotopoulou, I. Papageorgiou, and M. Skoularidou, “Bayesian context trees: Modelling and exact inference for discrete time series,” *arXiv*, 2020. [Online]. Available: https://arxiv.org/pdf/2007.14900v1.pdf

[11] I. H. Witten, R. M. Neal, and J. G. Cleary, “Arithmetic coding for data compression,” *Commun. ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[12] “Context-tree based CSI compression.” [Online]. Available: https://miyamotohk.github.io/context-tree-compression

[13] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA, USA: Kluwer, 1992.

[14] S. S. Dragomir, R. P. Agarwal, and N. S. Barnett, “Inequalities for beta and gamma functions via some classical and new integral inequalities,” *RGMIA Res. Rep. Collection*, vol. 2, no. 3, 1999.

[15] A. Decurninge and M. Guillaud, “Cube-split: Structured quantizers on the Grassmannian of lines,” in *2017 IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2017, pp. 1–6.

The Geometry of Uncoded Transmission for Symmetric Continuous Log-Concave Distributions

Hui-An Shen*, Stefan M. Moser†, and Jean-Pascal Pfister*
*University of Bern and University of Zurich ({huian.shen, jeanpascal.pfister}@unibe.ch)
†ETH Zurich and NYCU Taiwan (moser@isi.ee.ethz.ch)

***Abstract*—We present a geometric picture for optimal single-letter uncoded transmission for *source-channel duals*, where the source and distortion measure are dual to the channel and cost function. In particular, we investigate an additive noise channel with the conditional channel distribution and capacity-achieving input distribution both being symmetric, continuous log-concave densities. We show that under these assumptions, a Gaussian source transmitted over an additive Gaussian channel is the only possible choice for optimal single-letter uncoded transmission. We explain the uniqueness of Gaussian uncoded transmission through a *homothetic property* for the channel input and output typical sets, and illustrate the geometry of single-letter uncoded transmission as opposed to communication based on the classical source-channel separation principle.**

I. Introduction

As proven in Shannon's source-channel separation theorem, it is possible to design information-theoretically optimal systems that achieve both the rate-distortion function of the source and the capacity-cost function of the channel. Unfortunately, in Shannon's approach, the optimal system employs a code with high complexity and infinite delay.

In the context of biological plausibility, [1] proposed optimal *almost* code-free information transmission, for which it is required that both encoder and decoder only perform linear scaling (or are identity functions). For brevity, we will refer to such schemes as *uncoded transmission*.

Uncoded transmission was studied systematically by means of *probabilistic matching* in [2], and it follows that there exist infinite quadruples of source, encoder, channel, and decoder that are optimal over single-letter transmission, i.e., with codewords of length 1 that cause no delay. Such single-letter transmission works when both the distortion measure and the cost function are probabilistically matched to the said quadruple.

Two well-known examples of uncoded transmission are a binary source over a binary symmetric channel, using Hamming distortion (see [1, Ex. 1, Sec. 4.1]), and a Gaussian source over the Gaussian channel with squared-error distortion and second-moment constraint (see [1, Ex. 2, Sec. 4.2]). Note that both cases are information-theoretic *source-channel duals*.[1] This duality permits the aforementioned encoder and decoder to be functionally equivalent (see, e.g., [3]).

It is a curious fact that, for continuous sources and channels, no other source-channel duals for optimal single-letter uncoded transmission have been discovered other than the aforementioned Gaussian case. In the following we are going to partially explain why this is the case. We develop, under mild assumptions, the geometry of optimal uncoded transmission over continuous additive noise channels with its dual source. The assumptions we take are that a) both the channel and the capacity-achieving input distribution are symmetric continuous log-concave distributions, and b) that the aforementioned encoder and decoder are identical linear functions acting on a single letter. The main "geometric reasoning" for restricting to log-concave distributions is to make use of concentration inequalities for log-concave random variables, and to associate its density function with a convex body. The restriction for the encoder and decoder being identical functions arises from the functional equivalence between the source encoder and channel decoder in dual problems of source and channel coding, as introduced in the previous paragraph.

Our main theorems illustrate through a geometric perspective that the well-known Gaussian uncoded transmission is indeed the unique solution for source-channel duals where the latter is the family of all symmetric continuous log-concave additive noise channels. So, e.g., the corresponding situation of an Laplacian source and an ℓ_1-distortion measure that is transmitted over an additive Laplacian noise channel does *not* allow for uncoded transmission. In fact, any generalization of the ℓ_2 case (Gaussian case) to a general ℓ_p-normed case (for $p \neq 2$) does not work.

We show that uncoded transmission naturally arises when the input and output distributions of the channel yield not only "linearly equivalent[2]" but also "*homothetic*[3] typical sets", which we call the *homothetic property*. When the distortion measure is associated with a norm generated by an inner product, this allows for linear single-letter codes that are optimal.

This work was supported by the Swiss National Science Foundation grants PP00P3_179060 (J.-P.P.) and 31003A_175644 (H.-A.S.).

[1] Duality here is in the sense of source coding and channel coding. Therefore, strictly speaking, the duality is between the pairs "source and distortion measure" and "channel and cost function". See Section III for more details.

[2] Two sets $\mathcal{T}_1, \mathcal{T}_2$ are linearly equivalent when there is a nonsingular linear transformation ϕ such that $\phi(\mathcal{T}_1) = \mathcal{T}_2$.

[3] Here *homothetic* is based on the idea of *homothets* of a convex body, as we will define in Section II.

II. Definitions and Notation

We use capitalized Roman alphabets, e.g. X, to denote random variables (RV), with the exception of "V" that is exclusively reserved to denote a real vector space. We write $X \perp\!\!\!\perp Z$ to denote X and Z being independent RVs; f_X denotes the probability density function of the RV X. A density over $\mathbb{R}$ is said to be *symmetric* if $f_X(x) = f_X(-x)$, $\forall x \in \mathbb{R}$. For $k \in \mathbb{N}$, we define $[k] \triangleq \{1, 2, \ldots, k\}$.

We denote $\{\mathbf{e}_i\}_{i\in[n]}$ to be the standard orthonormal basis of an n-dimensional real vector space V. Bold font denotes a tuple or a vector, and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathbb{R}$, refers to either a point $\mathbf{x} \in \mathbb{R}^n$ or $\mathbf{x} = \sum_{i\in[n]} x_i\,\mathbf{e}_i \in V$. In this paper, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \in V$ are used interchangeably.

We use $\operatorname{conv}(\cdot)$ to denote the convex hull. Sets and convex bodies in $\mathbb{R}^n$ are denoted with calligraphic font, e.g. $\mathcal{K}$. A convex body $\mathcal{K} \subset \mathbb{R}^n$ is a compact convex set with nonempty interior. A *homothet* of a convex body $\mathcal{K} \subset \mathbb{R}^n$ is any set with the form $\mathbf{x} + \lambda\mathcal{K} = \{\mathbf{x} + \lambda\mathbf{t} \colon \mathbf{t} \in \mathcal{K}\}$ for some $\mathbf{x} \in \mathbb{R}^n$ and nonzero $\lambda \in \mathbb{R}$.

For an o-symmetric (i.e., centrally symmetric with regard to the origin o) convex body $\mathcal{K}$ in $\mathbb{R}^n$, we use $\|\cdot\|_{\mathcal{K}}$ to denote the norm defined by (the gauge of) $\mathcal{K}$, i.e., for any $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_{\mathcal{K}} = \inf\{\lambda > 0 \colon \mathbf{x} \in \lambda\mathcal{K}\}. \tag{1}$$

In this paper, normed spaces are always real vector spaces.

III. Duality between Source and Channel Coding

This section defines *source-channel duals* as first encountered in Section I, in terms of a dual source coding problem to a channel coding problem. This is based on the classical information-theoretic duality (see, e.g., [3, Sec. III]).

Definition 1: For a channel coding problem with input variable X, output variable Y, conditional channel distribution $P_{Y|X}$, and capacity-achieving input distribution P_X^* inducing the channel output marginal $\bar{P}_Y$, the *dual source coding problem* is the rate-distortion problem over the source variable Y with distribution $\bar{P}_Y$, reconstruction variable X, and the single-letter distortion measure $d \colon \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_0^+$ taking the form

$$d(y, x) = -c_0 \log P_{Y|X}(y|x) + d_0(y), \tag{2}$$

for arbitrary $c_0 > 0$ and $d_0(\cdot)$.
The following theorem is a corollary to [2, Th. 6].

Theorem 2: For the channel coding problem stated in Definition 1, define the induced *backward channel* as

$$\bar{P}_{X|Y}(x|y) = \frac{P_X^*(x) P_{Y|X}(y|x)}{\bar{P}_Y(y)}. \tag{3}$$

Then for the dual source coding problem according to Definition 1, we have

$$\bar{P}_{X|Y}(x|y) = \underset{\substack{Q_{X|Y} \\ \mathsf{E}[d(Y,X)] \le D}}{\operatorname{argmin}} \mathrm{I}\big(\bar{P}_Y, Q_{X|Y}\big), \tag{4}$$

$$\mathrm{I}\big(P_X^*, P_{Y|X}\big) = \min_{\substack{Q_{X|Y} \\ \mathsf{E}[d(Y,X)] \le D}} \mathrm{I}\big(\bar{P}_Y, Q_{X|Y}\big), \tag{5}$$

where the distortion $D = \mathsf{E}_{Q^*}[d(Y, X)]$ with $Q^*(y, x) = P_X^*(x) P_{Y|X}(y|x)$.

IV. Uncoded Transmission in Normed Spaces

A. Preliminaries Regarding Normed and Inner Product Spaces

Definition 3 (Real normed linear spaces): A real normed linear space $(V, \|\cdot\|)$ is a real vector space V with a function (called *norm*) $\|\cdot\| \colon V \to \mathbb{R}$, $\mathbf{x} \mapsto \|\mathbf{x}\|$ satisfying the following properties:

(1) $\|\mathbf{x}\| \ge 0$, $\forall \mathbf{x} \in V$;
(2) $\|\mathbf{x}\| = 0$ if, and only if, $\mathbf{x} = \mathbf{0}$;
(3) $\|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\|$, $\forall \lambda \in \mathbb{R}$, $\forall \mathbf{x} \in V$;
(4) $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in V$.

Definition 4 (Real inner product spaces): A real inner product space $(V, \langle\cdot,\cdot\rangle)$ is a real vector space with a function (called *inner product*) $\langle\cdot,\cdot\rangle \colon V \times V \to \mathbb{R}$, $(\mathbf{x}, \mathbf{y}) \mapsto \langle\mathbf{x}, \mathbf{y}\rangle$ satisfying the following properties:

(1) $\langle\mathbf{x}, \mathbf{x}\rangle \ge 0$, $\forall \mathbf{x} \in V$;
(2) $\langle\mathbf{x}, \mathbf{x}\rangle = 0$ if, and only if, $\mathbf{x} = \mathbf{0}$;
(3) $\langle\mathbf{x}, \mathbf{y}\rangle = \langle\mathbf{y}, \mathbf{x}\rangle$, $\forall \mathbf{x}, \mathbf{y} \in V$;
(4) $\langle\lambda\mathbf{x}, \mathbf{y}\rangle = \lambda\langle\mathbf{x}, \mathbf{y}\rangle$, $\forall \lambda \in \mathbb{R}$, $\forall \mathbf{x}, \mathbf{y} \in V$;
(5) $\langle\mathbf{x} + \mathbf{z}, \mathbf{y}\rangle = \langle\mathbf{x}, \mathbf{y}\rangle + \langle\mathbf{z}, \mathbf{y}\rangle$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$.

Theorem 5: For any real inner product space $(V, \langle\cdot,\cdot\rangle)$, define

$$\|\mathbf{x}\| \triangleq \sqrt{\langle\mathbf{x}, \mathbf{x}\rangle}, \quad \forall \mathbf{x} \in V. \tag{6}$$

Then $(V, \|\cdot\|)$ is a real normed linear space.

Definition 6: Let $\|\cdot\|$ be a norm on V. We say "$(V, \|\cdot\|)$ is an inner product space" when there exists an inner product $\langle\cdot,\cdot\rangle$ such that (6) is satisfied.

B. Preliminaries Regarding Log-concavity

Definition 7 (Log-concave functions): A function $f \colon \mathbb{R}^n \to [0, +\infty)$ is said to be *log-concave* if it has the form

$$f = e^{-g} \tag{7}$$

where $g \colon \mathbb{R}^n \to (-\infty, \infty]$ is a convex function.

Definition 8: We say "X is a log-concave random variable" when f_X is a log-concave function.

Proposition 9 (Log-concavity preserved over convolution [4]): If X, Z are independent log-concave random variables, then $X + Z$ is also log-concave.

C. Main Theorem 1

Definition 10: The differential entropy of a probability density function f is given as

$$\mathrm{h}(f) \triangleq -\int_{-\infty}^{\infty} f(s) \log f(s)\,\mathrm{d}s. \tag{8}$$

Definition 11: Let Ψ be the family of continuous symmetric log-concave probability density functions over $\mathbb{R}$, and Ω be the collection of compact convex sets in $\mathbb{R}^n$ with the collection of their boundaries $\partial\Omega$. For $f \in \Psi$, define $\Phi_n \colon \Psi \to \partial\Omega$, $f \mapsto \Phi_n(f)$ as follows:

$$\Phi_n(f) \triangleq \left\{\mathbf{x} \in \mathbb{R}^n \colon -\frac{1}{n}\sum_{i=1}^{n} \log f(x_i) = \mathrm{h}(f)\right\}. \tag{9}$$

Theorem 12 (Main Theorem 1): For any $n \geq 2$, let V be an n-dimensional real vector space, and let $\mathcal{B}_n(f)$ denote the convex hull of $\Phi_n(f)$. For a log-concave random variable $Z \in \mathbb{R}$ with a symmetric, continuous density f_Z, define the family $\mathcal{F}$ of log-concave random variables $X \in \mathbb{R}$ where $X \perp\!\!\!\perp Z$ and f_X is symmetric, continuous and satisfies

$$t\Phi_n(f_X) = \Phi_n(f_Z), \qquad \text{for some } t > 0. \tag{10}$$

Then the following three conditions are equivalent:

(i) For all $X \in \mathcal{F}$, $\big(V, \|\cdot\|_{\mathcal{K}}\big)$ is an inner product space for $\mathcal{K} = \mathcal{B}_n(f_X)$.

(ii) For all $X \in \mathcal{F}$, there exists some $\alpha_t > 0$ which only depends on t, such that

$$\Phi_n(f_{X+Z}) = \alpha_t \Phi_n(f_X). \tag{11}$$

(iii) For all $X \in \mathcal{F}$ and $\mathcal{K} = \mathcal{B}_n(f_X)$, there exists some $\alpha_t > 0$ which only depends on t, such that for any pair of $\mathbf{y} \in \Phi_n(f_{X+Z})$, $\mathbf{x} \in \Phi_n(f_X)$,

$$\|\mathbf{y} - \mathbf{x}\|_{\mathcal{K}} = \left\| \frac{1}{\alpha_t}\mathbf{y} - \alpha_t \mathbf{x} \right\|_{\mathcal{K}} \tag{12}$$

and $\alpha_t \neq 1$.

Proof: See Appendices A and B. ■

Remark 13: Property (ii) describes the "homothetic property" mentioned in Section I, and Property (iii) describes the geometry of uncoded transmission with a linear single-letter code over the source-channel duals. We will discuss these conditions further in Sections IV-D and IV-E.

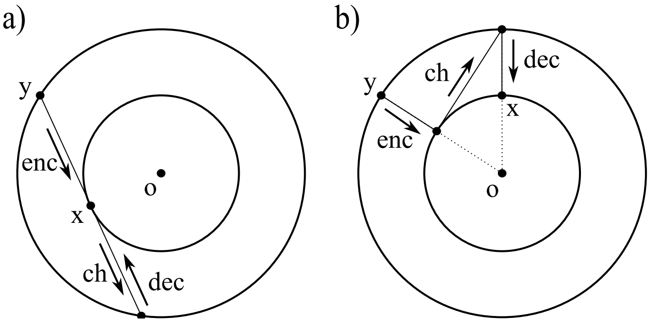


Fig. 1. A schematic comparison of a) optimal infinite-letter coded transmission, and b) optimal single-letter (scaled) uncoded transmission. Both source-channel communication systems consist sequentially of an encoder (enc), a channel (ch), and a decoder (dec). In a), the encoder and decoder are jointly typical encoder/decoders that act on infinite-length sequences. In b), the encoder and decoder can act on single symbols by scaling due to the homothetic property.

D. Main Theorem 2: Strengthening Theorem 12

In this section, we relax two assumptions in Theorem 12 to obtain Theorem 14: we do not require that f_X and f_Z yield linearly equivalent sets by Φ_n (i.e., assumption (10) is relaxed), and we also do not require in the third condition that $\mathcal{K}$ is a convex body associated with f_X (i.e., $\mathcal{K} = \mathcal{B}_n(f_X)$ in Condition (iii) of Theorem 12 is relaxed).

Theorem 14 (Main Theorem 2): For any $n \geq 2$, let V be an n-dimensional real vector space. Let $Z \in \mathbb{R}$ be a log-concave random variable with a symmetric, continuous density f_Z. Define the family $\mathcal{F}$ of log-concave random variables over $\mathbb{R}$ with symmetric, continuous densities satisfying for any $X_i, X_j \in \mathcal{F}$,

$$u\Phi_n(f_{X_i}) = \Phi_n(f_{X_j}), \qquad \text{for some } u > 0; \tag{13}$$

and $X \perp\!\!\!\perp Z$ for any $X \in \mathcal{F}$. Then the following three conditions are equivalent:

(I) The RV Z and all RVs in the family $\mathcal{F}$ are Gaussian.

(II) For all $X_i \in \mathcal{F}$, there exists some $\alpha_i > 0$, such that

$$\Phi_n(f_{X+Z}) = \alpha_i \Phi_n(f_X). \tag{14}$$

(III) For all $X_i \in \mathcal{F}$, and for any norm $\|\cdot\|$ generated by an inner product on V, there exists some $\alpha_i > 0$, such that for any pair of $\mathbf{y} \in \Phi_n(f_{X+Z})$, $\mathbf{x} \in \Phi_n(f_X)$,

$$\|\mathbf{y} - \mathbf{x}\| = \left\| \frac{1}{\alpha_i}\mathbf{y} - \alpha_i \mathbf{x} \right\| \tag{15}$$

and $\alpha_i \neq 1$.

Proof: We omit the proof of Theorem 14 for it is similar to that of Theorem 12. We only remark that for showing (II) $\Longrightarrow$ (I), we make use of the following proposition.

Proposition 15: For $X \perp\!\!\!\perp Z$, if f_X and f_{X+Z} are both zero-mean Gaussian distributions, then f_Z is also Gaussian. ■

The graphical representation of Gaussian uncoded transmission, which uniquely satisfies Properties (i)–(iii) and (I)–(III) in Theorem 12 and 14, is given in Fig. 1b).

E. Interpreting the Main Theorems for the Geometry of Uncoded Transmission

To see how the main theorems, Theorems 12 and 14, explain the geometry of uncoded transmission, we first introduce the following theorem.

Theorem 16 ([5, Th. 2]): Let $\mathbf{X}^{(n)}$ be a random vector in $\mathbb{R}^n$ with log-concave density f. Then for any $0 \leq t \leq 2$,

$$\Pr\left[\frac{1}{n}\left|\log f\big(\mathbf{X}^{(n)}\big) - \mathsf{E}\left[\log f\big(\mathbf{X}^{(n)}\big)\right]\right| \geq t\right] \leq 4\, e^{-ct^2 n} \tag{16}$$

where $c \geq \frac{1}{16}$.

Intuitively, Theorem 16 describes how the random vector realizations concentrate near a 'thin shell' as $n \to \infty$. In the special case when each element of the random vector is generated IID from f_X, a log-concave density over $\mathbb{R}$, this 'thin shell' lies around $\Phi_n(f_X)$ (see Definition 11) and can be understood as the *typical set* of X for some n sufficiently large. Note that in the main theorems, both X and $X + Z$ are log-concave random variables and thus Theorem 16 applies. In this context, loosely speaking, Property (ii) and (II) state that the typical sets of $X + Z$ and X are linearly equivalent as $n \to \infty$. We call this the "homothetic property", and depict it in Fig. 1 with the outer and inner sphere, respectively, representing the typical sets of $X + Z$ and X.

Let the channel coding problem be on input variable X with capacity-achieving distribution P_X^*, output variable $Y = X + Z$ with additive channel noise Z. Then using Definition 1 we have the dual source coding problem on source variable Y and reconstruction variable X. In this case we have the following communication system on a source-channel dual with length-n letters:

$$\mathbf{Y}^{(n)} \xrightarrow{\text{enc}} \mathbf{x}^{(n)} \xrightarrow{\text{ch}} \mathbf{Y}^{(n)} \xrightarrow{\text{dec}} \mathbf{X}^{(n)}, \tag{17}$$

where 'enc' stands for 'encoder', 'ch' for 'channel', and 'dec' for 'decoder'. As n becomes sufficiently large, we can understand the system in (17) as acting on sequences in the typical set for arbitrary $\epsilon > 0$:

$$\mathcal{T}_\epsilon^{(n)}(Y) \xrightarrow{\text{enc}} \mathcal{T}_\epsilon^{(n)}(X) \xrightarrow{\text{ch}} \mathcal{T}_\epsilon^{(n)}(Y) \xrightarrow{\text{dec}} \mathcal{T}_\epsilon^{(n)}(X). \tag{18}$$

The classical scenario for optimal source-channel dual communication system is shown in Fig. 1a), where the encoder is the *jointly typical encoder*, and the decoder is the *jointly typical decoder*, which 'undoes' the channel (i.e., we have reliable transmission). However, both the encoder and decoder in the classical scenario need to act on vectors of length n. In contrast, the homothetic property presented in the two main theorems allows for uncoded transmission, as shown in Fig. 1b). This means that the encoder can perform a linear scaling isotropically independent of the position of the source sequence $\mathbf{y}$ on the outer sphere. This makes it possible for the encoder to act on a single-letter allowing for uncoded transmission. Since the decoder is equivalent to the encoder due to the imposed source-channel duality, the same reasoning also applies to the decoder.

Appendix A "Homothetic Property" of Uncoded Transmission and the Inner Product Space

A. Characterizations of Inner Product Spaces

Since an inner product space is a normed linear space with extra structure (see also Theorem 5 and Definition 6), there are properties that characterize inner product spaces which help us determine whether the norm arises from an inner product in the sense of (6). One such characterization is the well-known *Parallelogram Law*[4] [6]. Another such characterization is presented as follows.

Theorem 17 ([7, Th. 2.2.]): Let $X = (V, \|\cdot\|)$ be a 2-dimensional real normed linear space and let its unit sphere be $\mathcal{S}_X \triangleq \{\mathbf{x} \in \mathbb{R}^2 \colon \|\mathbf{x}\| = 1\}$. Then X is an inner product space if, and only if, $\mathcal{S}_X$ is an ellipse.

In the following section, we use the characterization presented in Theorem 17 to give a proof for (i) $\implies$ (ii) of Theorem 12. We isolate this particular part of the proof of Theorem 12 to illustrate the underlying geometric property (as in Theorem 17) of an inner product space, before we present the rest of the proof for Theorem 12 in Appendix B.

B. Proof for (i) $\implies$ (ii) in Theorem 12

1) For $n = 2$, we apply Theorem 17 to see that the unit sphere $\Phi_2(f_X)$ is an ellipse. Taking $n = 2$ for (9), we immediately see that this ellipse $\Phi_2(f_X)$ must be a two-dimensional ℓ_2-sphere. Furthermore, using the fact that f_X is a symmetric log-concave density, we can write it in the form

$$f_X(x) = c_1\, e^{-c_0 x^2}, \quad \text{where } c_1, c_0 > 0. \tag{19}$$

2) For $n > 2$, since (i) holds, let $\langle\cdot,\cdot\rangle_V$ be an inner product satisfying

$$\|\mathbf{x}\|_{\mathcal{K}} = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_V}, \quad \forall \mathbf{x} \in V. \tag{20}$$

Let $V' = \operatorname{span}(\{\mathbf{e}_1, \mathbf{e}_2\})$, then

$$\langle \mathbf{x}', \mathbf{y}'\rangle_{V'} \triangleq \langle \mathbf{x}', \mathbf{y}'\rangle_V, \quad \forall \mathbf{x}', \mathbf{y}' \in V', \tag{21}$$

is an inner product on this subspace. Thus, $(V', \langle\cdot,\cdot\rangle_{V'})$ is an inner product space with the unit sphere

$$\mathcal{S}' = \Phi_n(f_X) \cap V' \tag{22}$$

$$= \left\{ \mathbf{x} \in \mathbb{R}^2 \colon \frac{-1}{n}\left(\sum_{i=1}^{2} \log f_X(x_i) + \sum_{i=3}^{n} \log f_X(0) \right) = \mathrm{h}(f_X) \right\}. \tag{23}$$

Since $\dim V' = 2$, it follows from Theorem 17 that $\mathcal{S}'$ is an ellipse, and therefore, similarly to 1), we can write f_X in the form of (19).

Combining 1) and 2) then yields (19) for any $n \geq 2$. Using (19), (10) and that X, Z are independent, we obtain that for any $t > 0$,

$$\Phi_n(f_{X+Z}) = \sqrt{1 + t^2}\, \Phi_n(f_X). \tag{24}$$

Letting $\alpha_t = \sqrt{1 + t^2} > 0$ in (24) we conclude (i) $\implies$ (ii) for any $n \geq 2$.

Appendix B Proof of Theorem 12

The proof consists of three parts that together prove the equivalence of Conditions (i)–(iii). The standard dot-product for $\mathbf{x}, \mathbf{y} \in V$ is denoted $\mathbf{x} \cdot \mathbf{y}$ and defined as

$$V \times V \to \mathbb{R},\ (\mathbf{x}, \mathbf{y}) \mapsto \sum_{i \in [n]} x_i\, y_i, \tag{25}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

A. (ii) $\implies$ (i)

We start proving (ii) $\implies$ (19). Pick an arbitrary $X \in \mathcal{F}$ and define $Y_0 = X$, let $k \in \mathbb{N}$, and for all $i \in [k]$, let Y_{i-1}, Z_i be independent RVs, where $\{Z_i\}$ are IID and where

$$Y_i = Y_{i-1} + Z_i. \tag{26}$$

[4] A normed linear space $(V, \|\cdot\|)$ is an inner product space if, and only if,

$$2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{x} + \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in V.$$

By Proposition 9, Y_i has log-concave (symmetric) density, so we can apply (10) and (ii) recursively (starting from $i = 1$) to obtain a sequence of $\{\alpha_{t_i}\}_{i\in[k]}$ satisfying

$$\Phi_n(f_{Y_i}) = \alpha_{t_i}\Phi_n(f_{Y_{i-1}}) \quad \forall i \in [k]. \tag{27}$$

From this we then obtain

$$\Phi_n(f_{Y_k}) = \left(\prod_{i=1}^{k}\alpha_{t_i}\right)\Phi_n(f_{Y_0}). \tag{28}$$

By letting $\beta_k \triangleq 1/\prod_{i=1}^{k}\alpha_{t_i}$, we can rewrite (28) as

$$\Phi_n(f_X) = \beta_k\Phi_n(f_{Y_k}), \quad \beta_k > 0. \tag{29}$$

Note that by the central limit theorem, $Y_k/k = X/k + \frac{1}{k}\sum_{i=1}^{k} Z_i$ converges to a zero-mean Gaussian RV as k becomes sufficiently large. (Note that the central limit theorem applies because a log-concave density has finite moments of all orders and thus also finite variance.) Also, for a Gaussian RV U it holds that $\Phi_n(f_{U/k}) = \gamma_k\Phi_n(f_U)$ for some factor γ_k. Thus, and since (29) holds for all $k \in \mathbb{N}$, we take $k \to \infty$ and conclude that $\Phi_n(f_X)$ is an ℓ_2-sphere and that f_X either takes the form of (19) or is a Dirac delta. However the latter is excluded due to the continuity assumption of f_X.

Next, we are going to prove that (19) $\Longrightarrow$ (i). Using (19) and the definition of (9) we obtain

$$\Phi_n(f_X) = \left\{\mathbf{x} \in \mathbb{R}^n \colon \sum_{i=1}^{n} x_i^2 = \frac{n}{2c_0}\right\} \tag{30}$$

to be the unit sphere of the normed space $(V, \|\cdot\|_{\mathcal{K}})$. Now we define an inner product $\langle\cdot,\cdot\rangle$ on V (which satisfies all five properties in Definition 4) as

$$\langle\mathbf{x},\mathbf{y}\rangle \triangleq \frac{\mathbf{x}\cdot\mathbf{y}}{r_0^2}, \quad \forall\mathbf{x},\mathbf{y} \in V, \tag{31}$$

with $r_0 \triangleq \sqrt{\frac{n}{2c_0}}$. Then by applying (1),

$$\begin{aligned}
&\|\mathbf{x}\|_{\mathcal{K}} = \inf\{\lambda > 0 \colon \mathbf{x} \in \lambda\mathcal{B}_n(f_X)\}\\
&\Longrightarrow \|\mathbf{x}\|_{\mathcal{K}}\mathbf{v} = \mathbf{x} \quad \text{where } \mathbf{v} \in \Phi_n(f_X) && (32)\\
&\Longrightarrow \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathcal{K}}} \in \Phi_n(f_X). && (33)
\end{aligned}$$

Using (33), (30) and (31) we get

$$\|\mathbf{x}\|_{\mathcal{K}} = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{r_0^2}} = \sqrt{\frac{\mathbf{x}\cdot\mathbf{x}}{r_0^2}} = \sqrt{\langle\mathbf{x},\mathbf{x}\rangle}, \tag{34}$$

i.e., $\|\mathbf{x}\|_{\mathcal{K}} = \sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$, $\forall\mathbf{x} \in V$. Therefore, by Definition 6, we say that "$(V, \|\cdot\|_{\mathcal{K}})$ is an inner product space", which holds for arbitrary $X \in \mathcal{F}$ and $\mathcal{K} = \Phi_n(f_X)$, concluding our proof.

B. (i) $\Longrightarrow$ (iii) (by way of (ii))

Since (i) holds, there exists an inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ such that $\|\mathbf{x}\|_{\mathcal{K}} = \sqrt{\langle\mathbf{x},\mathbf{x}\rangle_{\mathcal{K}}}$, $\forall\mathbf{x} \in V$. Using this and the properties of the inner product we can write

$$\left\|\frac{1}{\alpha_t}\mathbf{y} - \alpha_t\mathbf{x}\right\|_{\mathcal{K}}^2 = \left\langle\frac{1}{\alpha_t}\mathbf{y} - \alpha_t\mathbf{x}, \frac{1}{\alpha_t}\mathbf{y} - \alpha_t\mathbf{x}\right\rangle_{\mathcal{K}} \tag{35}$$

$$\begin{aligned}
&= \alpha_t^2\langle\mathbf{x},\mathbf{x}\rangle_{\mathcal{K}}^2 + \frac{\langle\mathbf{y},\mathbf{y}\rangle_{\mathcal{K}}^2}{\alpha_t^2} - 2\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{K}} && (36)\\
&= \alpha_t^2\|\mathbf{x}\|_{\mathcal{K}}^2 + \frac{\|\mathbf{y}\|_{\mathcal{K}}^2}{\alpha_t^2} - 2\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{K}} && (37)\\
&= \alpha_t^2 + \frac{\|\mathbf{y}\|_{\mathcal{K}}^2}{\alpha_t^2} - 2\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{K}}. && (38)
\end{aligned}$$

Similarly we have

$$\|\mathbf{y}-\mathbf{x}\|_{\mathcal{K}}^2 = \langle\mathbf{y}-\mathbf{x},\mathbf{y}-\mathbf{x}\rangle_{\mathcal{K}} = 1 + \|\mathbf{y}\|_{\mathcal{K}}^2 - 2\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{K}}. \tag{39}$$

Because in Section A-B we already proved (i) $\Longrightarrow$ (ii), we see that (ii) holds, which implies $\|\mathbf{y}\|_{\mathcal{K}}^2 = \alpha_t^2$, where $\alpha_t > 0$ only depends on t. Applying this to (38) and (39) we get

$$\left\|\frac{1}{\alpha_t}\mathbf{y} - \alpha_t\mathbf{x}\right\|_{\mathcal{K}}^2 = \alpha_t^2 + 1 - 2\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{K}} = \|\mathbf{y}-\mathbf{x}\|_{\mathcal{K}}^2, \tag{40}$$

where $\alpha_t > 0$ only depends on t. Using Property (1) of the norm in Definition 3 for (40) we obtain (12), which concludes the proof.

C. (iii) $\Longrightarrow$ (ii)

Since (12) holds for any $\mathbf{x} \in \Phi_n(f_X)$, $\mathbf{y} \in \Phi_n(f_{X+Z})$, we know that (12) also holds for any $\mathbf{y}_0 \in \Phi_n(f_{X+Z})$ and $\mathbf{x}_0 \triangleq \mathbf{y_0}/\|\mathbf{y}_0\|_{\mathcal{K}} \in \Phi_n(f_X)$. Taking $\mathbf{x} = \mathbf{x}_0$, $\mathbf{y} = \mathbf{y}_0$ in (12) we obtain

$$\begin{aligned}
&\left\|\mathbf{y}_0 - \frac{\mathbf{y}_0}{\|\mathbf{y}_0\|_{\mathcal{K}}}\right\|_{\mathcal{K}} = \left\|\frac{1}{\alpha_t}\mathbf{y}_0 - \frac{\alpha_t}{\|\mathbf{y}_0\|_{\mathcal{K}}}\mathbf{y}_0\right\|_{\mathcal{K}}\\
&\Longrightarrow \left|1 - \frac{1}{\|\mathbf{y}_0\|_{\mathcal{K}}}\right| = \left|\frac{1}{\alpha_t} - \frac{\alpha_t}{\|\mathbf{y}_0\|_{\mathcal{K}}}\right| && (41)\\
&\Longrightarrow \|\mathbf{y}_0\|_{\mathcal{K}} = \alpha_t, \quad \text{where } \alpha_t > 0. && (42)
\end{aligned}$$

In (42) $\|\mathbf{y}_0\|_{\mathcal{K}} = -\alpha_t$ is not valid because $\|\mathbf{y}_0\|_{\mathcal{K}} \geq 0$.

Because (42) holds for any $\mathbf{y}_0 \in \Phi_n(f_{X+Z})$ and $\Phi_n(f_X)$ is the unit sphere of $(V, \|\cdot\|_{\mathcal{K}})$, we obtain

$$\Phi_n(f_{X+Z}) = \alpha_t\Phi_n(f_X), \tag{43}$$

which concludes the proof.

Acknowledgment

We thank Dr. Simone Carlo Surace for drawing the authors' attention to the optimal Gaussian source channel communication system, which helped commence this project.

References

[1] T. Berger, "Living information theory," in *Proc. IEEE Int. Symp. Inf. Theory*, Lausanne, Switzerland, Jun. 30 – Jul. 5, 2002.

[2] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[3] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1181–1203, May 2003.

[4] M. Merkle, "Convolutions of logarithmically concave functions," *Publikacije Elektrotehničkog fakulteta — Serija Matematika*, vol. 9, pp. 1543–1549, 1998.

[5] S. Bobkov and M. Madiman, "An equipartition property for high-dimensional log-concave distributions," in *Proc. 50th Allerton Conf. Commun., Control Comput.*, Monticello, IL, USA, Oct. 1–5, 2012, pp. 482–488.

[6] P. Jordan and J. v. Neumann, "On inner products in linear, metric spaces," *Ann. Math.*, vol. 36, no. 3, pp. 719–723, Jul. 1935.

[7] M. M. Day, "Some characterizations of inner-product spaces," *Trans. Amer. Math. Soc.*, vol. 62, no. 2, pp. 320–337, Sept. 1947.

A Rate-Distortion-Perception Theory for Binary Sources

Jingjing Qian[†], George Zhang[⋆], Jun Chen[†], Ashish Khisti[⋆]
[†]McMaster University
[⋆]University of Toronto

Abstract—**Building upon a series of recent works on perception-constrained lossy compression, we develop a rate-distortion-perception theory for binary sources under Hamming distortion and TV perception losses. It includes a closed-form expression of the rate-distortion-perception function in the one-shot setting, a complete characterization of the distortion-perception region for an arbitrary representation, partially tight upper and lower bounds on the minimum rate penalty for universal representations, a necessary and sufficient condition for point-wise successive refinement, and a sufficient condition for the successive refinability of universal representations.**

I. Introduction

Recently, there has been an upsurge of research on perception-constrained lossy compression for images or videos. Within traditional compression, the well-established rate-distortion formulation is to minimize some notion of distortion under the condition that the given bit rate is not exceeded. In contrast, perceptually-constrained lossy compression takes into account the notion of *perceptual quality*, which turns out to be distinct from the notion of distortion, as well. The motivation for considering both the traditional distortion and the perceptual quality comes from the fact that in many cases, minimizing distortion does not produce visually pleasing results. Such a fact was exemplified by many remarkable deep learning enhanced lossy compression works capable of operating at extremely low rates, such as [1], [2]. In perception-constrained lossy compression, instead, the target is to find the best tradeoff among three quantities —- rate, distortion and perception. Following the success of deep learning for lossy compression, a mathematical view of this topic was initiated and investigated by Blau and Michaeli [3].

Concretely, perceptual quality aims to quantify the degree of visual satisfaction as measured by the human visual perception system and unlike distortion is taken to be fully no-reference (i.e., not with respect to any *particular* source sample, such as a single image or video). Blau and Michaeli adopt a notion of perceptual quality defined by the divergence (e.g., the Kullback-Leibler divergence, the Wasserstein distance, and the total variation (TV) distance) between the *distribution* of the original source and that of the reconstruction, with the property that perfect perceptual quality is obtained only when the two distributions are identical. By basing this measure on the distributions, we again emphasize that the perceptual quality is in fact a global and inherently no-reference measure of the reconstruction quality. In contrast, the distortion is a local one, expressed in terms of the symbol-by-symbol "distance". As mentioned previously, it may not be possible to attain both low distortion and high perceptual quality at the same time, in the sense that one quality must be sacrificed to improve the other one [2]. Optimizing the tradeoff between distortion and perception, incorporated with distribution-preserving lossy compression [4], is the central idea in studying the rate-distortion-perception tradeoff in the seminal work [3]. However, we note that various versions of distribution-constrained lossy compression have been studied before [3] in information theory literature (e.g., [5]–[7]).

In this paper, we investigate the tradeoff among rate, distortion, and perception for binary sources. The distortion considered here is the Hamming distortion and the perception quality is measured by the TV distance. We first derive a closed-form expression for the rate-distortion-perception tradeoff in the one-shot setting. This is followed by a complete characterization of the achievable distortion-perception region for a general representation. We then consider the universal setting [8] in which the encoder is one-size-fits-all, and derive upper and lower bounds on the minimum rate penalty. Finally, we study successive refinement for both point-wise and set-wise versions of perception-constrained lossy compression. A necessary and sufficient condition for point-wise successive refinement and a sufficient condition for the successive refinability of universal representations are provided.

II. Problem Definitions and Known Results

A. Rate-Distortion-Perception Function and Universal Representation

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a distortion measure and $\omega : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ be a divergence, where $\mathcal{X}$ is the source/reconstruction alphabet and $\mathcal{P}(\mathcal{X})$ denotes the set of distributions defined on $\mathcal{X}$. We assume that ω is convex in its second argument. Let Θ be a non-empty set of (D, P) pairs with each pair being a distortion-perception objective.

Definition 1 (One-Shot Rate-Distortion-Perception Function): A rate R is said to be *one-shot achievable* with respect to Θ for the source variable X if we can find a random seed U (which is independent of X) and an encoder $p_{V|XU}$ with $H(V|U) \leq R$ such that for every $(D, P) \in \Theta$, a decoder $p_{\hat{X}|VU}$ can be constructed to meet the constraints $\mathbb{E}[d(X, \hat{X})] \leq D$ and $\omega(p_X, p_{\hat{X}}) \leq P$, where the joint distribution $p_{XV\hat{X}U}$ is assumed to factor as $p_X p_U p_{V|XU} p_{\hat{X}|VU}$. The infimum of such R is denoted by $R^*(\Theta)$. In the case where

Θ consists of a single (D,P) pair, we simply write $R^*(\Theta)$ as $R^*(D,P)$ and refer to it as the *one-shot rate-distortion-perception function.*

The random seed U acts as a shared source of randomness, which plays an important role in our formulation. Note that with U available at both the encoder and decoder, V can be losslessly represented by approximately $H(V|U)$ bits using variable-length codes. This provides an operational justification of the rate constraint $H(V|U) \leq R$.

Let $\mathcal{P}_{Z|X}(\Theta)$ denote the set of conditional distributions $p_{Z|X}$ such that for every $(D,P) \in \Theta$, there exists a conditional distribution $p_{\hat{X}|Z}$ satisfying $\mathbb{E}[d(X,\hat{X})] \leq D$ and $\omega(p_X, p_{\hat{X}}) \leq P$, where the joint distribution $p_{XZ\hat{X}}$ is assumed to factor as $p_X p_{Z|X} p_{\hat{X}|Z}$. Define

$$R(\Theta) \triangleq \inf_{p_{Z|X} \in \mathcal{P}_{Z|X}(\Theta)} I(X;Z).$$

Theorem 1: It holds that $R(\Theta) \leq R^*(\Theta) \leq R(\Theta) + \log(R(\Theta)+1) + 4$. Moreover, in the case where Θ consists of a finite number of (D,P) pairs,

$$\begin{aligned} R^*(\Theta) = & \inf_{p_{\hat{X}_\Theta U|X}} H(\hat{X}_\Theta|U) & (1)\\ \text{subject to} \quad & I(X;U) = 0, & (2)\\ & H(\hat{X}_\Theta|X,U) = 0, & (3)\\ & \mathbb{E}[d(X,\hat{X}_{D,P})] \leq D, \quad (D,P) \in \Theta, & (4)\\ & \omega(p_X, p_{\hat{X}_{D,P}}) \leq P, \quad (D,P) \in \Theta, & (5) \end{aligned}$$

where $X_\Theta = \{\hat{X}_{D,P}\}_{(D,P)\in\Theta}$.

Proof: The first statement was established in [8] (see also [9] for the special case $\Theta = \{(D,P)\}$).The second statement follows by showing that there is no loss of optimality in setting $V = \hat{X}_\Theta$ and restricting it to be a deterministic function of (X,U). ■

Definition 2 (Asymptotic Rate-Distortion-Perception Function): A rate R is said to be *asymptotically achievable* with respect to Θ for the i.i.d. source sequence $\{X(t)\}_{t=1}^{\infty}$ with each component following the distribution p_X if for some positive integer n, we can find a random seed U and an encoder $p_{V|X^nU}$ with $\frac{1}{n}H(V|U) \leq R$ such that for every $(D,P) \in \Theta$, a decoder $p_{\hat{X}^n|VU}$ can be constructed to meet the constraints $\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[d(X(t),\hat{X}(t))] \leq D$ and $\omega(p_X, \frac{1}{n}\sum_{t=1}^{n} p_{\hat{X}(t)}) \leq P$, where the joint distribution $p_{X^nV\hat{X}^nU}$ is assumed to factor as $p_{X^n}p_U p_{V|X^nU} p_{\hat{X}^n|VU}$. The infimum of such R is denoted by $R^{(\infty)}(\Theta)$. In the case where Θ consists of a single (D,P) pair, we simply write $R^{(\infty)}(\Theta)$ as $R^{(\infty)}(D,P)$ and refer to it as the *asymptotic rate-distortion-perception function.*

As a consequence of Theorem 1, the following result holds [8] (see also [6], [7], [9] for the special case $\Theta = \{(D,P)\}$).

Theorem 2: We have $R^{(\infty)}(\Theta) = R(\Theta)$.

In view of Theorem 2, the asymptotic source coding rate is completely characterized by $R(\Theta)$, and such a quantity is expressed in terms of optimization over random variables Z satisfying certain conditions. Hence, we can interpret any random variable Z jointly distributed with X as a *representation* (or reconstruction random variable) of X.

Definition 3 (Universal Representation): Given a representation Z of X, its distortion-perception region, denoted by $\Pi(p_{Z|X})$, is the set of all (D,P) pairs for which there exists $p_{\hat{X}|Z}$ satisfying $\mathbb{E}[d(X,\hat{X})] \leq D$ and $\omega(p_X,p_{\hat{X}}) \leq P$, where the joint distribution $p_{XZ\hat{X}}$ is assumed to factor as $p_X p_{Z|X} p_{\hat{X}|Z}$. We say that Z is a Θ-universal representation of X if $\Theta \subseteq \Pi(p_{Z|X})$.

Note that $R^{(\infty)}$ is the minimum rate needed for a fixed encoder to cope with the distortion-perception objectives in Θ. In light of Theorem 2, it also coincides with the infimum of $I(X;Z)$ over all Θ-universal representations Z of X. On the other hand, $\sup_{(D,P)\in\Theta} R^{(\infty)}(D,P)$ is the rate required to meet the most demanding objective in Θ. As such, $\Delta(\Theta) \triangleq R^{(\infty)}(\Theta) - \sup_{(D,P)\in\Theta} R^{(\infty)}(D,P)$ characterizes the extra rate incurred by meeting all objectives in Θ with the encoder fixed. We can also interpret $\Delta(\Theta)$ equivalently as the minimum rate penalty for using Θ-universal representations as opposed to choosing an optimal representation for each objective in Θ. We are particularly interested in the case $\Theta = \Theta(R)$, where $\Theta(R)$ is the set of distortion-perception objectives achievable with dedicated encoders at rate R, i.e., $\Theta(R) \triangleq \{(D,P) : R^{(\infty)}(D,P) \leq R\}$. It will be seen that for the binary case studied in Section III, $\Delta(\Theta(R))$ is negligible compared to R, namely, objective-agnostic encoders/representations can be (almost) as rate-efficient as objective-aware encoders/representations.

B. Two-Stage Coding and Successive Refinement

Let Θ_1 and Θ_2 be two non-empty sets of (D,P) pairs.

Definition 4 (One-Shot Version): A rate pair (R_1,R_2) is said to be *one-shot successively achievable* with respect to (Θ_1,Θ_2) for the source variable X if we can find a random seed U and an encoder pair $(p_{V_1|XU}, p_{V_2|XV_1U})$ with U independent of X, $H(V_1|U) \leq R_1$, and $H(V_2|V_1,U) \leq R_2$ such that for every $(D_1,P_1) \in \Theta_1$ and $(D_2,P_2) \in \Theta_2$, a decoder pair $(p_{\hat{X}_1|V_1U}, p_{\hat{X}_1|V_1V_2U})$ can be constructed to meet the constraints $\mathbb{E}[d(X,\hat{X}_i)] \leq D_i$ and $\omega(p_X,p_{\hat{X}_i}) \leq P_i$, $i=1,2$, where the joint distribution $p_{XV_1V_2\hat{X}_1\hat{X}_2U}$ is assumed to factor as $p_X p_U p_{V_1|XU} p_{V_2|XV_1U} p_{\hat{X}_1|V_1U} p_{\hat{X}_2|V_2U}$. The closure of the set of such (R_1,R_2) is denoted by $\mathcal{R}^*(\Theta_1,\Theta_2)$.

Let $\mathcal{P}_{Z_1Z_2|X}(\Theta_1,\Theta_2)$ denote the set of $p_{Z_1Z_2|X}$ such that for every $(D_1,P_1) \in \Theta_1$ and $(D_2,P_2) \in \Theta_2$, there exists $(p_{\hat{X}_1|Z_1}, p_{\hat{X}_2|Z_2})$ satisfying $\mathbb{E}[d(X,\hat{X}_i)] \leq D_i$ and $\omega(p_X,p_{\hat{X}_i}) \leq P_i$, $i=1,2$, where the joint distribution $p_{XZ_1Z_2\hat{X}_1\hat{X}_2}$ is assumed to factor as $p_X p_{Z_1Z_2|X} p_{\hat{X}_1|Z_1} p_{\hat{X}_2|Z_2}$. Define

$$\begin{aligned} \underline{\mathcal{R}}(\Theta_1,\Theta_2) \triangleq & \bigcup_{p_{Z_1Z_2|X}\in\mathcal{P}_{Z_1Z_2|X}(\Theta_1,\Theta_2)} \{(R_1,R_2)\in\mathbb{R}_+^2 : \\ & R_1 \geq I(X;Z_1) + \log(I(X;Z_1)+1) + 4, \\ & R_1 + R_2 \geq I(X;Z_1,Z_2) + \log(I(X;Z_1)+1) \\ & \quad + \log(I(X;Z_2|Z_1)+1) + 8\}, \\ \overline{\mathcal{R}}(\Theta_1,\Theta_2) \triangleq & \bigcup_{p_{Z_1Z_2|X}\in\mathcal{P}_{Z_1Z_2|X}(\Theta_1,\Theta_2)} \{(R_1,R_2)\in\mathbb{R}_+^2 : \\ & R_1 \geq I(X;Z_1), \end{aligned}$$

$$R_1 + R_2 \geq I(X; Z_1, Z_2)\}.$$

Similarly to Theorem 1, we have the following theorem [8].

Theorem 3: We have $\mathrm{cl}(\underline{\mathcal{R}}(\Theta_1, \Theta_2)) \subseteq \mathcal{R}^*(\Theta_1, \Theta_2) \subseteq \mathrm{cl}(\overline{\mathcal{R}}(\Theta_1, \Theta_2))$.

Definition 5 (Asymptotic Version): A rate pair (R_1, R_2) is said to be *asymptotically successively achievable* with respect to (Θ_1, Θ_2) for the i.i.d. source sequence $\{X(t)\}_{t=1}^{\infty}$ if we can find a random seed U and an encoder pair $(p_{V_1|X^nU}, p_{V_2|X^nV_1U})$ with $\frac{1}{n}H(V_1|U) \leq R_1$ and $\frac{1}{n}H(V_2|V_1, U) \leq R_2$ such that for every $(D_1, P_1) \in \Theta_1$ and $(D_2, P_2) \in \Theta_2$, a decoder pair $(p_{\hat{X}_1^n|V_1U}, p_{\hat{X}_1^n|V_1V_2U})$ can be constructed to meet the constraints $\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[d(X(t), \hat{X}_i(t))] \leq D_i$ and $\omega(p_X, \frac{1}{n}\sum_{t=1}^{n} p_{\hat{X}_i(t)}) \leq P_i$, $i = 1, 2$, where the joint distribution $p_{X^nV_1V_2\hat{X}_1^n\hat{X}_2^nU}$ is assumed to factor as $p_{X^n}p_Up_{V_1|X^nU}p_{V_2|X^nV_1U}p_{\hat{X}_1^n|V_1U}p_{\hat{X}_2^n|V_2U}$. The closure of the set of such (R_1, R_2) is denoted by $\mathcal{R}^{(\infty)}(\Theta_1, \Theta_2)$. We say that *successive refinement* from Θ_1 to Θ_2 is feasible if $(R^{(\infty)}(\Theta_1), R^{(\infty)}(\Theta_2) - R^{(\infty)}(\Theta_1)) \in \mathcal{R}^{(\infty)}(\Theta_1, \Theta_2)$.

The following result [8] is a corollary of Theorem 3.

Theorem 4: We have $\mathcal{R}^{(\infty)}(\Theta_1, \Theta_2) = \mathrm{cl}(\overline{\mathcal{R}}(\Theta_1, \Theta_2))$. Moreover, successive refinement from Θ_1 to Θ_2 is feasible if and only if $(R^{(\infty)}(\Theta_1), R^{(\infty)}(\Theta_2) - R^{(\infty)}(\Theta_1)) \in \mathrm{cl}(\overline{\mathcal{R}}(\Theta_1, \Theta_2))$.

III. Main Results

Throughout this section, we assume $\mathcal{X} = \{0, 1\}$ and $X \sim \mathrm{Bern}(q)$ (i.e., X is a binary source with $p_X(1) = 1 - p_X(0) = q \in (0, \frac{1}{2}]$); moreover, we focus on the Hamming distortion $d(x, \hat{x}) = 1\{x \neq \hat{x}\}$ and the TV distance $w(p_X, p_{\hat{X}}) = \frac{1}{2}\|p_X - p_{\hat{X}}\|_1$.

We first consider the case $\Theta = \{(D, P)\}$ and characterize the one-shot rate-distortion-perception function. Without loss of generality, it is assumed that $P \in [0, q]$.

Theorem 5: For a binary source $X \sim \mathrm{Bern}(q)$, under Hamming distortion and TV perception losses,

$$R^*(D, P) = \begin{cases} \frac{q-D}{q} H_b(q) & 0 \leq D \leq P, \\ \frac{(1-q)+P-\frac{D+P}{2q}}{1-q} H_b(q) & P < D \leq D', \\ 0 & \text{otherwise}, \end{cases}$$

where $H_b(\cdot)$ denotes the binary entropy function and $D' = 2q(1-q) - (1-2q)P$.

For comparison, we present the asymptotic rate-distortion-perception function [3] in the following theorem.

Theorem 6: For a binary source $X \sim \mathrm{Bern}(q)$, under Hamming distortion and TV perception losses,

$$R^{(\infty)}(D, P) = \begin{cases} H_b(q) - H_b(D), & D \in \mathcal{S}_1, \\ 2H_b(q) + H_b(q-P) & \\ \quad -H_t(\frac{D-P}{2}, q) & \\ \quad -H_t(\frac{D+P}{2}, 1-q), & D \in \mathcal{S}_2, \\ 0, & D \in \mathcal{S}_3. \end{cases}$$

where $H_t(\alpha, \beta)$ denotes the entropy of a ternary random variable with probability values $(\alpha, \beta, 1-\alpha-\beta)$. Here, $\mathcal{S}_1 = [0, D_1]$, $\mathcal{S}_2 = [D_1, D_2]$, and $\mathcal{S}_3 = [D_2, \infty)$ with $D_1 = \frac{P}{1-2(q-P)}$ and $D_2 = 2q(1-q) - (1-2q)P$.

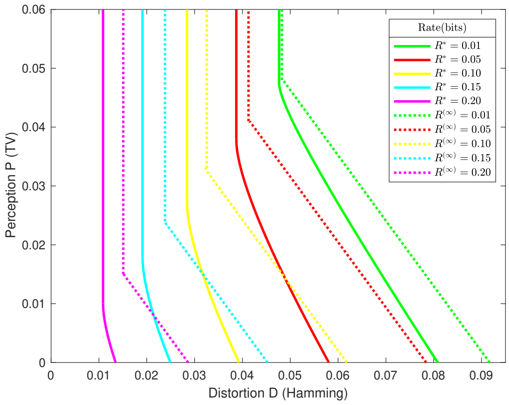


Fig. 1. Plots of perception-distortion curves for different bit rates, where solid curves denote the asymptotic case while dotted curves denote the one-shot case.

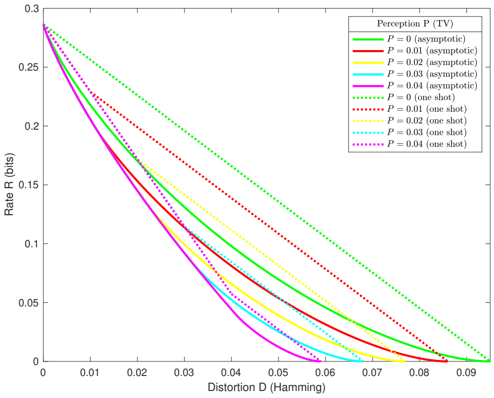


Fig. 2. Plots of rate-distortion curves for different perception qualities, where solid curves denote the asymptotic case while dotted curves denote the one-shot case.

Fig. 1 plots perception-distortion curves for different rates, comparing the asymptotic case and one-shot case under the same bit rate. Note that the trade-off curves for the asymptotic case always lie below their counterparts for the one-shot case. A similar phenomenon can be seen from Fig. 2, which plots rate-distortion curves for different perception qualities.

Let Z be a representation of a binary source $X \sim \mathrm{Bern}(q)$ with $p_Z(i) = q_i$ and $p_{X|Z}(1|i) = \epsilon_i$, $i \in [n]$, where $\sum_{i=1}^{n} q_i = 1$ and $\sum_{i=1}^{n} q_i\epsilon_i = q$. Without loss of generality, we assume that the values of $q_i|1-2\epsilon_i|, i \in [n]$ are in ascending order as i increases. Let j_k with $k \in [m]$ denote the k-th index at which $\epsilon_{j_k} \leq 0.5$. Denote $k^+ \in [m]$ as the first index such that $j_{k^+} > k$ and $\epsilon_{j_{k^+}} \leq 0.5$, if it exists. Define k^* to be the first positive integer satisfying $\frac{\sum_{i=k^*}^{n} q_i(1-\epsilon_i) - \sum_{i=(k^*)^+}^{m} q_{j_i}}{q_k^*} \leq 1 - \epsilon_{k^*}$.

Theorem 7: Let Z be a representation of a binary source $X \sim \mathrm{Bern}(q)$ as specified above. Under Hamming distortion and TV perception losses, the lower boundary of $\Pi(p_{Z|X})$ is piecewise linear with k^* turning points $\{(D_k, P_k)\}_{k=1}^{k^*}$ given by

$$\begin{aligned}
D_k &= \sum_{i=1}^{n} q_i(1-\epsilon_i) + \sum_{i=1}^{k} q_i(2\epsilon_i - 1)(1-\epsilon_i) \\
&\quad + \sum_{i=k^+}^{m} q_{j_i}(2\epsilon_{j_i} - 1), \quad k = 1, \ldots, k^* - 1 \\
P_k &= \left| \sum_{i=k+1}^{n} q_i(1-\epsilon_i) - \sum_{i=k^+}^{m} q_{j_i} \right|, \quad k = 1, \ldots, k^* - 1, \\
D_{k^*} &= \sum_{i=1}^{n} q_i(1-\epsilon_i) + \sum_{i=1}^{k^*-1} q_i(2\epsilon_i - 1)(1-\epsilon_i) \\
&\quad + (2\epsilon_{k^*} - 1)\left(\sum_{i=k^*}^{n} q_i(1-\epsilon_i) - \sum_{i=(k^*)^+}^{m} q_{j_i} \right) \\
&\quad + \sum_{i=(k^*)^+}^{m} q_{j_i}(2\epsilon_{j_i} - 1), \\
P_{k^*} &= 0.
\end{aligned}$$

Next consider the case $\Theta = \Theta(R)$. We start by introducing some quantities which are needed for bounding $R^{(\infty)}(\Theta(R))$. Let $D_1 = D_1(R)$ and $D_2 = D_2(R)$ be respectively the solutions of

$$\begin{aligned}
R &= H_b(q) - H_b(D_1), \\
R &= 3H_b(q) - H_t(\frac{D_2}{2}, q) - H_t(\frac{D_2}{2}, 1-q).
\end{aligned}$$

In fact, D_1 and D_2 correspond to the D_1 and D_2 in Theorem 6, but here expressed in terms of R, rather than in terms of P. Define

$$\begin{aligned}
R_{LB} = (1-q) \sum_{i,j \in \{0,1\}} p_{ij|0} \log \frac{p_{ij|0}}{(1-q)p_{ij|0} + qp_{ij|1}} \\
+ q \sum_{i,j \in \{0,1\}} p_{ij|1} \log \frac{p_{ij|1}}{(1-q)p_{ij|0} + qp_{ij|1}},
\end{aligned}$$

where

$$\begin{aligned}
p_{00|0} &= 1 - \frac{D_2}{2(1-q)}, \\
p_{01|0} &= \frac{(D_2 - D_1)(2q - 2D_1 + D_2)}{2(1-q)(q - 2D_1 + D_2)}, \\
p_{10|0} &= 0, \\
p_{11|0} &= \frac{(2D_1 - D_2)(q - D_1)}{2(1-q)(q - 2D_1 + D_2)}, \\
p_{00|1} &= \frac{D_2}{2q}, \\
p_{01|1} &= \frac{(D_2 - D_1)(2D_1 - D_2)}{2q(q - 2D_1 + D_2)}, \\
p_{10|1} &= 0, \\
p_{11|1} &= \frac{(q - D_1)(2q - 2D_1 + D_2)}{2q(q - 2D_1 + D_2)}.
\end{aligned} \tag{6}$$

Moreover, define

$$\begin{aligned}
R_{UB} = (1-q) \sum_{i,j \in \{0,1\}} p'_{ij|0} \log \frac{p'_{ij|0}}{(1-q)p'_{ij|0} + qp'_{ij|1}} \\
+ q \sum_{i,j \in \{0,1\}} p'_{ij|1} \log \frac{p'_{ij|1}}{(1-q)p'_{ij|0} + qp'_{ij|1}},
\end{aligned}$$

where

$$\begin{aligned}
p'_{00|0} &= 1 - \frac{D_2}{2(1-q)}, \\
p'_{01|0} &= \frac{D_2 - D_1 + P_{UB}}{2(1-q)}, \\
p'_{10|0} &= 0, \\
p'_{11|0} &= \frac{D_1 - P_{UB}}{2(1-q)}, \\
p'_{00|1} &= \frac{D_2}{2q}, \\
p'_{01|1} &= \frac{D_1 - D_2 + P_{UB}}{2q}, \\
p'_{10|1} &= 0, \\
p'_{11|1} &= \frac{2q - D_1 - P_{UB}}{2q},
\end{aligned} \tag{7}$$

and

$$P_{UB} = \kappa(D_1 - D_2), \tag{8}$$

$$\kappa = \kappa(D_2) \triangleq \frac{-\log \frac{D_2}{2} + \frac{1}{2}\log(1 - q - \frac{D_2}{2}) + \frac{1}{2}\log(q - \frac{D_2}{2})}{\log \frac{q}{1-q} + \frac{1}{2}\log(1 - q - \frac{D_2}{2}) - \frac{1}{2}\log(q - \frac{D_2}{2})}. \tag{9}$$

Theorem 8: For a binary source $X \sim \text{Bern}(q)$, under Hamming distortion and TV perception losses,

$$R_{LB} \leq R^{(\infty)}(\Theta(R)) \leq R_{UB}$$

and consequently

$$R_{LB} - R \leq \Delta(\Theta(R)) \leq R_{UB} - R.$$

Moreover, the upper and lower bounds coincide if and only if

$$\frac{q}{2D_1 - D_2 - q} \geq \kappa. \tag{10}$$

Fig. 3 shows when $R \gtrapprox 0.08$, the dashed lines coincide with the dotted lines, implying that the upper bound meets the lower bound. Fig. 4 provides a direct visualization in the rate domain, that is when $R \gtrapprox 0.08$, the two bounds match.

We next proceed to study successive refinement. Let $\mathcal{A}$ be the regime where both the distortion and perception constraints are active, i.e., $\mathcal{A} \triangleq \{(D, P) : D \in [\frac{P}{1-2(q-P)}, 2q(1-q) - (1-2q)P), P \in [0, q]\}$.

Theorem 9: For a binary source $X \sim \text{Bern}(q)$, under Hamming distortion and TV perception losses, successive refinement from $(D_1, P_1) \in \mathcal{A}$ to $(D_2, P_2) \in \mathcal{A}$ is feasible if and only if

$$q((D_1 - P_1) - (D_2 - P_2)) \geq D_1P_2 - D_2P_1, \tag{11}$$

$$(1-q)((D_1 + P_1) - (D_2 + P_2)) \geq D_2P_1 - D_1P_2. \tag{12}$$

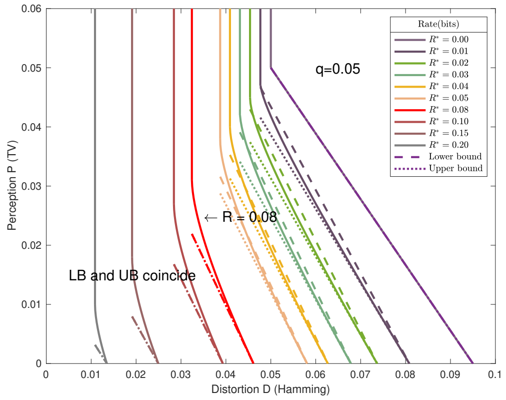


Fig. 3. Plots of perception-distortion curves for different bit rates under the lower bound and upper bound, where $q = 0.05$. When $R \gtrapprox 0.08$, the upper bound and lower bound coincide.

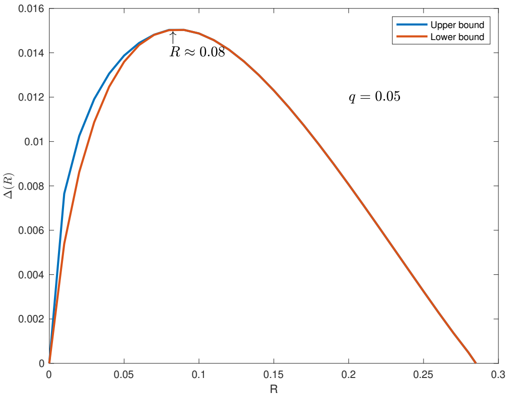


Fig. 4. Plots of $\Delta(R)$ with R for both lower bound and upper bound, where $q = 0.05$. When $R \gtrapprox 0.08$, the upper bound and lower bound coincide.

For $0 < R_1 < R_2$, we denote $D_1 = D_1(R_1), D_2 = D_2(R_1)$, and $D_1' = D_1(R_2), D_2' = D_2(R_2)$, where the functions $D_1(R), D_2(R)$ are defined above Theorem 8.

Theorem 10: Let $0 < R_1 < R_2$. For a binary source $X \sim \text{Bern}(q)$, under Hamming distortion and TV perception losses as well as the conditions that

$$\frac{q}{2D_1 - D_2 - q} \geq \kappa(D_2), \quad \frac{q}{2D_1' - D_2' - q} \geq \kappa(D_2') \quad (13)$$

with the function κ given in (9), successive refinement from $\Theta(R_1)$ to $\Theta(R_2)$ is feasible if

$$2D_1 - D_2 \leq 2D_1' - D_2'.$$

Remark 1: Note that the conditions in (13) are the necessary and sufficient conditions for the bounds on $R^{(\infty)}(\Theta(R))$ in Theorem 8 to match at $R = R_1$ and $R = R_2$ respectively. Hence, the theorem above provides a sufficient condition for the feasibility of set-wise successive refinement, given that R_1 and R_2 are above the matching threshold.

Remark 2: In fact, we numerically verify that $2D_1 - D_2 \leq 2D_1' - D_2'$ automatically holds once the conditions in (13) are satisfied. In other words, successive refinement from $\Theta(R_1)$ to $\Theta(R_2)$ is feasible if R_1 and R_2 are above the matching threshold.

Acknowledgment

The authors would like to thank Lei Yu for addressing a technical issue in the proof of Theorem 8 and for helping prepare a draft of this paper.

References

[1] S. Santurkar, D. Budden, and N. Shavit, “Generative compression," in Proceedings of the Picture Coding Sympo- sium, 2018.

[2] Y. Blau and T. Michaeli, “The perception-distortion tradeoff," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6228-6237, 2018.

[3] Y. Blau and T. Michaeli, “Rethinking lossy compression: The rate-distortion-perception tradeoff," in International Conference on Machine Learning, pp. 675–685, 2019.

[4] M. Tschannen, E. Agustsson, and Lucic M, “Deep generative models for distribution-preserving lossy compression," in Proceedings of the Conference on Neural Information Processing Systems, 2018.

[5] R. Zamir and K. Rose, “Natural type selection in adaptive lossy compression," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 99—111, Jan. 2001.

[6] N. Saldi, T. Linder, and S. Yüksel, “Randomized quantization and source coding with constrained output distribution," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 91–106, Jan. 2015.

[7] N. Saldi, T. Linder, and S. Yüksel, “Output constrained lossy source coding with limited common randomness," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4984–4998, Sep. 2015.

[8] G. Zhang, J. Qian, J. Chen, and A. Khisti, “Universal rate-distortion-perception representations for lossy compression," 2021. arXiv:2106.10311. [Online]. Availble: https://arxiv.org/abs/2106.10311

[9] L. Theis and A. B. Wagner, “A coding theorem for the rate-distortion-perception function," *ICLR 2021 neural compression workshop*.

[10] W. H. R. Equitz and T. M. Cover, “Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269—275, Mar. 1991.

Reliability-Based Message Passing Decoding of Non-binary Low-Density Parity-Check Codes

Emna Ben Yacoub
Institute for Communications Engineering
Technical University of Munich, Germany
Email: emna.ben-yacoub@tum.de

***Abstract*—A message passing decoding algorithm for q-ary low-density parity-check codes over the q-ary symmetric channel is introduced. The exchanged messages are symbols from $\mathbb{F}_q$ together with their reliability scores. A density evolution analysis for irregular low-density parity-check code ensembles is developed and thresholds for selected ensembles are computed showing gains with respect to other algorithms in the literature. Finite-length simulation results confirm the asymptotic analysis.**

I. Introduction

The design of efficient low-complexity low-density parity-check (LDPC) decoding algorithms for high-throughput applications is receiving considerable interest. In his thesis [1], Gallager presented low-complexity decoding algorithms where the exchanged messages are binary. In [2], it was shown that significant gains can be achieved by allowing erasures in the decoding. Another class of decoding algorithms for binary LDPC codes has been studied in [3]–[6], where the variable nodes (VNs) exploit soft information from the channel and the exchanged messages are coarsely quantized.

Due to their high decoding complexity, several works considered reduced-complexity decoding algorithms for non-binary LDPC codes over the binary-input additive white Gaussian noise (biAWGN) channel [7]–[11] and the q-ary symmetric channel (QSC) [12]–[17]. Majority logic based algorithms were considered in [18]–[21].

In [16], a decoding algorithm referred to as symbol message passing (SMP) for non-binary LDPC codes over the QSC was introduced, where the exchanged messages are symbols from $\mathbb{F}_q$. The SMP decoder was extended in [22] to the scaled reliability list message passing (SRLMP) decoder. The exchanged messages in SRLMP are sets of symbols from $\mathbb{F}_q$. As shown in [22], the SRLMP decoder outperforms the algorithm presented in [17], which has the same message alphabet and a comparable decoding complexity. The gain is due to the VN update rule, where the incoming check node (CN) messages and the channel observations are converted to log-likelihood vectors. The performance of SRLMP improves by increasing the list size from 1 to 2 but the decoder data flow increases as well, especially for large field orders q.

In this work, we introduce a message passing algorithm for q-ary LDPC codes over the QSC, which we dub reliability-based symbol message passing (RSMP). We follow the approach in [16], [22] and convert the channel and the incoming CN messages to log-likelihood vectors at the VNs. This approach is based on modeling the extrinsic channel as a discrete memoryless channel (DMC) whose transition probabilities may be estimated via density evolution (DE) as proposed in [3]. To decrease the data flow, instead of passing a list of symbols as in SRLMP, the exchanged messages are symbols from $\mathbb{F}_q$ together with their reliability scores from $\{\mathtt{H},\mathtt{L}\}$. We improve the performance of SMP by including reliability scores in the decoding. The VN operation is similar to a voting system, where the channel and the neighboring CNs vote for the value of the code symbol. The votes have different weights. The weight of the channel observation depends on the error probability of the QSC. The weights of the incoming CN messages depend on the transition probabilities of the extrinsic DMC which change in each iteration and can be estimated via DE analysis. The VN selects then the element with the highest score. The VN update rule is different than in majority logic algorithms in [18]–[21]. For instance in [19], the VN uses the voting result form the previous iteration, the weights of the CN votes are kept constant over the iterations and no reliability scores are used.

II. Preliminaries

A. Q-ary Symmetric Channel

Consider a QSC with error probability ϵ, input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, with $\mathcal{X} = \mathcal{Y} = \{0, \alpha^0, \dots, \alpha^{q-2}\}$, where α is a primitive element of $\mathbb{F}_q$. Denote by $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ the channel input and channel output, respectively. The transition probabilities of a QSC with error probability ϵ are

$$P(y|x) = \begin{cases} 1-\epsilon & \text{if } y = x \\ \frac{\epsilon}{q-1} & \text{otherwise.} \end{cases} \tag{1}$$

The capacity of the QSC, in symbols per channel use, is

$$C = 1 + \epsilon \log_q \frac{\epsilon}{q-1} + (1-\epsilon)\log_q(1-\epsilon).$$

B. Log-Likelihood Vector

For a given channel output y of a DMC with input alphabet $\mathcal{X} = \mathbb{F}_q$, we introduce the normalized log-likelihood vector, also referred to as $\boldsymbol{L}$-vector,

$$\boldsymbol{L}(y) = [L_0(y), L_1(y), \dots, L_{\alpha^{q-2}}(y)] \tag{2}$$

whose elements are defined as

$$L_u(y) = \log P(y|u) \quad \forall u \in \mathbb{F}_q. \tag{3}$$

The $\boldsymbol{L}$-vector will be instrumental to the design of a message passing decoding algorithm for non-binary LDPC codes over the QSC. We consider a decoding algorithm where the exchanged messages are symbols from $\mathbb{F}_q$ together with their reliability scores from $\{\mathtt{H}, \mathtt{L}\}$. A message sent from a CN to a VN can be modeled as the observation of the random variable (RV) X after transmission over a q-ary input $2q$-ary output discrete memoryless extrinsic channel [23, Fig. 3]. The VNs convert the channel and the incoming CN messages to $\boldsymbol{L}$-vectors. The transition probabilities of the communication channel are given in (1) and the transition probabilities of the extrinsic channel can be estimated via DE analysis, as suggested in [3].

C. Non-binary LDPC Codes

Non-binary LDPC codes are characterized by an $m \times n$ sparse parity-check matrix $\boldsymbol{H}$ with coefficients in $\mathbb{F}_q$. The parity-check matrix can be represented by a Tanner graph with n VNs corresponding to codeword symbols and m CNs corresponding to parity checks. The edge label associated to the edge connecting the VN $\mathtt{v}$ and the CN $\mathtt{c}$ is denoted by $h_{\mathtt{v},\mathtt{c}}$, with $h_{\mathtt{v},\mathtt{c}} \in \mathbb{F}_q \setminus \{0\}$. The sets $\mathcal{N}(\mathtt{v})$ and $\mathcal{N}(\mathtt{c})$ denote the neighbors of VN $\mathtt{v}$ and CN $\mathtt{c}$, respectively. The degree of a VN $\mathtt{v}$ is the cardinality of the set $\mathcal{N}(\mathtt{v})$. Similarly, the degree of a CN $\mathtt{c}$ is the cardinality of the set $\mathcal{N}(\mathtt{c})$. The edge-oriented degree distribution polynomials of an LDPC code graph are given by $\lambda(x) = \sum_i \lambda_i x^{i-1}$ and $\rho(x) = \sum_i \rho_i x^{i-1}$ where λ_i and ρ_i correspond to the fraction of edges incident to VNs and CNs with degree i, respectively. An unstructured irregular LDPC code ensemble $\mathscr{C}_{q,n}^{\lambda,\rho}$ is the set of all q-ary LDPC codes with block length n, degree distributions $\lambda(x)$ and $\rho(x)$ and edge labels uniformly chosen in $\mathbb{F}_q \setminus \{0\}$.

III. RSMP Decoding

This section introduces the RSMP decoder. An exchanged message between a check and a variable node is a symbol from $\mathbb{F}_q$ together with its reliability score from $\{\mathtt{H}, \mathtt{L}\}$, where $\mathtt{H}$ and $\mathtt{L}$ correspond to symbols with high and low reliability, respectively. We denote by $(m^{(\ell)}_{\mathtt{c}\to\mathtt{v}}, r^{(\ell)}_{\mathtt{c}\to\mathtt{v}})$ the message sent from CN $\mathtt{c}$ to its neighboring VN $\mathtt{v}$. Similarly, $(m^{(\ell)}_{\mathtt{v}\to\mathtt{c}}, r^{(\ell)}_{\mathtt{v}\to\mathtt{c}})$ is the message sent from VN $\mathtt{v}$ to CN $\mathtt{c}$ at the ℓ-th iteration. We have $m^{(\ell)}_{\mathtt{c}\to\mathtt{v}}, m^{(\ell)}_{\mathtt{v}\to\mathtt{c}} \in \mathbb{F}_q$ and $r^{(\ell)}_{\mathtt{c}\to\mathtt{v}}, r^{(\ell)}_{\mathtt{v}\to\mathtt{c}} \in \{\mathtt{H}, \mathtt{L}\}$.

Each VN sends its channel observation y to its neighboring CNs

$$m^{(0)}_{\mathtt{v}\to\mathtt{c}} = y.$$

The reliability score of $m^{(0)}_{\mathtt{v}\to\mathtt{c}}$ is

$$r^{(0)}_{\mathtt{v}\to\mathtt{c}} = \begin{cases} \mathtt{H} & \text{if } \mathsf{D}_{\text{ch}} > \Delta \\ \mathtt{L} & \text{otherwise} \end{cases}$$

where

$$\mathsf{D}_{\text{ch}} = \log(1-\epsilon) - \log\left(\frac{\epsilon}{q-1}\right). \tag{8}$$

The real-valued parameter Δ is chosen to maximize the iterative decoding threshold and can be chosen for each iteration individually. In this work, we keep Δ constant over the iterations, i.e., we compute the iterative decoding thresholds for several values of Δ and choose the best one.

Consider a CN $\mathtt{c}$ and a VN $\mathtt{v}$ connected to it. The CN $\mathtt{c}$ computes the symbol that satisfies the parity check equation given the incoming VN messages. We assign to the outgoing symbol from $\mathtt{c}$ the reliability score $\mathtt{L}$ if any incoming symbols from the other neighboring VNs has low reliability and the reliability score $\mathtt{H}$ otherwise. Formally, the outgoing message is $(m^{(\ell)}_{\mathtt{c}\to\mathtt{v}}, r^{(\ell)}_{\mathtt{c}\to\mathtt{v}})$ with

$$m^{(\ell)}_{\mathtt{c}\to\mathtt{v}} = -\,h^{-1}_{\mathtt{v},\mathtt{c}} \sum_{\mathtt{v}'\in\mathcal{N}(\mathtt{c})\setminus\mathtt{v}} h_{\mathtt{v}',\mathtt{c}} m^{(\ell-1)}_{\mathtt{v}'\to\mathtt{c}} \tag{9}$$

and the reliability score of $m^{(\ell)}_{\mathtt{c}\to\mathtt{v}}$ is

$$r^{(\ell)}_{\mathtt{c}\to\mathtt{v}} = \begin{cases} \mathtt{H} & \text{if } r^{(\ell-1)}_{\mathtt{v}'\to\mathtt{c}} = \mathtt{H} \quad \forall \mathtt{v}' \in \mathcal{N}(\mathtt{c}) \setminus \mathtt{v} \\ \mathtt{L} & \text{otherwise.} \end{cases}$$

The multiplication and the sum in (9) are performed over $\mathbb{F}_q$ and $h^{-1}_{\mathtt{v},\mathtt{c}}$ is the inverse of $h_{\mathtt{v},\mathtt{c}}$ in $\mathbb{F}_q$.

At the ℓ-th iteration, each VN computes

$$\begin{aligned} \boldsymbol{L}^{(\ell)}_{\text{ex}} &= \left[L^{(\ell)}_{\text{ex},0}, L^{(\ell)}_{\text{ex},1}, \ldots, L^{(\ell)}_{\text{ex},\alpha^{q-2}}\right] \\ &= \boldsymbol{L}(y) + \sum_{\mathtt{c}'\in\mathcal{N}(\mathtt{v})\setminus\mathtt{c}} \boldsymbol{L}\left((m^{(\ell)}_{\mathtt{c}'\to\mathtt{v}}, r^{(\ell)}_{\mathtt{c}'\to\mathtt{v}})\right). \end{aligned} \tag{10}$$

Then, the VN determines the $\mathbb{F}_q$ symbol with the maximum entry in $\boldsymbol{L}^{(\ell)}_{\text{ex}}$. The outgoing symbol has high reliability if its corresponding entry in $\boldsymbol{L}^{(\ell)}_{\text{ex}}$ is greater by Δ than each of the other entries. Formally, the VN sends $(m^{(\ell)}_{\mathtt{v}\to\mathtt{c}}, r^{(\ell)}_{\mathtt{v}\to\mathtt{c}})$ with

$$m^{(\ell)}_{\mathtt{v}\to\mathtt{c}} = \arg\max_{u\in\mathbb{F}_q} L^{(\ell)}_{\text{ex},u} \tag{11}$$

and the reliability score of $m^{(\ell)}_{\mathtt{v}\to\mathtt{c}}$ is

$$r^{(\ell)}_{\mathtt{v}\to\mathtt{c}} = \begin{cases} \mathtt{H} & \text{if } \exists a \in \mathbb{F}_q : L^{(\ell)}_{\text{ex},a} > L^{(\ell)}_{\text{ex},u} + \Delta \quad \forall u \in \mathbb{F}_q \setminus \{a\} \\ \mathtt{L} & \text{otherwise.} \end{cases}$$

In (11), if multiple maximizing arguments exist the $\arg\max$ function outputs one of them uniformly at random.

In (10), the $\boldsymbol{L}$-vector $\boldsymbol{L}(y)$ corresponding to the QSC channel observation is obtained from (1) and (3). Moreover, we model each CN to VN message as an observation of the symbol X (associated to $\mathtt{v}$) at the output of an *extrinsic channel* with input alphabet $\mathcal{X} = \mathbb{F}_q$ and output alphabet $\mathcal{Z} = \mathbb{F}_q \times \{\mathtt{H}, \mathtt{L}\}$. The transition probabilities of the extrinsic channel are in general unknown. It was shown in [3], [4] that, for moderate to large block lengths, these probabilities can be accurately estimated via the DE presented in Section IV. They are then used to compute the $\boldsymbol{L}$-vectors of the CN messages in (2) and (3).

To estimate its codeword symbol each VN computes

$$\begin{aligned} \boldsymbol{L}^{(\ell)}_{\text{app}} &= \left[L^{(\ell)}_{\text{app},0}, L^{(\ell)}_{\text{app},1}, \ldots, L^{(\ell)}_{\text{app},\alpha^{q-2}}\right] \\ &= \boldsymbol{L}(y) + \sum_{\mathtt{c}'\in\mathcal{N}(\mathtt{v})} \boldsymbol{L}\left((m^{(\ell)}_{\mathtt{c}'\to\mathtt{v}}, r^{(\ell)}_{\mathtt{c}'\to\mathtt{v}})\right). \end{aligned}$$

$$s_{\mathcal{I}_0}^{(\ell)} = \frac{1}{q}\left[\rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_1}^{(\ell-1)}\right) + (q-1)\rho\left(p_{\mathcal{I}_0}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)}}{q-1}\right)\right] \tag{4}$$

$$s_{\mathcal{I}_1}^{(\ell)} = \frac{q-1}{q}\left[\rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_1}^{(\ell-1)}\right) - \rho\left(p_{\mathcal{I}_0}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)}}{q-1}\right)\right] \tag{5}$$

$$s_{\mathcal{I}_2}^{(\ell)} = \frac{1}{q}\left[1 - \rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_1}^{(\ell-1)}\right) - (q-1)\rho\left(p_{\mathcal{I}_0}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)}}{q-1}\right) + (q-1)\rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_2}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)} + p_{\mathcal{I}_3}^{(\ell-1)}}{q-1}\right)\right] \tag{6}$$

$$s_{\mathcal{I}_3}^{(\ell)} = \frac{q-1}{q}\left[1 - \rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_1}^{(\ell-1)}\right) + \rho\left(p_{\mathcal{I}_0}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)}}{q-1}\right) - \rho\left(p_{\mathcal{I}_0}^{(\ell-1)} + p_{\mathcal{I}_2}^{(\ell-1)} - \frac{p_{\mathcal{I}_1}^{(\ell-1)} + p_{\mathcal{I}_3}^{(\ell-1)}}{q-1}\right)\right] \tag{7}$$

The final decision is

$$\hat{x}^{(\ell)} = \arg\max_{u \in \mathbb{F}_q} L_{\mathsf{app},u}^{(\ell)}.$$

Note that we can easily include erasures in the decoding algorithm. We observed that both decoding algorithms (with and without erasures) have similar performance.

Remark 1. *The complexity of a message passing decoding algorithm can be studied from 2 perspectives: the cost of the arithmetic operations and the decoder data flow. The internal decoder data flow, defined as the number of bits that are processed in each iteration, scales linearly in the number of bits that represent the exchanged CN and VN messages [24]. This work targets this second complexity, i.e., the reduction of the internal data flow. The exchanged messages in belief propagation (BP) decoder are $(q-1)$-ary real valued vectors, whereas for RSMP the exchanged messages are symbols from $\mathbb{F}_q$ together with a reliability score from $\{\mathtt{H}, \mathtt{L}\}$. This approach substantially reduces the number of bits needed to represent the exchanged CN and VN messages and therefore the decoder data flow.*

IV. Density Evolution Analysis for RSMP

This section provides a DE analysis for RSMP for non-binary irregular LDPC code ensembles. In the DE, the probabilities of VN to CN and CN to VN messages are tracked as iterations progress. Due to symmetry and under the all-zero codeword assumption, we can partition $\mathbb{F}_q \times \{\mathtt{H}, \mathtt{L}\}$ into the following 4 disjoint sets

$$\begin{aligned}
\mathcal{I}_0 =& \{(0, \mathtt{H})\}\\
\mathcal{I}_1 =& \{(a, \mathtt{H}) : a \in \mathbb{F}_q \setminus \{0\}\}\\
\mathcal{I}_2 =& \{(0, \mathtt{L})\}\\
\mathcal{I}_3 =& \{(a, \mathtt{L}) : a \in \mathbb{F}_q \setminus \{0\}\}
\end{aligned}$$

where $(u, \mathtt{H})$ denotes a high reliable symbol u and $(u, \mathtt{L})$ denotes a low reliable symbol $u \in \mathbb{F}_q$. Note that $|\mathcal{I}_0| = |\mathcal{I}_2| = 1$, $|\mathcal{I}_1| = |\mathcal{I}_3| = q - 1$.
Let $p_{\mathcal{I}_k}^{(\ell)}$ be the probability that a VN to CN message belongs to the set $\mathcal{I}_k$ at the ℓ-th iteration. That means a VN to CN symbol takes the value $a \in \mathbb{F}_q$ and has the reliability score $r \in \{\mathtt{H}, \mathtt{L}\}$ with probability $p_{\mathcal{I}_k}^{(\ell)}/|\mathcal{I}_k|$ if $(a, r) \in \mathcal{I}_k$. Similarly $s_{\mathcal{I}_k}^{(\ell)}$ is the probability that a CN to VN message belongs to the set $\mathcal{I}_k$, where $k \in \{0, 1, 2, 3\}$.

Initially, we have

$$\begin{aligned}
p_{\mathcal{I}_0}^{(0)} =& \mathbb{I}(\mathsf{D}_{\mathsf{ch}} > \Delta)(1 - \epsilon)\\
p_{\mathcal{I}_1}^{(0)} =& \mathbb{I}(\mathsf{D}_{\mathsf{ch}} > \Delta)\epsilon\\
p_{\mathcal{I}_2}^{(0)} =& \mathbb{I}(\mathsf{D}_{\mathsf{ch}} \leq \Delta)(1 - \epsilon)\\
p_{\mathcal{I}_3}^{(0)} =& \mathbb{I}(\mathsf{D}_{\mathsf{ch}} \leq \Delta)\epsilon
\end{aligned}$$

where $\mathbb{I}(\mathcal{A})$ is an indicator function that takes the value 1 if the proposition $\mathcal{A}$ is true and 0 otherwise.

For the CN to VN messages, we have $s_{\mathcal{I}_0}^{(\ell)}$,$s_{\mathcal{I}_1}^{(\ell)}$, $s_{\mathcal{I}_2}^{(\ell)}$ and $s_{\mathcal{I}_3}^{(\ell)}$ are given in (4), (5), (6) and (7), respectively. The extrinsic channel has input alphabet $\mathcal{X} = \mathbb{F}_q$, output alphabet $\mathcal{Z} = \mathbb{F}_q \times \{\mathtt{H}, \mathtt{L}\}$ and transition probabilities

$$P(z|u) = \begin{cases} s_{\mathcal{I}_0}^{(\ell)} & \text{if } z = (u, \mathtt{H}) \\ \frac{s_{\mathcal{I}_1}^{(\ell)}}{q-1} & \text{if } z = (e, \mathtt{H}) \quad e \in \mathbb{F}_q \setminus \{u\} \\ s_{\mathcal{I}_2}^{(\ell)} & \text{if } z = (u, \mathtt{L}) \\ \frac{s_{\mathcal{I}_3}^{(\ell)}}{q-1} & \text{if } z = (e, \mathtt{L}) \quad e \in \mathbb{F}_q \setminus \{u\}. \end{cases} \tag{12}$$

Consider now the VN to CN messages. Define the random vector $\boldsymbol{F}^{(\ell)} = \left(F_{(0,\mathtt{H})}^{(\ell)}, \ldots, F_{(\alpha^{q-2},\mathtt{H})}^{(\ell)}, F_{(0,\mathtt{L})}^{(\ell)}, \ldots, F_{(\alpha^{q-2},\mathtt{L})}^{(\ell)}\right)$ where $F_{(u,r)}^{(\ell)}$, for $u \in \mathbb{F}_q$ and $r \in \{\mathtt{H}, \mathtt{L}\}$, denotes the RV associated to the number of incoming CN messages to a degree d VN that are equal to (u, r) at the ℓ-th iteration. Let $\boldsymbol{f}^{(\ell)}$ be the realization of $\boldsymbol{F}^{(\ell)}$. The entries of $\boldsymbol{L}\left((m_{\mathsf{c}'\to\mathsf{v}}^{(\ell)}, r_{\mathsf{c}'\to\mathsf{v}}^{(\ell)})\right)$ in (10) are given by

$$L_u\left((m_{\mathsf{c}'\to\mathsf{v}}^{(\ell)}, r_{\mathsf{c}'\to\mathsf{v}}^{(\ell)})\right) = \log\left(P((m_{\mathsf{c}'\to\mathsf{v}}^{(\ell)}, r_{\mathsf{c}'\to\mathsf{v}}^{(\ell)})|u)\right)$$

where $m_{\mathsf{c}'\to\mathsf{v}}^{(\ell)} \in \mathbb{F}_q, r_{\mathsf{c}'\to\mathsf{v}}^{(\ell)} \in \{\mathtt{H}, \mathtt{L}\}, u \in \mathbb{F}_q$ and $P(z|u)$ can be computed from (4), (5), (6), (7) and (12) $\forall z \in \mathbb{F}_q \times \{\mathtt{H}, \mathtt{L}\}$. Hence, the elements $L_{\mathsf{ex},u}^{(\ell)}$ of the aggregated extrinsic $\boldsymbol{L}$-vector in (10) are related to $f_u^{(\ell)}$ and the channel observation y by

$$L_{\mathsf{ex},u}^{(\ell)} = \mathsf{D}_{\mathtt{H}}^{(\ell)} f_{(u,\mathtt{H})}^{(\ell)} + \mathsf{D}_{\mathtt{L}}^{(\ell)} f_{(u,\mathtt{L})}^{(\ell)} + \mathsf{D}_{\mathsf{ch}}\delta_{uy} + K \quad \forall u \in \mathbb{F}_q$$

where δ_{ij} is the Kronecker delta function, D_{ch} is given in (8) and we have

$$\mathsf{D}_{\mathtt{H}}^{(\ell)} = \log(s_{\mathcal{I}_0}^{(\ell)}) - \log\left(\frac{s_{\mathcal{I}_1}^{(\ell)}}{q-1}\right)$$

$$
\begin{aligned}
\mathsf{D}_{\mathsf{L}}^{(\ell)} &= \log(s_{\mathcal{I}_2}^{(\ell)}) - \log\left(\frac{s_{\mathcal{I}_3}^{(\ell)}}{q-1}\right)\\
K &= \log\left(\frac{\epsilon}{q-1}\right) + \sum_{a\in\mathbb{F}_q} f_{(a,\mathrm{H})}^{(\ell)} \log\left(\frac{s_{\mathcal{I}_1}^{(\ell)}}{q-1}\right)\\
&\quad + \sum_{a\in\mathbb{F}_q} f_{(a,\mathrm{L})}^{(\ell)} \log\left(\frac{s_{\mathcal{I}_3}^{(\ell)}}{q-1}\right).
\end{aligned}
\tag{13}
$$

Note that K in (13) can be ignored in the VN update rule since it is independent of the symbol u. We obtain

$$
\begin{aligned}
p_{\mathcal{I}_0}^{(\ell)} &= \sum_d \lambda_d \sum_{y\in\mathbb{F}_q} \Pr\{Y=y\} \sum_{\boldsymbol{f}^{(\ell)}} \Pr\{\boldsymbol{F}^{(\ell)} = \boldsymbol{f}^{(\ell)}\}\times\\
&\qquad \prod_{u\in\mathbb{F}_q\setminus\{0\}} \mathbb{I}(L_{\mathsf{ex},0}^{(\ell)} > L_{\mathsf{ex},u}^{(\ell)} + \Delta)\\
p_{\mathcal{I}_1}^{(\ell)} &= \sum_d \lambda_d \sum_{a\in\mathbb{F}_q\setminus\{0\}} \sum_{y\in\mathbb{F}_q} \Pr\{Y=y\}\times\\
&\qquad \sum_{\boldsymbol{f}^{(\ell)}} \Pr\{\boldsymbol{F}^{(\ell)} = \boldsymbol{f}^{(\ell)}\} \prod_{u\in\mathbb{F}_q\setminus\{a\}} \mathbb{I}(L_{\mathsf{ex},a}^{(\ell)} > L_{\mathsf{ex},u}^{(\ell)} + \Delta)\\
p_{\mathcal{I}_2}^{(\ell)} &= \sum_d \lambda_d \sum_{y\in\mathbb{F}_q} \Pr\{Y=y\} \sum_{\boldsymbol{f}^{(\ell)}} \Pr\{\boldsymbol{F}^{(\ell)} = \boldsymbol{f}^{(\ell)}\}\times\\
&\qquad \left[\mathbb{I}(\mathcal{S}_0 \neq \emptyset) \prod_{u\in\mathbb{F}_q\setminus\{0\}} \mathbb{I}(L_{\mathsf{ex},0}^{(\ell)} > L_{\mathsf{ex},u}^{(\ell)}) + \frac{\mathbb{I}(0\in\mathcal{U})}{|\mathcal{U}|}\right]\\
p_{\mathcal{I}_3}^{(\ell)} &= \sum_d \lambda_d \sum_{a\in\mathbb{F}_q\setminus\{0\}} \sum_{y\in\mathbb{F}_q} \Pr\{Y=y\}\times\\
&\qquad \sum_{\boldsymbol{f}^{(\ell)}} \Pr\{\boldsymbol{F}^{(\ell)} = \boldsymbol{f}^{(\ell)}\}\times\\
&\qquad \left[\mathbb{I}(\mathcal{S}_a \neq \emptyset) \prod_{u\in\mathbb{F}_q\setminus\{a\}} \mathbb{I}(L_{\mathsf{ex},a}^{(\ell)} > L_{\mathsf{ex},u}^{(\ell)}) + \frac{\mathbb{I}(a\in\mathcal{U})}{|\mathcal{U}|}\right]
\end{aligned}
$$

where the inner sum is over all length $2q$ integer vectors $\boldsymbol{f}^{(\ell)}$ whose entries are non-negative and sum to $d-1$. For all $u \in \mathbb{F}_q$, we have

$$
\begin{aligned}
\mathcal{S}_u &= \{e \in \mathbb{F}_q : L_{\mathsf{ex},u}^{(\ell)} - \Delta \leq L_{\mathsf{ex},e}^{(\ell)} < L_{\mathsf{ex},u}^{(\ell)}\}\\
\mathcal{U} &= \{e \in \mathbb{F}_q : L_{\mathsf{ex},e}^{(\ell)} = \max_{u\in\mathbb{F}_q} L_{\mathsf{ex},u}^{(\ell)}\}
\end{aligned}
$$

$$
\begin{aligned}
\Pr\{\boldsymbol{F}^{(\ell)} = \boldsymbol{f}^{(\ell)}\} &= \binom{d-1}{f_{(0,\mathrm{H})}^{(\ell)}, \ldots, f_{(\alpha^{q-2},\mathrm{L})}^{(\ell)}} \prod_{k=0}^{3} \left(\frac{s_{\mathcal{I}_k}^{(\ell)}}{|\mathcal{I}_k|}\right)^{f_{\mathcal{I}_k}^{(\ell)}}\\
f_{\mathcal{I}_k}^{(\ell)} &= \sum_{(a,r)\in\mathcal{I}_k} f_{(a,r)}^{(\ell)} \quad \forall k \in \{0,\ldots,3\}.
\end{aligned}
$$

The iterative decoding threshold $\epsilon^{\star}$ is defined as the maximum channel error probability such that $p_{\mathcal{I}_0}^{(\ell)} \to 1$ as $\ell \to \infty$.

V. Numerical Results

A first set of results is related to the asymptotic performance of RSMP decoding. Tables I, II and III compare the iterative decoding thresholds $\epsilon^{\star}$ of RSMP, SMP, SRLMP (for maximum list size $\Gamma = 1$ and $\Gamma = 2$) and BP decoding $\epsilon^{\star}_{\mathsf{BP}}$ for $(4,8)$, $(3,6)$ and $(3,4)$ regular ensembles and several q values.

TABLE I
Decoding thresholds $\epsilon^{\star}$ of the $(4,8)$ regular LDPC code ensembles

q	SMP [16]	SRLMP [22] $\Gamma=1$	SRLMP [22] $\Gamma=2$	$\epsilon^{\star}$	$\epsilon^{\star}_{\mathsf{BP}}$	ϵ_{Sh}
2	0.0516	0.0656	-	0.0687	0.076	0.110
4	0.0814	0.0923	0.1075	0.1041	0.134	0.189
8	0.1064	0.1151	0.1332	0.1321	0.175	0.247
16	0.137	0.1389	0.1533	0.1481	0.204	0.2897
32	0.1636	0.1636	0.1673	0.1697	0.226	0.3217
64	0.1758	0.1758	0.1758	0.1866	0.241	0.3462

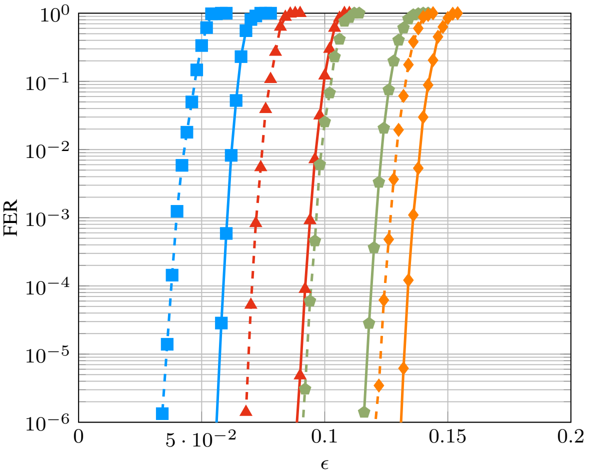


Fig. 1. FER versus channel error probability ϵ for regular $(4,8)$ LDPC codes with $n = 12000$ for SMP (-■-, -▲-, -⬟-, -◆-) and RSMP (—■—, —▲—, —⬟—, —◆—) for $q = 2$ (——), $q = 4$ (——), $q = 8$ (——) and $q = 16$ (——).

The tables also give the Shannon limit ϵ_{Sh} and the thresholds of the list message passing algorithm in [17] for maximum list size $\Gamma = 1$ and $\Gamma = 2$. Observe that our algorithm outperforms SMP decoding. This gain is due to including reliability scores in the decoding process. For RSMP, the size of the alphabet of the messages is $2q$ which is much smaller than the alphabet size of SRLMP in [22] and the list message passing [17] for maximum list size 2, which is equal to $1 + q(q+1)/2$. Remarkably, for some values of q and degree distributions, RSMP outperforms both SRLMP and the algorithm in [17] for maximum list size 2 and with reduced complexity and data flow.

To check the finite-length performance under RSMP, we consider the performance of a regular $(4,8)$ code where we set the maximum number of iterations $\ell_{\max} = 50$. The code has a block length $n = 12000$ and its Tanner graph is obtained via the progressive edge-growth (PEG) algorithm [25] and edge labels uniformly chosen in $\mathbb{F}_q \setminus \{0\}$. Finite-length simulation results for $q \in \{2,4,8,16\}$ are shown in Fig. 1 in terms of FER versus the QSC error probability ϵ. We use $\Delta = 1.6$ for $q = 2$ and 8, $\Delta = 1.5$ for $q = 4$ and $\Delta = 1.8$ for $q = 16$. Due to space limitations, the parameters $\mathsf{D}_{\mathsf{H}}^{(\ell)}, \mathsf{D}_{\mathsf{L}}^{(\ell)}$ are not provided but are obtained as a byproduct of DE analysis. As a reference, we provide the simulation results under SMP decoding [16].

TABLE II
DECODING THRESHOLDS $\epsilon^{\star}$ OF THE $(3, 6)$ REGULAR LDPC CODE ENSEMBLES

q	SMP [16]	[17] $\Gamma = 1$	[17] $\Gamma = 2$	SRLMP [22] $\Gamma = 1$	SRLMP [22] $\Gamma = 2$	$\epsilon^{\star}$	$\epsilon^{\star}_{\mathsf{BP}}$	ϵ_{Sh}
2	0.0395	0.039	-	0.0707	-	0.0741	0.084	0.110
4	0.0890	0.072	0.111	0.0946	0.1203	0.1102	0.149	0.189
8	0.1039	0.073	0.137	0.1086	0.1411	0.1390	0.196	0.247
16	0.1075	0.075	0.148	0.122	0.1517	0.1676	0.231	0.2897
32	0.1092	-	-	0.1387	0.1560	0.1814	0.26	0.3217
64	0.1101	-	-	0.1576	0.1585	0.1915	0.279	0.3462

TABLE III
DECODING THRESHOLDS $\epsilon^{\star}$ OF THE $(3, 4)$ REGULAR LDPC CODE ENSEMBLE

q	SMP [16]	[17] $\Gamma = 1$	[17] $\Gamma = 2$	SRLMP [22] $\Gamma = 1$	SRLMP [22] $\Gamma = 2$	$\epsilon^{\star}$	$\epsilon^{\star}_{\mathsf{BP}}$	ϵ_{Sh}
2	0.1069	0.106	-	0.1439	–	0.1448	0.167	0.2145
4	0.1724	0.123	0.222	0.1842	0.2390	0.2213	0.280	0.3546
8	0.1867	0.124	0.269	0.2096	0.2790	0.2791	0.355	0.4480
16	0.1930	0.120	0.287	0.2481	0.2977	0.3138	0.407	0.5120
32	0.1960	-	-	0.2893	0.3110	0.3382	0.444	0.5570
64	0.1974	-	-	0.3128	0.3175	0.354	0.475	0.5894

VI. CONCLUSIONS

A reduced-complexity decoding algorithm for q-ary LDPC codes on the QSC was presented. A DE analysis for irregular non-binary LDPC ensembles was developed. The presented DE yields the asymptotic iterative decoding thresholds and estimates the transition probabilities of the extrinsic channel needed for the VN update rule. Numerical results show that our algorithm outperforms competing algorithms with comparable complexity.

REFERENCES

[1] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[2] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, 2001.

[3] G. Lechner, T. Pedersen, and G. Kramer, "Analysis and design of binary message passing decoders," *IEEE Trans. Commun.*, vol. 60, no. 3, pp. 601–607, Mar. 2012.

[4] E. Ben Yacoub, F. Steiner, B. Matuz, and G. Liva, "Protograph-based LDPC code design for ternary message passing decoding," in *Proc. ITG Int. Conf. Syst. Commun. Coding (SCC)*, 2019, pp. 1–6.

[5] F. Steiner, E. Ben Yacoub, B. Matuz, G. Liva, and A. G. i. Amat, "One and two bit message passing for SC-LDPC codes with higher-order modulation," *J. Lightwave Technol.*, vol. 37, no. 23, pp. 5914–5925, Dec 2019.

[6] E. Ben Yacoub, "Matched quantized min-sum decoding of low-density parity-check codes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Riva del Garda, Italy, Apr. 2021.

[7] L. Barnault and D. Declercq, "Fast decoding algorithm for LDPC over GF(2^q)," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Paris, Mar. 2003, pp. 70–73.

[8] D. Declercq and M. Fossorier, "Decoding algorithms for nonbinary LDPC codes over GF(q)," *IEEE Trans. Commun.*, vol. 55, no. 4, pp. 633–643, Apr. 2007.

[9] X. Ma, K. Zhang, H. Chen, and B. Bai, "Low complexity X-EMS algorithms for nonbinary LDPC codes," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 9–13, 2012.

[10] S. Zhao, Z. Lu, X. Ma, and B. Bai, "A variant of the EMS decoding algorithm for nonbinary LDPC codes," *IEEE Commun. Lett.*, vol. 17, no. 8, pp. 1640–1643, 2013.

[11] L. Song, Q. Huang, and Z. Wang, "Set min-sum decoding algorithm for non-binary LDPC codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2016, pp. 3008–3012.

[12] M. G. Luby and M. Mitzenmacher, "Verification-based decoding for packet-based low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 120–127, 2005.

[13] B. Matuz, G. Liva, E. Paolini, and M. Chiani, "Verification-based decoding with map erasure recovery," in *Proc. ITG Int. Conf. Syst. Commun. Coding (SCC)*, 2013, pp. 1–6.

[14] A. Shokrollahi, "Capacity-approaching codes on the q-ary symmetric channel for large q," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2004, pp. 204–208.

[15] F. Zhang and H. D. Pfister, "Analysis of verification-based decoding on the q-ary symmetric channel for large q," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6754–6770, 2011.

[16] F. Lazaro, A. Graell i Amat, G. Liva, and B. Matuz, "Symbol message passing decoding of nonbinary low-density parity-check codes," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, 2019, pp. 1–5.

[17] B. M. Kurkoski, K. Yamaguchi, and K. Kobayashi, "Density evolution for GF(q) LDPC codes via simplified message-passing sets," in *Proc. Inf. Theory and Applications Workshop*, 2007, pp. 237–244.

[18] J. J. Metzner, "Majority-logic-like decoding of vector symbols," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1227–1230, Oct. 1996.

[19] C. Chen, B. Bai, X. Wang, and M. Xu, "Nonbinary LDPC codes constructed based on a cyclic MDS code and a low-complexity nonbinary message-passing decoding algorithm," *IEEE Commun. Lett*, vol. 14, no. 3, pp. 239–241, 2010.

[20] C.-Y. Chen, Q. Huang, C.-c. Chao, and S. Lin, "Two low-complexity reliability-based message-passing algorithms for decoding non-binary LDPC codes," *IEEE Trans. Commun.*, vol. 58, no. 11, pp. 3140–3147, 2010.

[21] F. Garcia-Herrero, M. J. Canet, J. Valls, and M. F. Flanagan, "Serial symbol-reliability based algorithm for decoding non-binary LDPC codes," *IEEE Commun. Lett.*, vol. 16, no. 6, pp. 909–912, 2012.

[22] E. Ben Yacoub, "List message passing decoding of non-binary low-density parity-check codes," *IEEE Int. Sym. Inf. Theory (ISIT)*, Jul. 2021.

[23] A. Ashikhmin, G. Kramer, and S. ten Brink, "Extrinsic information transfer functions: Model and erasure channel properties," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2657–2673, 2004.

[24] B. P. Smith, A. Farhood, A. Hunt, F. R. Kschischang, and J. Lodge, "Staircase Codes: FEC for 100 Gb/s OTN," *J. Lightwave Technol.*, vol. 30, no. 1, pp. 110–117, Jan. 2012.

[25] X.-Y. Hu, E. Eleftheriou, and D. M. Arnold, "Progressive edge-growth tanner graphs," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 2, Nov 2001, pp. 995–1001 vol.2.

On the Properties of Error Patterns in the Constant Lee Weight Channel

Jessica Bariffi*†, Hannes Bartz*, Gianluigi Liva*, and Joachim Rosenthal†
*Institute of Communication and Navigation, German Aerospace Center, 82234 Wessling, Germany
Email:{jessica.bariffi,hannes.bartz,gianluigi.liva}@dlr.de
†Institute of Mathematics, University of Zurich, CH-8057 Zürich, Switzerland
Email: rosenthal@math.uzh.ch

Abstract—**The problem of scalar multiplication applied to vectors is considered in the Lee metric. Unlike in other metrics, the Lee weight of a vector may be increased or decreased by the product with a nonzero, nontrivial scalar. This problem is of particular interest for cryptographic applications, like for example Lee metric code-based cryptosystems, since an attacker may use scalar multiplication to reduce the Lee weight of the error vector and thus to reduce the complexity of the corresponding generic decoder. The scalar multiplication problem is analyzed in the asymptotic regime. Furthermore, the construction of a vector with constant Lee weight using integer partitions is analyzed and an efficient method for drawing vectors of constant Lee weight uniformly at random from the set of all such vectors is given.**

I. Introduction

In the late 1950s, relating to transmitting symbols from a finite prime field $\mathbb{F}_q$, the Lee metric was introduced in [1], [2]. Error correcting codes endowed with the Lee metric (like BCH codes, dense error-correcting codes or codes with maximum Lee distance) were constructed and applied in various different manners [3]–[9]. Recently, the Lee metric was applied to DNA storage systems [10] and considered for cryptographic applications [11]. New families of error correcting codes endowed with the Lee metric together with an iterative decoding algorithm were proposed [12] while information set decoding (ISD) in the Lee metric has been analyzed [11], [13].

ISD is one way to solve the well-known generic (syndrome) decoding problem, which aims at decoding an arbitrary linear code efficiently without knowing or using the structure of the code. This problem is fundamental for code-based cryptography and was shown to be NP-complete in both the Hamming metric [14], [15] and the Lee metric [16]. The desirable feature of generic (syndrome) decoding is to succeed in correcting an error vector $\mathbf{e}$ as long as its corresponding weight is small, where small refers to the Gilbert-Varshamov bound [17], [18]. In fact, syndrome decoding has an exponential complexity in the weight of the error for both the Hamming and the Lee weight. From an adversarial point of view, the goal is to reduce the weight of the introduced error vector in order to make the generic (syndrome) decoding problem more feasible. In fact, while the Hamming weight of a vector with entries from a finite field is invariant under multiplication with a nonzero scalar, the Lee weight of a vector can be increased or decreased by the product with a scalar. Understanding under which conditions (and with what probability) the Lee weight on the error vector $\mathbf{e}$ is reduced represents a key preliminary step in the design of Lee metric code-based cryptosystems. We will refer to this problem as *scalar multiplication problem.*

In this paper, we consider an additive channel model that adds an error vector of a fixed Lee weight to the transmitted codeword. We will refer to this channel as the *constant Lee weight channel.* We present an algorithm that draws a vector of length n and fixed Lee weight t over the ring of integers $\mathbb{Z}_m$ modulo m uniformly at random from the set of vectors with the same parameters. Introducing errors uniformly at random is important from a cryptographic point of view in order to hide the structure of the error pattern. We will then derive the marginal distribution of the constant Lee weight channel in the limit of large block lengths n. This result enables to analyze how the Lee weight of a given error vector changes when multiplied by a random nonzero scalar, in the asymptotic regime. We show that, under certain conditions, the Lee weight of such an error vector will not decrease after scalar multiplication with high probability.

The paper is organized as follows. Section II provides the notations and preliminaries needed for the course of the paper. In Section III we introduce the constant Lee weight channel and provide a uniform construction of an error vector of given Lee weight among all possible vectors of the same Lee weight. The scalar multiplication problem is introduced in Section IV. We state the problem in a finite length setting and analyze it in the asymptotic regime. Conclusions are stated in Section V.

II. Notation and Preliminaries

We denote by $\mathbb{Z}_m$ the ring of integers modulo m, where m is a positive integer. To simplify the reading, vectors will be denoted by boldface lower-case letters.

A. The Lee Metric

Definition 1. *The Lee weight of a scalar $a \in \mathbb{Z}_m$ is defined as*

$$\mathrm{wt}_{\mathrm{L}}(a) := \min(a, m-a).$$

The Lee weight of a vector $\mathbf{x} \in \mathbb{Z}_m^n$ of length n is defined as the sum of the Lee weights of its entries, i.e.

$$\mathrm{wt}_{\mathrm{L}}(\mathbf{x}) := \sum_{i=1}^{n} \mathrm{wt}_{\mathrm{L}}(x_i).$$

Note that the Lee weight of an element $a \in \mathbb{Z}_m$ is upper bounded by $\lfloor m/2 \rfloor$. Hence, the Lee weight of a length-n vector $\mathbf{x}$ over $\mathbb{Z}_m$ is at most $n \cdot \lfloor m/2 \rfloor$. To simplify the notation, we define

$$r := \lfloor m/2 \rfloor.$$

Furthermore, we observe that if $m \in \{2, 3\}$ the Lee weight is equivalent to the Hamming weight.

If we consider the elements of $\mathbb{Z}_m$ as points placed along a circle such that the circle is divided into m arcs of equal length, then the Lee distance between two distinct values a and b can be interpreted as the smallest number of arcs separating the two values. Therefore, the following property holds

$$\mathrm{wt}_{\mathsf{L}}(a) = \mathrm{wt}_{\mathsf{L}}(m-a) \quad \text{for every } a \in \{1, \ldots, r\}. \tag{1}$$

The Lee distance between two vectors is defined as follows.

Definition 2. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_m^n$. The Lee distance between $\mathbf{x}$ and $\mathbf{y}$ is given by the Lee weight of their difference, i.e.*

$$\mathrm{d}_L(\mathbf{x}, \mathbf{y}) := \mathrm{wt}_{\mathsf{L}}(\mathbf{x} - \mathbf{y}).$$

It is well-known that the Lee distance indeed induces a metric.

B. Useful Results from Information Theory

Let X be a random variable over an alphabet $\mathcal{X}$ with probability distribution P, where $P(x) := \mathbb{P}(X = x)$ with $x \in \mathcal{X}$. The entropy $H(X)$ is defined as

$$H(X) := -\sum_{x \in \mathcal{X}} P(x) \log(P(x)).$$

The Kullback-Leibler divergence between two distributions Q and P is denoted as

$$D(Q \,||\, P) := \sum_{x \in \mathcal{X}} Q(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

Theorem 1 (Conditional Limit Theorem [19, Theorem 11.6.2])**.** *Let E be a closed convex subset of probability distributions over a given alphabet $\mathcal{X}$ and let Q be a distribution not in E over the same alphabet $\mathcal{X}$. Consider $X_1, \ldots, X_n$ to be discrete random variables drawn i.i.d. $\sim Q$ and let $P^\star = \arg\min_{P \in E} D(P \,||\, Q)$. Denote by X^n the random sequence $(X_1, \ldots, X_n)$ and P_{X^n} its empirical distribution. Then for any $a \in \mathcal{X}$*

$$\mathbb{P}\left(X_1 = a \,|\, P_{X^n} \in E\right) \longrightarrow P^\star(a)$$

in probability as n grows large.

C. Combinatorics

Definition 3. *Let t and s be positive integers. An integer partition of t into s parts is an s-tuple $\lambda := (\lambda_1, \ldots, \lambda_s)$ of positive integers satisfying the following two properties:*

i. $\lambda_1 + \ldots + \lambda_s = t$,
ii. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$.

The elements λ_i are called parts and we say that s is the length of the partition λ.

Note that the order of the parts does not matter. This means that, for instance, the tuples $(1, 1, 2)$, $(1, 2, 1)$ and $(2, 1, 1)$ are all identical and represented only by $(2, 1, 1)$. We will denote by Π_λ the set of all permutations of an integer partition λ. Let n_i denote the number of occurrences of a positive integer i in an integer partition λ of t, where $i \in \{1, \ldots, t\}$, then $|\Pi_\lambda| = \binom{t}{n_1, \ldots, n_t} = \frac{t!}{n_1! \ldots n_t!}$.
In the following, we use $\mathcal{P}(t)$ to denote the set of integer partitions of t. We write $\mathcal{P}_k(t)$ instead, if we restrict $\mathcal{P}(t)$ to those partitions with part sizes not exceeding some fixed nonnegative integer value k. Note that for any $\lambda \in \mathcal{P}_k(t)$ its length ℓ_λ is bounded by $\lceil \frac{t}{k} \rceil \leq \ell_\lambda \leq t$.

We will now introduce a definition describing vectors whose Lee weight decomposition is based on a given integer partition.

Definition 4. *For a positive integer n and a given partition $\lambda \in \mathcal{P}_r(t)$ of a positive integer t, we say that a length-n vector $\mathbf{x}$ has weight decomposition λ over $\mathbb{Z}_m$ if there is a one-to-one correspondence between the Lee weight of the nonzero entries of $\mathbf{x}$ and the parts of λ.*

Example 1. *Let $n = 5$ and let $\lambda = (2, 1, 1)$ be an integer partition of $t = 4$ over $\mathbb{Z}_7$. All vectors of length n over $\mathbb{Z}_7$ consisting of one element of Lee weight 2 and two elements of Lee weight 1 have weight decomposition λ.*

We will denote the set of all vectors of length n of the same weight decomposition $\lambda \in \mathcal{P}(t)$ by $\mathcal{V}_{t,\lambda}^{(n)}$.

III. The Constant Lee Weight Channel

Let us consider a channel

$$\mathbf{y} = \mathbf{x} + \mathbf{e},$$

where $\mathbf{y}, \mathbf{x}$ and $\mathbf{e}$ are length-n vectors over $\mathbb{Z}_m$ and the channel introduces the error vector $\mathbf{e}$ uniformly at random from the set $\mathcal{S}_{t,m}^{(n)}$ of all vectors in $\mathbb{Z}_m^n$ with a fixed Lee weight t, i.e.

$$\mathcal{S}_{t,m}^{(n)} := \{\mathbf{e} \in \mathbb{Z}_m^n \mid \mathrm{wt}_{\mathsf{L}}(\mathbf{e}) = t\}.$$

A. Marginal Channel Distribution

Since certain decoder types (e.g., iterative decoders employed for low-density parity-check codes defined over integer rings) require the knowledge of the channel's marginal conditional distribution, our goal is to describe the marginal distribution P_e, for a generic element E of the error.

Lemma 1. *The marginal distribution of a constant Lee weight channel over $\mathbb{Z}_m$ is given by*

$$P_e^\star = \frac{1}{\sum_{j=0}^{m-1} \exp(-\beta\, \mathrm{wt}_{\mathsf{L}}(j))} \exp\left(-\beta\, \mathrm{wt}_{\mathsf{L}}(e)\right),$$

for some constant $\beta > 0$.

Proof. Following [19, Ch. 12], we are looking for a distribution $\mathbf{P} = (P_0, \ldots, P_{m-1})$ that maximizes the entropy function

$$\mathrm{H}_e(\mathbf{P}) := -\sum_{e=0, P_e \neq 0}^{m-1} P_e \log P_e$$

under the constraint that the Lee weight of the vector is t, or equivalently, that the normalized Lee weight of the error vector is $\delta := t/n$, i.e.

$$\sum_{e=0}^{m-1} \mathrm{wt}_{\mathsf{L}}(e) P_e = \delta.$$

Let us introduce a Lagrange multiplier $\beta > 0$, which is the solution to

$$\delta = \frac{(k-1)\mathrm{e}^{(k+1)\beta} - k\mathrm{e}^{k\beta} + \mathrm{e}^{\beta}}{(\mathrm{e}^{\beta k} - 1)(\mathrm{e}^{\beta} - 1)}$$

with $k = r + 1$. Then the optimization problem has the following solution

$$P_e^\star = \kappa \exp\left(-\beta \, \mathrm{wt}_{\mathsf{L}}(e)\right), \tag{2}$$

where κ is a normalization constant enforcing $\sum_e P_e^\star = 1$. □

The solution (2) is closely related to the problem in statistical mechanics of finding the distribution of the energy state of a given system [19]–[21]. Here, we may interpret the energy value of the particles as the Lee weight $\mathrm{wt}_{\mathsf{L}}(e)$ of an element $e \in \mathbb{Z}_m$. Note that for the channel law determined by Lemma 1, the optimum decoder will seek for the codeword at minimum Lee distance from the channel output $\mathbf{y}$.

B. Error Pattern Construction

In the following we will present an algorithm that draws a vector uniformly at random from $\mathcal{S}_{t,m}^{(n)}$ for given parameters n, t and m. The idea is inspired by the algorithm presented in [12]. We start from partitioning the desired Lee weight t into integer parts of size at most r, since the Lee weight of any $a \in \mathbb{Z}_m$ is at most r. The main difference to the algorithm presented in [12, Lemmas 2 and 3], and crucial to design the vector uniformly at random from $\mathcal{S}_{t,m}^{(n)}$, is that the integer partition of t is not chosen uniformly at random from the set of all integer partitions $\mathcal{P}_r(t)$ of t. In fact, picking a partition uniformly at random from $\mathcal{P}_r(t)$ yields that some of the vectors in $\mathcal{S}_{t,m}^{(n)}$ are more probable than others. Therefore, we need to understand the number of vectors with weight decomposition λ, for a fixed partition $\lambda \in \mathcal{P}_r(t)$. The following result gives an answer to this question.

Lemma 2. *Let n, m and t be positive integers with $t \leq n$ and consider the set of partitions $\mathcal{P}_r(t)$ of t with part sizes not exceeding r. For any $\lambda \in \mathcal{P}_r(t)$ the number of vectors of length n over $\mathbb{Z}_m$ with weight decomposition λ is given by*

$$\left|\mathcal{V}_{t,\lambda}^{(n)}\right| = \begin{cases} 2^{\ell_\lambda} \left|\Pi_\lambda\right| \binom{n}{\ell_\lambda} & \text{if } m \text{ is odd,} \\ 2^{\ell_\lambda - c_{r,\lambda}} \left|\Pi_\lambda\right| \binom{n}{\ell_\lambda} & \text{else} \end{cases}$$

where $c_{r,\lambda} = |\{i \in \{1, \dots, \ell_\lambda\} \mid \lambda_i = r\}|$.

Proof. Recall from Definition 4 that $\mathcal{V}_{t,\lambda}^{(n)}$ consists of all length n vectors $\mathbf{x}$ whose nonzero entries are in one-to-one correspondence with the parts of λ. Let $x_{i_1}, \dots, x_{i_{\ell_\lambda}}$ denote the nonzero positions of $\mathbf{x}$ and let us first consider the case where

$$\mathrm{wt}_{\mathsf{L}}(x_{i_1}) = \lambda_1, \ \dots, \ \mathrm{wt}_{\mathsf{L}}(x_{i_{\ell_\lambda}}) = \lambda_{\ell_\lambda}. \tag{3}$$

Finding the number of such vectors relies on the "selection with repetition" problem [22, Section 1.2], which implies that this number is exactly $\binom{\text{number of zeros } + \text{ free spaces } - 1}{\text{free spaces } - 1}$, i.e.

$$\binom{(n - \ell_\lambda) \ + \ (\ell_\lambda + 1) \ - \ 1}{(\ell_\lambda + 1) \ - \ 1} = \binom{n}{\ell_\lambda},$$

where with "free spaces" we mean all the possible gaps in front, between and at the end of the parts of λ.

If m is odd, the number n_i of elements in $\mathbb{Z}_m$ having a nonzero Lee weight i is always 2 for every possible Lee weight $i \in \{1, \dots, r\}$. Hence, there are $2^{\ell_\lambda}\binom{n}{\ell_\lambda}$ vectors satisfying (3). On the other hand, if m is even, then $n_i = 2$ for $i \in \{1, \dots, r-1\}$ and $n_r = 1$. If we define $c_{r,\lambda} = |\{i \in \{1, \dots, \ell_\lambda\} \mid \lambda_i = r\}|$ to be the number of parts of λ equal to r, then the number of parts of λ that can be flipped is $2^{\ell_\lambda - c_{r,\lambda}}$. Hence, the number of vectors satisfying (3) is $2^{\ell_\lambda - c_{r,\lambda}}\binom{n}{\ell_\lambda}$.

Finally, since the ordering of the nonzero elements of $\mathbf{x}$ is not necessarily the same as the order of the parts of λ, we multiply $\binom{n}{\ell_\lambda}$ by the number of permutations $|\Pi_\lambda|$ of λ and obtain the desired result. □

Finally, the actual vector construction over $\mathbb{Z}_m$, described in Algorithm 1, mainly consists of picking a partition $\lambda \in \mathcal{P}_r(t)$ of the Lee weight t with part sizes not exceeding r. The probability of $\mathbf{x} \in \mathcal{S}_{t,m}^{(n)}$ with weight decomposition $\lambda \in \mathcal{P}_r(t)$ is given by

$$p_\lambda := \frac{\left|\mathcal{V}_{t,\lambda}^{(n)}\right|}{\sum_{\tilde{\lambda} \in \mathcal{P}_{r(t)}} \left|\mathcal{V}_{t,\tilde{\lambda}}^{(n)}\right|}.$$

The idea is to choose the integer partition according to the probability mass function $\mathcal{X}_{t,m}^{(n)}$ defined by the probabilities p_λ, for $\lambda \in \mathcal{P}_r(t)$. We will denote this procedure by

$$\lambda \xleftarrow{\mathcal{X}_{t,m}^{(n)}} \mathcal{P}_r(t).$$

We then randomly flip the elements of the partition modulo m and assign these values to randomly chosen positions of the error vector. Choosing an element a uniformly at random from a given set $\mathcal{A}$ will be denoted by $a \xleftarrow{\$} \mathcal{A}$. We want to emphasize at this point that for fixed parameters n, t and m the computation of $\mathcal{X}_{t,m}^{(n)}$ needs to be done only once at the beginning, since the distribution is only dependent on these parameters and does not change anymore.

Theorem 2. *Let n, m and t be positive integers. Algorithm 1 draws a vector uniformly at random among $\mathcal{S}_{t,m}^{(n)}$.*

Proof. First note that $\mathcal{S}_{t,m}^{(n)} = \bigsqcup_{\lambda \in \mathcal{P}_r(t)} \mathcal{V}_{t,\lambda}^{(n)}$, where $\bigsqcup$ denotes the disjoint union of sets. Hence, we want to pick $\lambda \in \mathcal{P}_r(t)$ such that all the vectors in $\mathcal{S}_{t,m}^{(n)}$ are equally probable to be drawn. The choice of λ is decisive for the set $\mathcal{V}_{t,\lambda}^{(n)}$. Since

Algorithm 1 Drawing a vector uniformly at random from $\mathcal{S}_{t,m}^{(n)}$

```
Require: $n, m, t \in \mathbb{N}_{>0}$, distribution $\mathcal{X}_{t,m}^{(n)}$
Ensure: $\mathbf{e} \overset{\$}{\leftarrow} \mathcal{S}_{t,m}^{(n)}$
$\lambda \xleftarrow{\mathcal{X}_{t,m}^{(n)}} \mathcal{P}_r(t)$
$F = \{f_1, \dots, f_{\ell_\lambda}\} \overset{\$}{\leftarrow} \{\pm 1\}^{\ell_\lambda}$
$\mathrm{supp}(\mathbf{e}) \overset{\$}{\leftarrow} \{S \subset \{1, \dots, n\} : |S| = \ell_\lambda\}$
for $i = 1, \dots, n$ do
    if $i \in \mathrm{supp}(\mathbf{e})$ then
        $e_i \leftarrow f_i \cdot \lambda_i$
    else
        $e_i = 0$
    end if
end for
return random_permutation($\mathbf{e}$)
```

$\left|\mathcal{V}_{t,\lambda}^{(n)}\right|$ changes with λ, we pick λ according to distribution p_λ from $\mathcal{X}_{t,m}^{(n)}$ using Lemma 2 and the result follows. □

IV. Scalar Multiplication Problem

While we know that the Hamming weight of a vector over a finite field is invariant under multiplication with a nonzero scalar, the Lee weight can possibly change. In this section, we analyze the behavior of the Lee weight of a vector when multiplied by a scalar. Recalling that the Lee metric coincides with the Hamming metric over $\mathbb{Z}_2$ and $\mathbb{Z}_3$, in the following we will focus only on the case where the Lee weight is different from the Hamming weight, i.e. we focus on $\mathbb{Z}_m$ with $m > 3$.

Remark 1. *Even though we will not discuss the following, we want to emphasize at this point that the Hamming weight is* not *invariant under multiplication with a nonzero scalar when working over a finite integer ring that is not a field.*

A. Problem Statement

We now establish bounds on the probability of reducing the Lee weight of a random vector by multiplying it with a random nonzero scalar.

Problem 1. *Consider the ring of integers $\mathbb{Z}_m$, with $m > 3$. Given a random vector $\mathbf{x} \in \mathbb{Z}_m^n$ with Lee weight $\mathrm{wt}_{\mathsf{L}}(\mathbf{x}) = t$ uniformly distributed in $\mathcal{S}_{t,m}^{(n)}$. Let a be chosen uniformly at random from $\mathbb{Z}_m \backslash \{0\}$. Find the probability that the Lee weight of $a \cdot x$ is less than the Lee weight t of x, i.e.*

$$\mathbb{P}\left(\mathrm{wt}_{\mathsf{L}}(a \cdot x) < t\right).$$

For simplicity, let us define the following event

$$F := \{\mathrm{wt}_{\mathsf{L}}(a \cdot \mathbf{x}) < t\}.$$

We denote by $Q_{\mathbf{x}}$ the empirical distribution of the entries of $\mathbf{x}$. Recall the distribution $P^\star$ defined in (2). We will rewrite $\mathbb{P}(F)$ by distinguishing between vectors $\mathbf{x}$ with $Q_{\mathbf{x}}$ close to $P^\star$ and all others, where by "close" we mean with respect to the Kullback-Leibler divergence, i.e. $Q_{\mathbf{x}}$ satisfies $D(Q_{\mathbf{x}} \,||\, P^\star) < \varepsilon$ for some $\varepsilon > 0$ small. We have that

$$\begin{aligned} \mathbb{P}(F) \le\, & \mathbb{P}\left(\mathrm{wt}_{\mathsf{L}}(a \cdot \mathbf{x}) < t \,|\, D(Q_{\mathbf{x}} \,||\, P^\star) < \varepsilon\right) \\ & + \mathbb{P}\left(D(Q_{\mathbf{x}} \,||\, P^\star) \ge \varepsilon\right). \end{aligned} \tag{4}$$

Note that the probability $\mathbb{P}(F)$ is dependent on three parameters: the length n of the constructed vector $\mathbf{x}$, the size m of the integer ring and the given Lee weight t of $\mathbf{x}$. The evaluation of the bound (4) is challenging for $m > 3$, finite n and generic t. In the following subsection we will describe how to attack the problem for n large.

B. Asymptotic Analysis

Let us focus now on the asymptotic regime, i.e. where the block length n tends to infinity. Note here that we let $\mathrm{wt}_{\mathsf{L}}(\mathbf{x}) = t$ grow linearly with n. Let us denote by $U(\mathbb{Z}_m)$ the uniform distribution over $\mathbb{Z}_m$ and let E be the set of probability distributions over $\mathbb{Z}_m$ with an average Lee weight $\delta := t/n$, i.e.

$$E := \left\{ p = (p_0, \dots, p_{m-1}) \Big| \sum_{i=0}^{m-1} p_i = 1 \text{ and } \sum_{i=0}^{m-1} p_i \mathrm{wt}_{\mathsf{L}}(i) = \delta \right\}$$

Hence, a straightforward application of Theorem 1 yields the following corollary.

Corollary 1. *Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{Z}_m^n$ a random vector drawn uniformly from $\mathcal{S}_{\delta n,m}^{(n)}$. Then, for every $\varepsilon > 0$ it holds*

$$\mathbb{P}\left(D(Q_{\mathbf{x}} \,||\, P^\star) \ge \varepsilon\right) \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Proof. Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{Z}_m^n$ be a random vector whose entries are independent and uniformly distributed in $\mathbb{Z}_m$. The distribution of $\mathbf{x}$ is uniform on $\mathbb{Z}_m^n$, and hence on $\mathcal{S}_{\delta n,m}^{(n)}$. We have that

$$P^\star = \arg\min_{P \in E} D(P \,||\, U(\mathbb{Z}_m)).$$

Then, by Theorem 1, we obtain the desired result. □

In fact, Theorem 1 allows to assume that the entries of a sequence $\mathbf{x}$ drawn uniformly in $\mathcal{S}_{\delta n,m}^{(n)}$ are distributed according to $P^\star$ as n grows large. Hence, in the asymptotic regime, Problem 1 reduces to estimating the probability $\mathbb{P}\left(\mathrm{wt}_{\mathsf{L}}(a \cdot \mathbf{x}) \le \mathrm{wt}_{\mathsf{L}}(\mathbf{x}) \,|\, D(Q_{\mathbf{x}} \,||\, P^\star) < \varepsilon\right)$. In that case, we apply Definition 1 for the Lee weight of a vector $\mathbf{x}$. Then the assumption that the entries of $\mathbf{x}$ are distributed as in (2) yields, in the limit of n large, the following equivalent description of the desired probability

$$\begin{aligned} \lim_{n \longrightarrow \infty} \mathbb{P}(F) = \mathbb{P}\Big(& \sum_{i=1}^{m-1} \mathrm{e}^{-\beta \,\mathrm{wt}_{\mathsf{L}}(i)}\, \mathrm{wt}_{\mathsf{L}}([a \cdot i]_m) \\ & < \sum_{i=1}^{m-1} \mathrm{e}^{-\beta \,\mathrm{wt}_{\mathsf{L}}(i)}\, \mathrm{wt}_{\mathsf{L}}(i)\Big) \end{aligned} \tag{5}$$

By Property (1), we can run the sum only up to r. Nevertheless we need to distinguish between even or odd ring order m. In particular, for m odd we rewrite (5) as

$$\lim_{n \longrightarrow \infty} \mathbb{P}(F) = \mathbb{P}\Big(0 < \sum_{i=1}^{r} \mathrm{e}^{-\beta i}(i - \mathrm{wt}_{\mathsf{L}}([a \cdot i]_m))\Big) \tag{6}$$

whereas for m even (5) is equivalent to

$$\lim_{n\longrightarrow\infty}\mathbb{P}(F)=\mathbb{P}\Big(0<\sum_{i=1}^{r-1}2\mathrm{e}^{-\beta i}(i-\mathrm{wt}_{\mathsf{L}}([a\cdot i]_m)) \\ +\mathrm{e}^{-\beta r}(r-\mathrm{wt}_{\mathsf{L}}([a\cdot r]_m))\Big) \quad (7)$$

where $[a\cdot i]_m$ denotes the reduction of $a\cdot i \mod m$.

Since we want $\mathbb{P}(F)$ to be small (or equal to zero), we need to understand under which circumstances the sums in (6) and (7) are non-positive. Note that both $\sum_{i=1}^{r}\mathrm{e}^{-\beta i}(i-\mathrm{wt}_{\mathsf{L}}([a\cdot i]_m))$ and $\sum_{i=1}^{r-1}2\mathrm{e}^{-\beta i}(i-\mathrm{wt}_{\mathsf{L}}([a\cdot i]_m))+\mathrm{e}^{-\beta r}(r-\mathrm{wt}_{\mathsf{L}}([a\cdot r]_m)$ are dependent on m and β, where β depends on δ. If we fix these parameters, we are able to compute the sum and hence (5). We therefore fix m and evaluate the two expressions for different values of δ. Let $\delta^\star$ denote the largest normalized Lee weight such that (6) or rather (7) are equal to zero for every $\delta<\delta^\star$. Table I shows the values of the threshold $\delta^\star$ for different ring orders m.

TABLE I

MAXIMAL NORMALIZED LEE WEIGHT $\delta^\star$ OVER $\mathbb{Z}_m$ SUCH THAT $\mathbb{P}(F)=0$ AS $n\longrightarrow\infty$, FOR SOME VALUES OF m COMPARED TO THE MAXIMAL POSSIBLE NORMALIZED LEE WEIGHT r.

m	5	7	8	9	11	15	16	31	32	53
r	2	3	4	4	5	7	8	15	16	26
$\delta^\star$	1.2	1.714	2	1.962	2.727	3.310	4	7.741	8	13.245

Observe from Table I that for m an odd prime power and for $\delta^\star=(m^2-1)/4m$ (i.e. the average Lee weight when choosing an element uniformly from $\mathbb{Z}_m$ [23]) the Lee weight of a vector $\mathbf{x}\in\mathbb{Z}_m^n$ can never be reduced when multiplied by a nonzero scalar. This fact can be established by observing that the multiplication of a random variable X in $\mathbb{Z}_m$ by $a\in\mathbb{Z}_m\setminus\{0\}$ induces a permutation of the distribution. Moreover, if X is distributed according to $P^\star$ with $\beta>0$, the permutation that maximizes $\mathbb{E}(\mathrm{wt}_{\mathsf{L}}(aX))$ is the identity, i.e., $a=1$. On the contrary, if $\beta<0$, the identity permutation ($a=1$) minimizes $\mathbb{E}(\mathrm{wt}_{\mathsf{L}}(aX))$. The result follows by observing that $\beta>0$ implies that the average Lee weight is $\delta<(m^2-1)/4m$.

Note that the same result follows for any m if $a\in\mathbb{Z}_m^\times$ is a unit modulo m. Moreover, if m is a power of 2, the threshold is

$$\delta^\star=m/4.$$

V. Conclusions

In this work we have introduced an algorithm for the construction of error patterns over $\mathbb{Z}_m^n$ of a fixed Lee weight. The algorithm is efficient compared to straightforward approaches, which are more involved in terms of computation and memory. The proposed algorithm is based on the idea of subdividing the tasks into subtasks, which are more or less easy to solve. The procedure is dominated by the computation of the distribution used to choose the underlying integer partition of a vector's Lee weight decomposition. For a fixed Lee weight t, this distribution can be pre-computed. We have shown that the presented algorithm draws a vector uniformly at random among all vectors of the same length and Lee weight. This property is important for cryptographic applications in the context of Lee metric code-based cryptography in order to avoid information leakage on the structure of the error pattern. Additionally, the results on the constant-weight Lee channel together with the random construction of sequences of fixed Lee weight were used to derive the probability of reducing the Lee weight of a vector over $\mathbb{Z}_m^n$ when multiplying it by a random nonzero element of $\mathbb{Z}_m$, for the limit case where the sequence length grows large. An open problem is to characterize this probability in the finite sequence length regime.

References

[1] W. Ulrich, "Non-binary error correction codes," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1341–1388, 1957.

[2] C. Lee, "Some properties of nonbinary error-correcting codes," *IRE Trans. Inf. Theory*, vol. 4, no. 2, pp. 77–82, 1958.

[3] E. Prange, "The use of coset equivalene in the analysis and decoding of group codes," Air Force Cambridge Research Labs, Tech. Rep., 1959.

[4] E. R. Berlekamp, "Negacyclic codes for the Lee metric," North Carolina State University. Dept. of Statistics, Tech. Rep., 1966.

[5] S. W. Golomb and L. R. Welch, "Algebraic coding and the Lee metric," *Error Correcting Codes*, pp. 175–194, 1968.

[6] J. C.-Y. Chiang and J. K. Wolf, "On channels and codes for the Lee metric," *Information and Control*, vol. 19, no. 2, pp. 159–173, 1971.

[7] R. M. Roth and P. H. Siegel, "Lee-metric BCH codes and their application to constrained and partial-response channels," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1083–1096, Apr. 1994.

[8] T. Etzion, A. Vardy, and E. Yaakobi, "Dense error-correcting codes in the Lee metric," in *Proc. IEEE Information Theory Workshop*, Sep. 2010.

[9] T. L. Alderson and S. Huntemann, "On maximum Lee distance codes," *Journal of Discrete Mathematics*, 2013.

[10] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4982–4995, Aug. 2017.

[11] A.-L. Horlemann-Trautmann and V. Weger, "Information set decoding in the Lee metric with applications to cryptography," *arXiv preprint arXiv:1903.07692*, 2019.

[12] P. Santini, M. Battaglioni, F. Chiaraluce, M. Baldi, and E. Persichetti, "Low-Lee-Density Parity-Check Codes," in *Proc. 2020 IEEE International Conference on Communications (ICC)*, June 2020.

[13] V. Weger, M. Battaglioni, P. Santini, F. Chiaraluce, M. Baldi, and E. Persichetti, "Information set decoding of Lee-metric codes over finite rings," *arXiv preprint arXiv:2001.08425*, 2020.

[14] S. Barg, "Some new NP-complete coding problems," *Problemy Peredachi Informatsii*, vol. 30, no. 3, pp. 23–28, 1994.

[15] E. Berlekamp, R. McEliece, and H. Van Tilborg, "On the inherent intractability of certain coding problems (corresp.)," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 384–386, 1978.

[16] V. Weger, M. Battaglioni, P. Santini, A.-L. Horlemann-Trautmann, and E. Persichetti, "On the hardness of the Lee syndrome decoding problem," *arXiv e-prints*, 2020.

[17] E. N. Gilbert, "A comparison of signalling alphabets," *The Bell System Technical Journal*, vol. 31, no. 3, pp. 504–522, 1952.

[18] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," *Docklady Akad. Nauk, SSSR*, vol. 117, pp. 739–741, 1957.

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. New York: Wiley, 2006.

[20] L. Boltzmann, "Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten (studies of the equilibrium and the life force between material points)," *Wien. Ber*, vol. 58, p. 517, 1868.

[21] J. W. Gibbs, *Elementary principles in statistical mechanics: developed with special reference to the rational foundation of thermodynamics*. Dover Publications, 1902.

[22] S. Jukna, *Extremal combinatorics: with applications in computer science*. Springer Science & Business Media, 2011.

[23] A. D. Wyner and R. L. Graham, "An upper bound on minimum distance for a k-ary code," *Inf. Control.*, vol. 13, no. 1, pp. 46–52, 1968.

Error Probability Bounds for Coded-Index DNA Storage Channels

Nir Weinberger
The Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200004, Israel
nirwein@technion.ac.il

***Abstract*—The DNA storage channel is considered, where each codeword is comprised of M unordered DNA molecules. At reading time, the molecules are sampled N times with replacement, and then sequenced. A coded-index concatenated-coding scheme is proposed, in which the mth molecule of the codeword is restricted to an inner code, unique for each index. A low-complexity decoder is proposed that is based on separated decoding of each molecule (inner code), followed by decoding the sequence of molecules (outer code). Mild assumptions are made on the sequencing channel, in the form of the existence of an inner code and decoder with vanishing error probability. The error probability of a random code for the storage system is analyzed and shown to decay exponentially with N. This establishes the importance of high coverage depth N/M for achieving low error probability.**

I. Introduction

Various authors have recently proposed and analyzed coding methods for data storage systems based on a Deoxyribonucleic acid (DNA) medium (see a survey in [1]). In this channel model, information is stored in a pool of M DNA molecules, where each molecule is comprised of two complementary length-L strands of four nucleotides (Adenine, Cytosine, Guanine, and Thymine). The M molecules cannot be spatially ordered, and during reading, N molecules are independently sampled from the DNA pool, with replacement. Then, each of these sampled molecules is *sequenced* in order to obtain a length-L vector describing the synthesized nucleotides, and the N sequenced molecules is the channel output. Roughly speaking, the impairments of this channel include: (1) *Molecule errors* – e.g., the event in which some of the molecules are not sampled at all (erased). (2) *Symbol errors* – modeled by a channel $W^{(L)}$ which specifies the probability of sequencing some L-symbol vector conditioned that the information was (possibly other) L symbols. In this paper, we propose a random coding ensemble and a low-complexity decoder for this channel model, and analyze the average error probability.

In terms of fundamental limits, it was the capacity of such a channel which was first addressed [1], with the general conclusion that the capacity is positive only when $L = \beta \log M$, with $\beta > 1$. Under this scaling, [1]–[4], have derived bounds on the capacity, assuming a constant *coverage depth* N/M, and a discrete memoryless sequencing channel. In this paper, we focus on a somewhat different model for the following reasons: First, the tightest achievable bound for a discrete memoryless channel (DMC) [4] require a computationally intensive decoder, which is difficult to implement in practice. Second, in practice, the sequencing channel is not a DMC, and may include deletions and insertions [5], or constraints on the codeword symbols [6], [7]. Third, as was also established in [4], the error probability is dominated by molecule errors (erasures), and so the error probability decays as $e^{-\Theta(M)}$ rather than the $e^{-\Theta(ML)}$ decay rate anticipated from a blocklength of ML. This slow decay of the error probability is significant for practical systems of finite blocklength.

Accordingly, and following [2], in this paper, we theoretically analyze the error probability of a simple, yet general, coding method. The scheme follows a practical approach [8]–[11] in which the lack of order of the molecules is resolved by an *index*. The simplest version of indexing-based schemes uses the first $\log_2 M$ bits of each DNA molecule to specify its index $m \in [M]$, and is capacity achieving for noiseless sequencing channels, despite its rate loss of $1/\beta$, which seems to be inevitable [1], [3], [4], [12], [13]. Nonetheless, if the payload bits (the last $(\beta - 1)\log_2 M$ bits of the molecule) are arbitrary, then an erroneous ordering of the molecules can be caused by a single channel bit flip. This motivates us to consider in this paper *coded-indexing* based schemes for noisy sequencing channels. In such a scheme, the possible molecules of the codeword are chosen from a inner code – a sub-code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ of all possible molecules. Moreover, this inner code is further partitioned into M equal size sub-codes $\mathcal{B}_m^{(L)}$ so that the mth molecule of a codeword is chosen only from $\mathcal{B}_m^{(L)}$. The inner code $\mathcal{B}^{(L)}$ thus also protects the index from sequencing errors. An outer code then specifies the valid sets of molecules.

Our proposed decoder is based on a decoder for the inner code $\mathcal{B}^{(L)}$, which is used to *independently* decode each of the N sequenced molecules to a sequence in $\mathcal{B}^{(L)}$. Since the decoder operates on a molecule-by-molecule basis, future design of codes based on this scheme is a feasible goal (L is typically on the order of $10^2 - 10^3$), and is much simpler than the decoder of [4] (the clustering-based decoder [12] also has a low complexity of $\Theta(N)$, but there are no guarantees on the decay rate of the error probability). A decoder for the outer code is then used to resolve molecule erasures and undetected errors. Hence, the proposed coded-index based scheme is practically oriented, and its analysis is general, in the sense that very little is assumed on the sequencing channel. It is only required that a decoder for the inner code exists whose error probability decays to zero with increasing L. This addresses the first two issues raised above.

Regarding the third issue, as explained in [4], for fixed coverage depth ($N = \alpha M$ for some fixed $\alpha > 1$) the slow $e^{-\Theta(M)}$ decay rate of the error probability is the result of molecule errors (erasures), rather than sequencing errors. So, apparently, faster decay rate is only possible by increasing N. In accordance, we consider in this paper the scaling $N = \alpha_M M$, where α_M is (possibly) an increasing function of M (though rather slowly). Our main result is a single-letter upper bound on the error probability which decays as $e^{-\Theta(N)}$, achieved by a coded-index based scheme. An important consequence of this result is that operating at a large coverage depth N/M is of importance for low error probability. This is in opposed to capacity analysis, for which large N/M only provides marginal capacity gains [1, Sec. I]. We remark that our scheme is not capacity achieving under the DMC and fixed α model studied in [1]–[4], as it does not exploit multiple observations of the same molecule to increase the rate. However, the rate loss is small for sequencing channels which are fairly clean, as multiple observations only marginally increase the capacity in this case. Anyway, adapting our scheme to achieve capacity is an important open problem.

Previously, [12] has considered an (explicit) coded-indexing and concatenated coding scheme, whose decoder is based on (hard) output clustering, and so is mainly tailored to the binary symmetric channel. As described above, we consider here general sequencing channels, and focus on error probability analysis and simple decoding (see [14] for a detailed comparison with this, as well as with additional related work [15]). In our context, the conclusion is that the loss is more profound for small β. The rest of the paper is organized as follows. In Sec. II we establish notation conventions, formulate the DNA storage channel and coded-index based systems. In Sec. III we state our main result, and in Sec. IV we outline the proof. All proofs and further results and discussions are available in a full version of the paper [14]).

II. Problem Formulation

We begin with notation conventions. 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 super-scripted by their dimension. The probability of the event $\mathcal{E}$ will be denoted by $\mathbb{P}(\mathcal{E})$, and its indicator function will be denoted by $\mathbb{1}(\mathcal{E})$. The expectation operator will be denoted by $\mathbb{E}[\cdot]$. Logarithms and exponents will be understood to be taken to the natural base. The binary Kullback–Leibler (KL) divergence d_b: $[0,1]\times(0,1)\to\mathbb{R}^+$ by $d_b(a||b) := a\log\frac{a}{b} + (1-a)\log\frac{(1-a)}{(1-b)}$. The number of *distinct* elements of a finite multiset $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. The equivalence relation will be denoted by $\equiv$, and will mainly be used to simplify notation. Asymptotic Bachmann–Landau notation will be used. For a positive integer N we will denote $[N] := \{0, 1, \ldots, N-1\}$, where scalar multiplications of these sets will be used, e.g., as $\frac{1}{N}[N+1] = \{0, \frac{1}{N}, \ldots \frac{N-1}{N}, 1\}$.

Next, we formulate a sequence of channels, encoders and decoders for the DNA storage channel, indexed by M, the number of molecules in a codeword.

The channel model (reading mechanism): A DNA molecule is a sequence of $L \equiv L_M \in \mathbb{N}_+$ nucleotides (symbols) chosen from an arbitrary alphabet $\mathcal{X}$ (in physical systems $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, and in some previous works [1]–[3] a binary alphabet $\mathcal{X} = \{0, 1\}$ was assumed for simplicity). Thus, each molecule is uniquely represented by a sequence $x^L \in \mathcal{X}^L$. An input to the DNA channel is a sequence of M molecules, $x^{LM} = (x_0^L, \ldots x_{M-1}^L)$ where $x_m^L \in \mathcal{X}^L$ for all $m \in [M]$. A message is synthesized into a sequence of M molecules, x^{LM}. The DNA storage channel model is determined by the number of molecule samples $N \equiv N_M \in \mathbb{N}_+$, and by the sequencing channel $W^{(L)}\colon \mathcal{X}^L \to \mathcal{Y}^L$. The operation of the channel on the stored codeword is modeled as a two-stage process:

1) Sampling: N molecules are sampled uniformly from the M molecules of x^{LM}, independently, with replacement. Let $U^N \sim$ Uniform$([M]^N)$ be such that U_n is the sampled molecule at sampling event $n \in [N]$. The result of the sampling stage is the vector $(x_{U_0}^L, x_{U_1}^L, \ldots, x_{U_{N-1}}^L) \in (\mathcal{X}^L)^N$. We also denote by S_m the number of times that molecule m was sampled, to wit $S_m = \sum_{n\in[N]} \mathbb{1}\{U_n = m\}$, the empirical count of U^N. It then holds that $S^M = (S_0, \ldots, S_{M-1}) \sim$ Multinomial$(N; (\frac{1}{M}, \frac{1}{M}, \ldots \frac{1}{M}))$.

2) Sequencing: For each $n \in [N]$, $x_{U_n}^L$ is sequenced to $Y_n^L \in \mathcal{Y}^L$, where the sequencing of $x_{U_n}^L$ is independent for all $n \in [N]$. Denoting the channel output by $Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L) \in (\mathcal{Y}^L)^N$ it thus holds that

$$\mathbb{P}\left[Y^{LN} = y^{LN} \mid x^{LM},\, U^N\right] = \prod_{n\in[N]} W^{(L)}\left(y_n^L \mid x_{U_n}^L\right). \tag{1}$$

We make the following assumptions on the channel: (1) $L \equiv L_M = \beta \log M$ where $\beta > 1$ is the *molecule length parameter*. (2) N/M where $\alpha \equiv \alpha_M > 1$ is the *coverage depth scaling function*.

The encoder: A codebook is a set of different possible codewords $\mathcal{C} = \{x^{LM}(j)\}$. We propose the following restricted set of *coded-index based codebooks*:

Definition 1. Let $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$ be a collection of pairwise disjoint sets $\mathcal{B}_m^{(L)} \subset \mathcal{X}^L$ of equal cardinality, and let $\mathcal{B}^{(L)} := \cup_{m\in[M]}\mathcal{B}_m^{(L)}$ be their union. A DNA storage codebook is said to be a *coded-index* based codebook if $x_m^L(j) \in \mathcal{B}_m^{(L)}$ for all $m \in [M]$ and all $j \in [|\mathcal{C}|]$.

To wit, a codeword contains exactly a single molecule from each of the M sets $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$. The identity of the set from which $x_m^L(j)$ was chosen from is considered an “index” of the molecule that is used by the decoder to order the molecules that has been decoded. A coded-index based codebook, can be thought of as a concatenated code. The set $\mathcal{B}^{(L)}$ is an inner-code, which is used to clean the output molecules from sequencing errors, and the dependency between molecules of different index m can be considered an outer-code which is used to cope with erasures (mainly due to the sampling stage).

The decoder: A general decoder is a mapping $\mathsf{D}\colon (\mathcal{Y}^L)^N \to [|\mathcal{C}|]$. We propose the following class of decoders, which are suitable for coded-index based codebooks. A decoder from this class is equipped with an inner-code decoder $\mathsf{D}_b\colon \mathcal{Y}^L \to \mathcal{B}^{(L)}$, and a threshold $T \in \mathbb{R}^+$, and decodes the channel output y^{LN} in three steps:

1) Correction of individual molecules: The decoder employs the inner-code decoder for each of the received molecules y_n^L, for each $n \in [N]$, and set $z_n^L = \mathsf{D}_b(y_n^L)$. Following this stage, it holds that $z^{LN} = (z_0^L, \dots, z_{N-1}^L)$ is such that $z_n^L \in \mathcal{B}^{(L)}$ for all $n \in [N]$.

2) Threshold for each index: For each index $m \in [M]$, if there exists a $b^L \in \mathcal{B}_m^{(L)}$ such that

$$\sum_{n\in[N]} \mathbb{1}\{z_n^L = b^L\} \geq T > \max_{\tilde{b}^L \in \mathcal{B}_m^{(L)} \setminus \{b_l^L\}} \sum_{n\in[N]} \mathbb{1}\{z_n^L = \tilde{b}^L\} \tag{2}$$

then the decoder sets $\hat{x}_m^L = b^L$. That is, $\hat{x}_m^L = b^L$ if b^L is a unique molecule in $\mathcal{B}_m^{(L)}$ whose number of appearances in z^{LN} is larger than T. Otherwise $\hat{x}_m^L = \mathsf{e}$, where e is a symbol representing an *erasure*.

3) Codeword decoding: Let

$$j^* = \arg\min_{j\in[|\mathcal{C}|]} \rho(\hat{x}^{LM}, x^{LM}(j)) \tag{3}$$

where (with a slight abuse of notation)

$$\rho(\hat{x}^L, x^L) := \begin{cases} \mathbb{1}\{\hat{x}^L \neq x^L\}, & \hat{x}^L \neq \mathsf{e} \\ 0, & \hat{x}^L = \mathsf{e} \end{cases} \tag{4}$$

and $\rho(\hat{x}^{LM}, x^{LM}) := \sum_{m\in[M]} \rho(\hat{x}_m^L, x_m^L)$, which is a Hamming distance with zero contribution in case of erasures.

The DNA storage channel is thus indexed by M and parameterized by $(\alpha_M, \beta, \{W^{(L)}\}_{L\in\mathbb{N}_+})$. The (storage) rate of the codebook $\mathcal{C}$ is given by $R = \frac{\log|\mathcal{C}|}{ML}$, and the error probability of D given that $x^{LM}(j) \in \mathcal{C}$ was stored is given by

$$\mathsf{pe}(\mathcal{C}, \mathsf{D} \mid x^{LM}(j)) := \mathbb{P}\left[\mathsf{D}(y) \neq j \mid x^{LM}(j)\right]. \tag{5}$$

Let $\psi_M\colon \mathbb{N}_+ \to \mathbb{N}_+$ be a monotonic strictly increasing sequence. An *error exponent* $E(R)$ *w.r.t. scaling* ψ_M is achievable for channel DNA at rate R, if there exists a sequence $\{\mathcal{C}_M, \mathsf{D}_M\}_{M\in\mathbb{N}_+}$ so that the average error probability is bounded as

$$-\log\left[\frac{1}{|\mathcal{C}_M|}\sum_{j\in[|\mathcal{C}_M|]} \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M \mid x^{LM}(j))\right] \geq \psi_M \cdot E(R) - o(\psi_M). \tag{6}$$

In this paper, we will obtain single-letter expressions for error exponents achieved under coded-index codebook and the class of decoders defined above. Throughout, we only make the following assumptions on the sequencing channel: 1) Inner code rate: $R_b := \frac{1}{L}\log|\mathcal{B}^{(L)}| > 1/\beta$. 2) Vanishing inner code (maximal) error probability:

$$\mathsf{pe}_b(\mathcal{B}^{(L)}) := \max_{b^L\in\mathcal{B}^{(L)}} W^{(L)}\left[\mathsf{D}_b(y^L) \neq b^L \mid b^L\right] = o(1). \tag{7}$$

As $|\mathcal{B}_m^{(L)}| = \frac{e^{R_b L}}{M} = \frac{\exp[R_b\beta\log M]}{\exp[\log M]}$, the assumption on R_b assures that $\mathcal{B}_m^{(L)}$ is not empty. The assumption on the error probability assures that the error probability at the first decoding step tends to zero as $L = \beta\log M \to \infty$. Thus, if the sequencing channel $W^{(L)}$ has capacity $C(W^{(L)})$ (with rate normalized to single symbol), then it must hold that $R_b \leq C(W^{(L)})$. For example, for sequencing DMC, the error probability decays as $e^{-E(R_b)\cdot L}$, where $E(R_b)$ is the error exponent. For general sequencing channels, the decay rate could be slower even for optimal codes. Thus, for concreteness, we set $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^\zeta)}$, where $\zeta > 0$, and as we shall see, ζ will not affect the achievable exponent of the DNA storage system. Therefore, even sub-optimal codes can be used, for example, *polar codes*, whose error scales as $e^{-\Theta(\sqrt{N})}$ for standard DMCs [16], and of $e^{-\Theta(N^{1/3})}$ for channel which include insertions, deletions, and substitutions [17].

Our achievable error exponent will be based on the following *coded-index based random coding ensemble*:

Definition 2. Following Definition 1, let $\mathcal{C} = \{X^{LM}(j)\}$ be a random coded such that $X_m^L(j)$ is chosen uniformly at random from $\mathcal{B}_m^{(L)}$ independently for all $m \in [M]$ and all $j \in [|\mathcal{C}|]$.

III. Main Result

Our main result is as follows:

Theorem 3. *Let an inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$, and let D_b be a decoder which satisfy the assumptions on the inner code ($R_b > 1/\beta$, $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^\zeta)}$). Then, there exists a sequence of codebooks $\{\mathcal{C}_M\}$ and corresponding threshold-based decoders $\{\mathsf{D}_M\}$ (as described in Sec. II) so that the following holds: If $N/M = \Theta(1)$ then for any $R < (R_b - 1/\beta)(1 - e^{-\frac{N}{M}})$,*

$$-\log\mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq M \cdot d_b\left(1 - \frac{R}{R_b - 1/\beta} \,\middle|\middle|\, e^{-\frac{N}{M}}\right) - O\left(\frac{M}{\log M}\right). \tag{8}$$

If $N/M = \omega(1)$ then for any $R < R_b - 1/\beta$,

$$-\log\mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq \frac{N}{2}\left[1 - \frac{R}{R_b - 1/\beta}\right] - O(M) \tag{9}$$

if $\frac{N}{ML} < 2(R_b - 1/\beta)$, and

$$-\log\mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq ML\left[R_b - 1/\beta - R\right] - O\left(\frac{N}{\log M}\right) \tag{10}$$

if $2(R_b - 1/\beta) \leq \frac{N}{ML}$.

Discussion:

1) The bound is not continuous in N (that is, there is a phase transition), and a the behavior is different between $N=\Theta(M)$ and $N=\omega(M)$. As stems from the proof, in both regimes, the threshold is chosen as $T\equiv T_M=o(M)$. This follows since the error probability of the inner code is $e^{-\Theta(L^{\zeta})}=e^{-\Theta(\log^{\zeta} M)}$, and so the number of erroneously sequenced molecules is $o(M)$, with an average of less than a single erroneous molecule per index.

2) The result does not depend on ζ, the assumed scaling of the inner code error probability ($\mathsf{pe}_b(\mathcal{B}^{(L)})=e^{-\Theta(L^{\zeta})}$), and manifests the fact that sampling events dominate the error probability, compared to sequencing error events.

3) For the standard channel coding problem over DMCs with blocklength N, the method of types leads to random coding and expurgated bounds which tend to their asymptotic values up to a $O((\log N)/N)$ term (this can be avoided for Gallager's method [18, Ch. 5], see also [19, Problem 10.33]). Here, it is evident that the decay is much slower, and could be as slow as $O(1/\log M)$. As discussed in [4, Sec. VII] this seems to be an inherent property of this channel.

4) Proving tightness of Theorem 3 is challenging, even for optimal decoders. The main difficulty is in the *Poissonization of the multinomial* effect which is used to upper bound the large-deviations behavior of the number of under-sampled number of molecules in Lemma 4 to follow (as proposed in [1], [12]). This upper bound is tight at the center of the multinomial distribution, but may be loose at its tails. Developing lower bounds on the error probability is thus an open problem.

5) An expurgated bound is also proved in [14], which improves the error probability at the regime $\frac{N}{ML}>4(R_b-1/\beta)$.

IV. Main Steps of the Proof

The proof begins by analyzing the probability of channel-related events, and specifically, the event in which some of the molecules are not sampled enough times, or the event of excessive number of sequencing errors. Let the threshold $T\equiv T_\tau:=\frac{N}{M}(1-\sqrt{2\tau})$ of the decoder D be parameterized by a parameter $\tau\in(0,1/2)$. In coded-index based coding, each codeword $x^{LM}(j)$ contains exactly a single molecule from each of the sub-codes $\mathcal{B}_m^{(L)}$, and the molecule $x_m^L(j)$ is sampled S_m times. Let $K_m\in[S_m+1]$ be the number of copies of the molecule $x_m^L(j)$ that have been erroneously sequenced, let $K:=\sum_{m\in[M]}K_m\in[N+1]$ be the total number of molecules which have been erroneously sequenced, and let $V_m\in[K+1]$ be the number of molecules $x_{m'}^L(j)$ for $m'\in[M]\backslash\{m\}$ which have been erroneously sequenced to have index m. Note that $\sum_{m\in[M]}V_m=K$ holds too. The event in which the molecule x_m^L was not decoded correctly in the second stage of the operation of the decoder is included in a union of the following events:

1) $S_m<T_\tau$, that is, the molecule have not been sampled enough times in the sampling stage.

2) $S_m\geq T_\tau$ yet $S_m-K_m<T_\tau$, that is, the molecule have been sampled enough times in the sampling stage step, but K_m sequencing errors have caused the number of appearances of $x_m^L(j)$ to drop below the threshold T.

3) $V_m\geq T_\tau$, that is, there are more than T molecules with index m, which are not the correct molecule $x_m^L(j)$.

On the face of it, the event $V_m\geq T_\tau$ can lead to a crude upper bound, since the V_m molecules which are erroneously mapped to index m are not likely to be the *exact* same molecule in $\mathcal{B}_m^{(L)}$. However, a more precise analysis of this event would require making assumptions on the structure of the sub-codes $\{\mathcal{B}_m^{(L)}\}$, which we avoid here altogether.

Corresponding to these events, we define the following sets:

$$\mathcal{M}_\sigma:=\{m\in[M]\colon S_m<T_\tau\} \tag{11}$$

$$\mathcal{M}_\kappa:=\{m\in[M]\colon S_m\geq T_\tau,\ S_m-K_m<T_\tau\} \tag{12}$$

$$\mathcal{M}_\nu:=\{m\in[M]\colon V_m\geq T_\tau\}, \tag{13}$$

The next lemma addresses the cardinality of $\mathcal{M}_\sigma$:

Lemma 4. *Let $x^{LM}(j)$ be a codeword from a coded-index codebook. Let $\tilde{S}\sim\text{Pois}(N/M)$ and*

$$\varphi_\tau:=-\frac{1}{N/M}\log\mathbb{P}\left[\tilde{S}\leq T_\tau\right]. \tag{14}$$

If $N/M=\Theta(1)$ then

$$\mathbb{P}\left[|\mathcal{M}_\sigma|\geq\sigma M\mid x^{LM}(j)\right]\leq 3\cdot\exp\left[-M\cdot d_b\left(\sigma\,\middle|\middle|\,e^{-\varphi_\tau\frac{N}{M}}\right)\right] \tag{15}$$

for $\sigma\in(e^{-\varphi_\tau\frac{N}{M}},1]$. If $N/M=\omega(1)$ then

$$\mathbb{P}\left[|\mathcal{M}_\sigma|\geq\sigma M\mid x^{LM}(j)\right]\leq 4e^{-\sigma\tau N} \tag{16}$$

for $\sigma\in(e^{-\tau\frac{N}{M}},1]$.

Proof outline: The empirical count vector S^M follows a multinomial distribution, whose components are dependent. The proof utilizes the *Poissonization* of the multinomial distribution effect [20, Thm. 5.6]: If $\tilde{N}\sim\text{Pois}(\lambda)$ and $\tilde{S}^M\sim\text{Multinomial}(\tilde{N},(\frac{1}{M},\ldots,\frac{1}{M}))$ conditioned on $\tilde{N}$, then $\{\tilde{S}_m\}_{m\in[M]}$ are independent and identically distributed (i.i.d.) $\tilde{S}_m\sim\text{Pois}(\frac{\lambda}{M})$ (unconditioned on $\tilde{N}$). Let $A\equiv\sum_{m\in[M]}\mathbb{1}\{S_m<T_\tau\}$ and $\tilde{A}\equiv\sum_{m\in[M]}\mathbb{1}\{\tilde{S}_m<T_\tau\}$. The Poissonization effect is used to prove the upper bound (see also [20, Exercise 5.14])

$$\mathbb{P}\left[A\geq\sigma M\right]\leq 2\cdot(1+o(1))\cdot\mathbb{P}\left[\tilde{A}\geq\sigma M\right], \tag{17}$$

and as $\tilde{A}$ is a sum of i.i.d. random variables $\{\tilde{S}_m\}$, the right probability is then evaluated by a standard Chernoff bound on the binomial distribution. ■

The following lemma is used to bound the total number of sequencing errors K, which, in turn, is used to bound the cardinalities of $\mathcal{M}_\kappa$ and $\mathcal{M}_\nu$:

Lemma 5. *Let K be the total number of erroneously sequenced molecules. Let $\mathcal{U}\subseteq[M]^N$ be a sampling event, and assume that $\mathsf{pe}_b(\mathcal{B}^{(L)})=e^{-c\cdot L^{\zeta}}$. Then, $\mathbb{P}[K\geq\kappa N\mid\mathcal{U}]\leq e^{-c\cdot\kappa NL^{\zeta}}$ for any $\kappa\in(0,1]$.*

Proof outline: The proof is based on a Chernoff bound over the N independent sequencing operations, for which the probability of error is at most $e^{-c\cdot L^{\zeta}}$. It requires, however, a more refined argument, since the sequencing errors are not be independent for a given codebook $\mathcal{B}^{(L)}$. ■

The channel/decoder operation is more directly defined by the set of erased molecules and the set of molecules with undetected errors as

$$\mathcal{M}_{\mathsf{e}} := \left\{m \in [M]\colon \hat{x}_m^L = \mathsf{e}\right\}, \tag{18}$$

$$\mathcal{M}_{\mathsf{u}} := \left\{m \in [M]\colon \hat{x}_m^L \neq \mathsf{e},\ \hat{x}_m^L \neq x_m^L(j)\right\}. \tag{19}$$

Lemmas 4, and 5 are utilized to analyze the cardinality of $\mathcal{M}_{\mathsf{e}}$ and $\mathcal{M}_{\mathsf{u}}$. As it turns out, the dominating event is $\mathbb{P}[|\mathcal{M}_{\sigma}|\geq \sigma M]$, to wit, the probability that the molecules have not been amplified enough times, which is on the exponential order of N, compared to the probability evaluated in Lemma 5 which are on the exponential order of $LN = N\beta \log M$.

Lemma 6. *Consider a decoder* D *for a coded-index based codebook. For the erasure set* $\mathcal{M}_{\mathsf{e}}$*: If* $N/M = \Theta(1)$ *then*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{e}}|\geq \theta M\right] \geq M d_b\left(\theta || e^{-\varphi_\tau \frac{N}{M}}\right) + o(M) \tag{20}$$

for all $\theta \in (e^{-\varphi_\tau \frac{N}{M}}, 1]$. *If* $N/M = \omega(1)$ *then*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{e}}|\geq \theta M\right] \geq \theta \tau N \cdot [1 + o(1)] \tag{21}$$

for all $\theta \in (e^{-\tau \frac{N}{M}}, 1]$. *For the undetected error set* $\mathcal{M}_{\mathsf{u}}$*:*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{u}}|\geq \theta M\right] \geq c \cdot (1 - \sqrt{2\tau})\theta N L^{\zeta}. \tag{22}$$

Proof outline: By deriving relations between K and $|\mathcal{M}_{\kappa}|, |\mathcal{M}_{\nu}|$, and then between these sets and $|\mathcal{M}_{\sigma}|$, to $|\mathcal{M}_{\mathsf{e}}|$ and $|\mathcal{M}_{\mathsf{u}}|$, and utilizing Lemmas 4 and 5. ■

Thus, as apparent from Lemma 6, and as discussed in the introduction, for coded-index based codebooks, the type of decoders, and the analysis in this paper, the effect of sequencing errors is much less profound compared to erasures.

The random coding analysis is based on the following lemma, which bounds the probability that an erroneous codeword will be decoded, conditioned on a given number of channel erasures and undetected errors.

Lemma 7. *Let* $\mathcal{C}$ *be drawn from the coded-index based random coding ensemble. Let* $X^{LM}(0) = x^{LM}(0)$ *be arbitrary, and let* $\hat{X}^{LM}$ *be the output of the decoder conditioned on the input* $x^{LM}(0)$*. Then, for* $\theta_{\mathsf{e}}, \theta_{\mathsf{u}} \in \frac{1}{M}[M+1]$ *such that* $\theta_{\mathsf{e}} + \theta_{\mathsf{u}} \leq 1$ *and any* $j \in [|\mathcal{C}|]\backslash\{0\}$ *it holds that*

$$\begin{aligned} -\frac{1}{M}\log \mathbb{P}\Bigg[&\rho(\hat{X}^{LM}, X^{LM}(j)) \leq \rho(\hat{X}^{LM}, x^{LM}(0)) \\ &\Bigg| |\mathcal{M}_{\mathsf{e}}|= \theta_{\mathsf{e}} M,\ |\mathcal{M}_{\mathsf{u}}|= \theta_{\mathsf{u}} M\Bigg] \\ &\geq (R_b\beta - 1)(1 - \theta_{\mathsf{e}} - \theta_{\mathsf{u}})\log M - \Theta(1). \end{aligned} \tag{23}$$

Proof outline: The proof is based on an argument which counts the relative number of competing codewords $\tilde{x}^{LM}$ in the coded-index based ensemble which have distance $\rho(\hat{X}^{LM}, \tilde{x}^{LM})$ smaller than $\rho(\hat{X}^{LM}, x^{LM}(0))$, followed by an analysis of its asymptotic behavior with M. ■

The proof of Theorem 3 then follows from Lemma 7, by conditioning over $\theta_{\mathsf{e}}, \theta_{\mathsf{u}}$, taking a clipped union bound over the probability that one of the $\lceil e^{MLR}\rceil - 1$ competing codewords causes an error, averaging over $\theta_{\mathsf{e}}, \theta_{\mathsf{u}}$ via Lemma 6, and analyzing the asymptotic behavior of the resulting expressions for the different regimes of $\alpha_M = N/M$.

References

[1] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3675–3689, 2021.

[2] A. Lenz, L. Welter, and S. Puchinger, "Achievable rates of concatenated codes in DNA storage under substitution errors," in *International Symposium on Information Theory and Its Applications*, pp. 269–273, IEEE, 2020.

[3] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *IEEE Information Theory Workshop*, pp. 1–5, IEEE, 2019.

[4] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," 2021. Available at https://arxiv.org/pdf/2109.12549.pdf.

[5] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[6] K. A. S. Immink and K. Cai, "Design of capacity-approaching constrained codes for DNA-based storage systems," *IEEE Communications Letters*, vol. 22, no. 2, pp. 224–227, 2017.

[7] Y. Wang, M. Noor-A-Rahim, E. Gunawan, Y. L. Guan, and C. L. Poh, "Construction of bio-constrained code for DNA data storage," *IEEE Communications Letters*, vol. 23, no. 6, pp. 963–966, 2019.

[8] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. S., "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[9] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.

[10] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[11] L. Organick, S. D. Ang, Y. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, and B. Nguyen, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[12] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the DNA storage channel," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8846–8850, IEEE, 2020.

[13] L. C. Meiser, P. L. Antkowiak, W. D. Koch, J.and Chen, A. X. Kohll, W. J. Stark, R. Heckel, and R. Grass, "Reading and writing digital data in DNA," *Nature protocols*, vol. 15, no. 1, pp. 86–101, 2020.

[14] N. Weinberger, "Error probability bounds for coded-index DNA storage channels," 2021. Available at https://drive.google.com/file/d/1tuEGj4852slCPvNq6xteTgZz8TgCx1ME/view?usp=sharing.

[15] M. Kovačević and V. Y. F. Tan, "Codes in the space of multisets – coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, 2018.

[16] S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke, "Rate-dependent analysis of the asymptotic behavior of channel polarization," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2267–2276, 2012.

[17] I. Tal, H. D. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 1362–1366, IEEE, 2019.

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

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

[20] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

Error Correction for FrodoKEM Using the Gosset Lattice

Charbel Saliba and Laura Luzzi
ETIS, UMR 8051,
CY Université, ENSEA, CNRS,
Cergy, France
Email: {charbel.saliba, laura.luzzi}@ensea.fr

Cong Ling
Department of Electrical
and Electronic Engineering
Imperial College London, U.K.
Email: cling@ieee.org

***Abstract*—We consider FrodoKEM, a lattice-based cryptosystem based on LWE, and propose a new error correction mechanism to improve its performance. Our encoder maps the secret key block-wise into the Gosset lattice E_8. We propose three sets of parameters for our modified implementation. Thanks to the improved error correction, the first implementation allows to reduce the bandwidth by 7% by halving the modulus q; the second outperforms FrodoKEM in terms of plausible security by 10 to 13 bits by increasing the error variance; and the third one allows to increase the key size. In all cases, the decryption failure probability is improved compared to the original FrodoKEM. Unlike some previous works on error correction for lattice-based protocols, we provide a rigorous error probability bound by decomposing the error matrix into blocks with independent error coefficients.**

I. Introduction

Quantum computers pose a threat since they are capable of breaking most of the cryptographic systems currently in use. Post-quantum cryptography refers to cryptographic algorithms believed to be secure against a cryptanalytic attack by a quantum computer. Lattice-based cryptographic constructions are particularly promising candidates for post-quantum cryptography because they offer strong theoretical security guarantees and can be implemented efficiently. Therefore, lattice-based cryptosystems are considered a safe avenue for replacing the currently used schemes based on RSA and the discrete logarithm. As of now, NIST is assessing and standardizing PQC algorithms. In the third round submissions, three of the four finalists in the public-key encryption and key-establishment algorithms are lattice-based schemes, along with the majority of the alternate candidates.

One of the most widely used cryptographic primitives based on lattices is the Learning With Errors problem (LWE), introduced by Regev [1], who proved a worst-case to average-case reduction from the shortest independent vector problem (SIVP) to LWE. It can be used to build a variety of cryptographic algorithms and provides guarantees in terms of IND-CPA and IND-CCA security. Later works introduced structured variants of LWE such as Ring-LWE [2] and Module-LWE [3] which involve ideal lattices and module lattices respectively. Their cryptographic applications are generally more efficient compared to LWE. However, in principle the additional algebraic structure might make these variants more vulnerable to attacks. Although currently there are no specific known attacks targeting Ring-LWE or Module-LWE, much progress has been made in recent works to exploit the structure of ideal lattices and module lattices to solve lattice problems [4]–[6]. Thus, although the Module-LWE based scheme Kyber [7] was selected as a finalist for the NIST PQC standardization Round 3, the plain-LWE scheme FrodoKEM [8] was selected as an alternate candidate which may provide longer-term security guarantees since it is less susceptible to algebraic attacks. From the NIST's perspective, although FrodoKEM can be used in the event that new cryptanalytic results targeting structured lattices emerge, the first priority for standardization is a KEM that would have acceptable performance across widely used applications.

In this paper, we aim at improving the bandwidth efficiency and/or security of FrodoKEM, or at increasing the key size, through an enhanced error correction mechanism. We note that although the current security estimate for FrodoKEM against known attacks is greater than the brute-force security (except for Frodo-1344), the plausible security [8, Section 5.2], which takes into account possible improvements in sieving algorithms, is not. Improving the plausible security would give FrodoKEM better guarantees for long-term security.

A modification of FrodoKEM has been proposed in [9] using Gray labeling and error correcting codes in order to improve the performance. However, the decryption failure analysis in [9] assumes that the coefficients of the error are independent. Unfortunately this assumption does not hold for FrodoKEM, and as shown in [10], it can lead to underestimating the decryption failure by a large exponential factor.

In this work, we propose a different approach where enhanced error correction is obtained through lattice encoding and decoding rather than using error-correcting codes. More precisely, our encoder maps the secret key block-wise into the Gosset lattice E_8. Lattice codes were used in previous works for Ring-LWE based cryptosystems, such as the reconciliation mechanism based on the $\tilde{D}_4$ lattice for NewHope [11]. Due to its optimal density and low-complexity quantization, the E_8 lattice was already used in KCL [12], a first round NIST candidate. In our previous work [13], the E_8 lattice was employed to improve the security of the Module-LWE based candidate KyberKEM.

The choice of an 8-dimensional lattice encoder is well-suited to the parameters of FrodoKEM. In fact, due to its particular structure, the error matrix can be decomposed into 8 blocks of 8 independent components, which makes a rigorous decryption error analysis possible. The encryption function used by the original FrodoKEM implicitly uses the cubic lattice $\mathbb{Z}^{64} \cong \left(\mathbb{Z}^8\right)^8$. Accordingly, switching from $\mathbb{Z}^8$ to E_8 allows us to improve the security or bandwidth. We propose three sets of parameters for our modified implementation. Thanks to the improved error correction, the first implementation allows to reduce the bandwidth by 7% by halving the modulus q, the second improves the security level by 10-13 bits by increasing the error variance, and the third allows to generate 192 bits from Frodo-640 instead of 128 bits, as well as 256-bit key instead of 192 bits in Frodo-976, with comparable security and error probability.

Organization: This paper is organized as follows. In section II, we provide essential mathematical and cryptographic background for our work, then we develop the proposed modification for FrodoKEM in section III. Section IV gives an upper bound for the decryption error probability for our algorithm, while section V derives its security analysis. In the last section, we show the improvements made with regard to security, bandwidth and key size.

II. Notation and preliminaries

a) Notation: Given a set $A \subseteq \mathbb{R}^n$, $|A|$ stands for its cardinality. All vectors and matrices are denoted in bold. The function $\operatorname{sign}(\cdot)$ outputs 1 for positive real input (including zero) and -1 for strictly negative one. For $\mathbf{x} \in \mathbb{R}^n$ we denote $\lfloor \mathbf{x} \rceil$ to be the rounding function of each component of $\mathbf{x}$, where $\pm 1/2$ is rounded to 0. We also denote $\rfloor \mathbf{x} \lceil$ to be the same as $\lfloor \mathbf{x} \rceil$ except that the worst component of $\mathbf{x}$ - that furthest from an integer - is rounded the wrong way. More formally, if $i_0 = \underset{i}{\operatorname{argmax}} |x_i - \lfloor x_i \rceil|$, then $\rfloor \mathbf{x} \lceil_i = \lfloor x_i \rceil + \operatorname{sign}(x_i) \cdot \operatorname{sign}(|x_i| - \lfloor |x_i| \rceil)$ if $i = i_0$ and $\rfloor \mathbf{x} \lceil_i = \lfloor x_i \rceil$ if not. A constant vector $(\alpha, \dots, \alpha) \in \mathbb{R}^n$ is denoted by $\boldsymbol{\alpha}$. For $a, b \in \mathbb{Z}$, the operation $(a+b) \bmod 2$ is simplified to $a \oplus b$.

b) Lattice definitions and properties: An n-dimensional lattice Λ is a discrete subgroup of $\mathbb{R}^n$ that can be defined as the set of integer linear combinations of n linearly independent vectors, called *basis vectors*. The closest lattice point to $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\mathrm{CVP}_\Lambda(\mathbf{x})$, and the *Voronoi region* $\mathcal{V}(\Lambda)$ is the set of all points $\mathbf{x} \in \mathbb{R}^n$ for which $\mathrm{CVP}_\Lambda(\mathbf{x}) = \mathbf{0}$. The *volume* of a lattice, which is a lattice constant, is defined to be the volume of its Voronoi region. The *Voronoi relevant vectors* of Λ are the vectors $\lambda \in \Lambda$ such that $\langle \mathbf{x}, \lambda \rangle < \|\mathbf{x}\|^2$ for all $\mathbf{x} \in \Lambda \setminus \{0, \lambda\}$. The minimal distance of the lattice is defined as $\lambda_1(\Lambda) := \min_{\mathbf{v} \in \Lambda \setminus \{0\}} ||\mathbf{v}||$.

c) The Gosset lattice: We introduce the 8-dimensional lattice E_8 [14, p.121] which will be used throughout this paper. This lattice has a unit volume, and is generated by the rows of the matrix

$$\mathbf{G}_{E_8} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \end{bmatrix}$$

The Voronoi relevant vectors of E_8 form two sets: VR_1 which contains the first type of the form $(\pm 1^2, 0^6)$, and VR_2 which contains $(\pm 0.5^8)$ as the second type. Note that $|\mathrm{VR}_1| = 112$ and $|\mathrm{VR}_2| = 128$, so that the total number of Voronoi relevant vectors is 240.

d) Error distribution: The error distribution required for the LWE problem defined in the next section is ideally a Gaussian-like distribution. Let $D_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\|x\|^2/2\sigma^2\right)$ denotes the probability density function of a zero-mean continuous Gaussian distribution with variance σ. A rounded Gaussian distribution Ψ_σ is obtained by rounding a sample from D_σ to the nearest integer.

As in the FrodoKEM specifications [8], we use a discrete and symmetric distribution χ on $\mathbb{Z}$, centered at zero and with finite support $\{-s, \dots, s\}$, which approximates a rounded Gaussian distribution. In our case, χ is generated for different values of σ and the support $\{-s, \dots, s\}$ depends on the chosen σ value. In a more detailed manner, given the target standard deviation σ, we first construct a function $\tilde{\chi}$ on $\{-s, \dots, s\} \subseteq \mathbb{Z}$ as follows:

$$\forall i \in \{-s, \dots, s\},\ \tilde{\chi}(i) = \frac{1}{2^{16}} \left\lfloor 2^{16} \cdot \int_{\left[i-\frac{1}{2}, i+\frac{1}{2}\right]} D_\sigma(x) dx \right\rceil .$$

The distribution χ is obtained from $\tilde{\chi}$ by making small changes in the numerator values of $\tilde{\chi}(i)$ in order to obtain a probability distribution (the whole sum ends up to be 1) The sampling algorithm for such a distribution is given in [8, Algorithm 5], and it is resistant to cache and timing side-channels. The distance between Ψ_σ and χ is measured according to the Rényi divergence, which indicates how far a discrete distribution P is from another distribution Q. More formally, for a given positive order $\alpha \neq 1$, the Rényi divergence between P and Q is defined as

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \ln \left(\sum_{x \in \mathrm{supp} P} P(x) \left(\frac{P(x)}{Q(x)} \right)^{\alpha - 1} \right).$$

The Rényi divergence can be used to relate the probabilities of an event according to P or Q [8, Lemma 5.5]. This justifies why replacing the rounded Gaussian with a distribution which is close in Rényi divergence will preserve the security reductions [8, Corollary 5.6]. We use the script `scripts/Renyi.py` in [11] to compute the Rényi divergence between our chosen distribution χ and the rounded Gaussian.

e) LWE problem: The security of FrodoKEM and our modified version is based on the hardness of the LWE problem [1]. Let n and q be positive integers, and χ an error distribution over $\mathbb{Z}$. Take $\mathbf{s}$ to be a uniform vector in $\mathbb{Z}_q^n$. The problem consists in distinguishing uniform samples $(\mathbf{a}, b) \leftarrow \mathbb{Z}_q^n \times \mathbb{Z}_q$ from $(\mathbf{a}, \langle \mathbf{a}, \mathbf{s} \rangle + e)$, where $\mathbf{a} \overset{\$}{\leftarrow} \mathbb{Z}_q^n$ is uniform and $e \overset{\chi}{\leftarrow} \mathbb{Z}_q$. We use a variant of the original LWE problem, for which the secret $\mathbf{s}$ is sampled from χ rather than $\mathcal{U}$. A polynomial reduction to the original decision LWE is given in [15].

f) FrodoPKE: This section presents the basic algorithm of FrodoPKE [8], which can be transformed into an IND-CCA secure KEM called FrodoKEM [8] using the Fujisaki-Okamoto (FO) transform [16], keeping the error probability unchanged. FrodoPKE is designed to guarantee IND-CPA security at three levels: Frodo-640, Frodo-976 and Frodo-1344. The security of these levels matches the brute-force security of AES-128, AES-192 and AES-256 respectively. Each level is parameterized by an integer dimension n such that $n \equiv 0 \bmod 8$, a variance σ and a discrete error distribution χ_{Frodo} which is close to the rounded Gaussian Ψ_σ in Rényi divergence. The LWE modulus q in FrodoPKE is either 2^{14} or 2^{15} depending on what level is adopted. A sketch of the algorithm is given in Table I. Alice generates $\mathbf{A} \overset{\$}{\leftarrow} \mathbb{Z}_q^{n\times n}$, then samples $\mathbf{S}, \mathbf{E} \leftarrow \chi_{\text{Frodo}}^{n\times\bar{n}}$, computes the LWE samples $\mathbf{B} = \mathbf{AS} + \mathbf{E}$ and outputs a public key $(\mathbf{A}, \mathbf{B})$. Bob chooses $\mathbf{S}', \mathbf{E}', \mathbf{E}'' \leftarrow \chi_{\text{Frodo}}^{\bar{n}\times n}$, then computes the LWE samples $\mathbf{U} = \mathbf{S}'\mathbf{A} + \mathbf{E}'$ and $\mathbf{V} = \mathbf{S}'\mathbf{B} + \mathbf{E}''$. A message $\mathbf{m}$ in $\{0,1\}^\ell$ is generated unilaterally on Bob's side and encoded into $\mathbb{Z}_q^{\bar{n}\times\bar{n}}$ using the function FRODO.ENCODE$(\cdot)$ [8, Algorithm 1]. Alice recovers $\mathbf{m}'$ using the decoding mechanism [8, Algorithm 2]. The two messages are the same except with probability $P_e = \mathbb{P}\{\mathbf{m}' \neq \mathbf{m}\}$. The number of message bits $\ell \in \{128, 192, 256\}$ depends on the assigned security level.

Parameters: q; $n \in \{640, 976, 1344\}$; $\bar{n} = 8$
FrodoKEM's distribution χ_{Frodo}

Alice (server)		**Bob (Client)**
$\mathbf{A} \overset{\$}{\leftarrow} \mathbb{Z}_q^{n\times n}$		
$\mathbf{S}, \mathbf{E} \leftarrow \chi_{\text{Frodo}}^{n\times\bar{n}}$		$\mathbf{S}', \mathbf{E}' \leftarrow \chi_{\text{Frodo}}^{\bar{n}\times n}$,
$\mathbf{B} := \mathbf{AS} + \mathbf{E} \in \mathbb{Z}_q^{n\times\bar{n}}$	$\xrightarrow{(\mathbf{A},\mathbf{B})}$	$\mathbf{E}'' \leftarrow \chi_{\text{Frodo}}^{\bar{n}\times\bar{n}}$
		$\mathbf{U} := \mathbf{S}'\mathbf{A} + \mathbf{E}' \in \mathbb{Z}_q^{\bar{n}\times n}$
		$\mathbf{V} := \mathbf{S}'\mathbf{B} + \mathbf{E}'' \in \mathbb{Z}_q^{\bar{n}\times\bar{n}}$
		$\mathbf{m} \overset{\$}{\leftarrow} \{0,1\}^\ell$
$\mathbf{V}' := \mathbf{C} - \mathbf{US} \in \mathbb{Z}_q^{\bar{n}\times\bar{n}}$	$\xleftarrow{(\mathbf{U},\mathbf{C})}$	$\mathbf{C} = \mathbf{V}+$FRODO.ENCODE$(\mathbf{m})$
$\mathbf{m}' =$ FRODO.DECODE$(\mathbf{V}')$		

TABLE I
SIMPLIFIED DESCRIPTION OF THE ORIGINAL FRODOPKE

III. PROPOSED MODIFICATION OF FRODOPKE

With the choice of parameter $\bar{n} = 8$ in FrodoKEM [8], the message $\mathbf{m} \in \{0,1\}^\ell$ is encoded into a point of $\mathbb{Z}_q^{64}$. In this section, we propose a modified version of FrodoPKE where the encoder maps the key into a suitably scaled version of the 64-dimensional lattice E_8^8, i.e. the product of 8 copies of the Gosset lattice. Since E_8 is the densest 8-dimensional packing, this results in a more efficient encoding. Since all integer operations in FrodoPKE are performed modulo q, we identify the lattice points that are equivalent modulo $q\mathbb{Z}^{64}$.

Referring to Table I, the main adjustments are made for the encryption and decryption algorithms FRODO.ENCODE$(\cdot)$ and FRODO.DECODE$(\cdot)$ respectively. Following the approach in [17], we search for a suitable scaling parameter β such that $q\mathbb{Z}^{64} \subseteq (\beta E_8)^8 \subseteq \mathbb{Z}^{64}$, knowing that $2\mathbb{Z}^8 \subseteq E_8 \subseteq \frac{1}{2}\mathbb{Z}^8$. Our aim is to define an encoding function from $\{0,1\}^\ell$ to $(\beta E_8)^8 / q\mathbb{Z}^{64} \subseteq \mathbb{Z}^{8\times 8}$. This function is one-to-one if the number of points in $(\beta E_8)^8 / q\mathbb{Z}^{64}$, which is $\text{Vol}\left(q\mathbb{Z}^{64}\right) / \text{Vol}\left((\beta E_8)^8\right)$, is greater than or equal to 2^ℓ. This condition is verified by setting $\beta = q/2^{\ell/64} \in \{q/4, q/8, q/16\}$ for $\ell \in \{128, 192, 256\}$.

The construction of the encoder is as follows. First, $\mathbf{m} \in \{0,1\}^\ell$ is partitioned into 8 substrings $\mathbf{m}_i \in \{0,1\}^{\ell/8}$, $i = 0,..,7$. Each substring is mapped into $\beta E_8 / q\mathbb{Z}^8 \subseteq \mathbb{Z}^8$. For simplification, each element in $\beta E_8 / q\mathbb{Z}^8$ is identified with the corresponding coset leader in $E_8 / 2^{\ell/64}\mathbb{Z}^8$. As an example, for $\ell = 128$, the value of β is $q/4$. Hence mapping 8 bits of information into $E_8/2\mathbb{Z}^8$ allows to map 16 bits into $E_8/4\mathbb{Z}^8$. Let $f : \{0,1\}^8 \longrightarrow E_8/2\mathbb{Z}^8$ that maps $\mathbf{b} = [b_1, b_2, \ldots, b_8] \in \{0,1\}^8$ as follows:

$$\begin{cases} f(\mathbf{b}) = [b_1, \ldots, b_7, -1] \cdot \mathbf{G}_{E_8} \bmod 2 & \text{if } b_1 = 0 \text{ \&\& } b_8 = 0 \\ f(\mathbf{b}) = [b_1, \ldots, b_7, 0] \cdot \mathbf{G}_{E_8} \bmod 2 & \text{if } b_1 = 0 \text{ \&\& } b_8 = 1 \\ f(\mathbf{b}) = [b_1, \ldots, b_7, 1] \cdot \mathbf{G}_{E_8} \bmod 2 & \text{if } b_1 = 1 \text{ \&\& } b_8 = 0 \\ f(\mathbf{b}) = [b_1, \ldots, b_7, 2] \cdot \mathbf{G}_{E_8} \bmod 2 & \text{if } b_1 = 1 \text{ \&\& } b_8 = 1 \end{cases}$$

One can verify that f is a bijective function. We can map 16 bits into the quotient $E_8/4\mathbb{Z}^8$ as follows: map the first 8 bits into $E_8/2\mathbb{Z}^8$, and the remaining ones into $2\mathbb{Z}^8/4\mathbb{Z}^8$. This last mapping is obtained by simply multiplying the input string by 2. This example can be extended to the cases $\ell = 192$ and $\ell = 256$ by considering the chain $E_8 \supseteq 2\mathbb{Z}^8 \supseteq 4\mathbb{Z}^8 \supseteq 8\mathbb{Z}^8 \supseteq 16\mathbb{Z}^8$. We denote the function that maps the remaining $\ell/8 - 8$ bits by g. The encoding function FRODO.ENCODE$(\cdot)$ can now be changed to E8.ENCODE$(\cdot)$ as shown in Algorithm 1.

Algorithm 1 Gosset Lattice Encoding

```
function E8.ENCODE($\mathbf{m} \in \{0,1\}^\ell$)
    $\mathbf{m}_{i\,:\,i=0,\ldots,7} = (m_{i(\ell/8)}, \ldots, m_{i(\ell/8)+\ell/8-1}) \in \{0,1\}^{\frac{\ell}{8}}$
    $\mathbf{X}_{i\,:\,i=0,\ldots,7} = f(\mathbf{m}_{i,0}, \ldots, \mathbf{m}_{i,7}) \in E_8/2\mathbb{Z}^8$
    $\mathbf{X}'_{i\,:\,i=0,\ldots,7} = g(\mathbf{m}_{i,8}, \ldots, \mathbf{m}_{i,\ell/8-1}) \in 2\mathbb{Z}^8/2^{\ell/64}\mathbb{Z}^8$
    $\mathbf{R}_{i\,:\,i=0,\ldots,7} = \mathbf{X}_i + \mathbf{X}'_i \in E_8/2^{\ell/64}\mathbb{Z}^8 \cong \beta E_8/q\mathbb{Z}^8$
    return $O_{i,j} = \left(R_{(8-i+j) \bmod 8, j}\right)_{\substack{0\le i\le 7 \\ 0\le j\le 7}}$
```

Note that each substring $\mathbf{m}_i$ is mapped into a vector in $\mathbb{Z}^8$, which is encoded in a block

$$\text{BLOCK}_i(\mathbf{O}) = (O_{i \bmod 8, 0}, \ldots, O_{i+7 \bmod 8, 7}) \tag{1}$$

of 8 components of the output matrix $\mathbf{O}$. Finally, E8.ENCODE is a bijection from $\{0,1\}^\ell$ to $(\beta E_8)^8 / q\mathbb{Z}^{64}$.

The decoding algorithm E8.DECODE uses the CVP$_{E_8}$ algorithm [18] presented in Algorithm 2 below.

Algorithm 2 Closest Vector Point in E_8

```
function CVP$_{E_8}$($\mathbf{x} \in \mathbb{R}^8$)
    $\mathbf{f} = \lfloor \mathbf{x} \rceil$ ; $\mathbf{g} = \rfloor \mathbf{x} \lceil$
    $\mathbf{y} = (1 \oplus \sum f_i)\,\mathbf{f} + (1 \oplus \sum g_i)\,\mathbf{g}$
    $\mathbf{f}' = \lfloor \mathbf{x} - \frac{1}{2} \rceil$ ; $\mathbf{g}' = \rfloor \mathbf{x} - \frac{1}{2} \lceil$
    $\mathbf{y}' = (1 \oplus \sum f'_i)\,\mathbf{f}' + (1 \oplus \sum g'_i)\,\mathbf{g}' + \frac{1}{2}$
    return $\underset{\mathbf{y}'' \in \{\mathbf{y}, \mathbf{y}'\}}{\text{argmin}} \|\mathbf{x} - \mathbf{y}''\|$
```

We describe the decoding protocol in Algorithm 3. It concatenates the outputs of CVP_{E_8} to form an element of $(\beta E_8)^8/q\mathbb{Z}^{64}$. Since our lattice E_8 is scaled by β, we use the fact that $\text{CVP}_{\beta E_8}(\mathbf{x}) = \beta \cdot \text{CVP}_{E_8}\left(\frac{1}{\beta}\mathbf{x}\right)$.

Algorithm 3 Gosset Lattice Decoding

```
function E8.Decode($\mathbf{N} \in \mathbb{R}_q^{8\times 8}$)
    $\mathbf{Y}_{i\,:\,1\le i\le 8} = \beta \cdot \text{CVP}_{E_8}\left(\frac{1}{\beta}\text{Block}_i(\mathbf{N})\right) \bmod q$
    $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_8] \in (\beta E_8)^8/q\mathbb{Z}^{64}$
    return $\mathbf{m}' = \text{E8.Encode}^{-1}(\mathbf{Y}) \in \{0,1\}^\ell$
```

IV. Reliability

In this section we aim to provide an upper bound for the decryption error probability for our algorithm. Clearly, an error occurs whenever the received message $\mathbf{m}'$ differs from the original one $\mathbf{m}$, i.e., $P_e = \mathbb{P}\{\mathbf{m} \neq \mathbf{m}'\}$. Following Table I, the expression of $\mathbf{V}'$ can be simplified as follows:

$$\begin{aligned}\mathbf{V}' &= \mathbf{C} - \mathbf{US} = \mathbf{V} + \text{E8.Encode}(\mathbf{m}) - (\mathbf{S}'\mathbf{A} + \mathbf{E}')\,\mathbf{S}\\ &= \mathbf{S}'(\mathbf{AS} + \mathbf{E}) \ + \mathbf{E}'' + \text{E8.Encode}(\mathbf{m}) - \mathbf{S}'\mathbf{AS} - \mathbf{E}'\mathbf{S}\\ &= \text{E8.Encode}(\mathbf{m}) + \underbrace{\mathbf{S}'\mathbf{E} + \mathbf{E}'' - \mathbf{E}'\mathbf{S}}_{\mathbf{E}'''}.\end{aligned}$$

From this we can express the decoded message $\mathbf{m}'$ as

$$\begin{aligned}\mathbf{m}' &= \text{E8.Decode}(\mathbf{V}')\\ &= \text{E8.Decode}\left(\text{E8.Encode}(\mathbf{m}) + \mathbf{E}'''\right)\\ &= \mathbf{m} + \text{E8.Decode}\left(\mathbf{E}'''\right).\end{aligned}$$

Each entry $E'''_{i,j}$ in the matrix $\mathbf{E}'''$ is the sum of $2n$ products of two independent samples from χ, adding to it another independent sample also from χ:

$$\forall\, 0 \le i,j \le 7,\ E'''_{i,j} = \sum_{k=0}^{n-1}\left(S'_{i,k}E_{k,j} - E'_{i,k}S_{k,j}\right) + E''_{i,j} \quad (2)$$

The distribution of $E'''_{i,j}$, denoted by χ', can be efficiently computed using the product of *probability generating functions*. Due to equation (2), two entries of the matrix $\mathbf{E}'''$ which are not on the same row or column are independent, and hence we can extract 8 identically distributed blocks of 8 independent coordinates from this error matrix, just as indicated in equation (1). Decoding is correct whenever $\text{E8.Decode}\left(\mathbf{E}'''\right) = 0$. For this it is sufficient to have $\text{Block}_k\left(\mathbf{E}'''\right) \in \mathcal{V}(\beta E_8)$ for all $k = 0,..,7$, i.e.,

$$\left\langle \text{Block}_k\left(\mathbf{E}'''\right), \mathbf{v}\right\rangle < \frac{\|\mathbf{v}\|_2^2}{2},\ \forall \mathbf{v} \in \beta\left(\text{VR}_1 \cup \text{VR}_2\right).$$

The error probability can thus be bounded by

$$\begin{aligned}P_e &\le \sum\nolimits_{i=0}^{7} \mathbb{P}\left\{\exists\, \mathbf{v}_1 \in \text{VR}_1 : \left\langle \text{Block}_k\left(\mathbf{E}'''\right), \mathbf{v}_1\right\rangle \ge \beta\|\mathbf{v}_1\|_2^2/2\right\}\\ &+ \sum\nolimits_{i=0}^{7} \mathbb{P}\left\{\exists\, \mathbf{v}_2 \in \text{VR}_2 : \left\langle \text{Block}_k\left(\mathbf{E}'''\right), \mathbf{v}_2\right\rangle \ge \beta\|\mathbf{v}_2\|_2^2/2\right\} \quad (3)\end{aligned}$$

Since the error probability is independent of the choice of Voronoi relevant vector for vectors of the same type (because the distribution of each entry of $\mathbf{E}'''$ is symmetric, centered at 0), without loss of generality we can choose $\mathbf{v}_1 = (1,1,0,0,0,0,0,0)$ and $\mathbf{v}_2 = (\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2})$. This reduces the computations to just two cases. Choosing the value of n and the modulus q, we can compute an upper bound for the above expression for different values of σ. The R.H.S. of equation (3) becomes:

$$8 \cdot 112 \cdot \mathbb{P}\left\{E'''_{0,0} + E'''_{1,1} \ge \beta\right\} + 8 \cdot 128 \cdot \mathbb{P}\left\{E'''_{0,0} + \cdots + E'''_{7,7} \ge 2\beta\right\}.$$

In order to upper bound P_e, we use the following.

Remark 1: A discrete distribution p taking values in $\mathbb{Z}$ is *unimodal* with mode 0 if $p(n+1) \le p(n-1)\ \forall n \ge 0$, and $p(n+1) \ge p(n)\ \forall n < 0$. The convolution of two symmetric discrete unimodal distributions is symmetric unimodal [19, Theorem 4.7].

Since the distribution χ' is symmetric unimodal, so are the distributions χ'_2, χ'_4, χ'_8 of the sum of two, four and eight independent copies of $E'''_{i,j}$ respectively. While χ'_2 and χ'_4 can be calculated efficiently, the computation of χ'_8 is slow. Thanks to unimodality, we can estimate the term $\mathbb{P}\left\{E'''_{0,0} + \cdots + E'''_{7,7} \ge 2\beta\right\}$ by upper bounding χ'_8 by a piecewise constant function.

V. Security

a) IND-CPA security: Our scheme only modifies the encoding and decoding functions, the choice of parameters q and σ, and the error distribution. As shown in [8], the IND-CPA security of FrodoPKE is upper bounded by the advantage of the decision-LWE problem for the same parameters and error distribution [Theorem 5.9, Theorem 5.10]. We note that the security proof relies on the pseudorandomness of the adversary's observation (similarly to [20, Theorem 3.2]) and thus the choice of encoding function has no effect on the security level, which is only affected by the parameters and error distribution. In terms of security against known attacks, the best known bound is given by the BKZ attacks, which involve both primal and dual attacks [21].

b) IND-CCA security: It was shown in [8] that applying the Fujisaki-Okamoto transformation to the IND-CPA secure protocol FrodoPKE yields an IND-CCA secure key encapsulation mechanism FrodoKEM, even if they use different error distributions, provided that the Rényi divergence between these error distributions is small. In particular, FrodoKEM using the finite support distribution χ_{Frodo} is IND-CCA secure provided that the FrodoPKE protocol using a rounded Gaussian distribution Ψ_σ is IND-CPA secure, and the classical IND-CCA advantage $\text{Adv}^{\text{ind-cca}}$ can be upper bounded by [8, Equation (3)] $\forall \alpha > 1$:

$$\frac{q_{\text{RO}}}{|\mathcal{M}|} + \left(\left(\frac{2\cdot q_{\text{RO}}+1}{|\mathcal{M}|} + q_{\text{RO}} \cdot P_e + 3 \cdot \text{Adv}^{\text{ind-cpa}}\right) \cdot e^{t \cdot D_\alpha(P||Q)}\right)^{1-\frac{1}{\alpha}}$$

where q_{RO} is the maximum number of oracle queries, $|\mathcal{M}| = 2^\ell$ is the cardinality of the set of keys, and $t = 2n(8+8)+64$ is the total number of samples (drawn from the error distribution χ) used to generate $\mathbf{E}, \mathbf{S}, \mathbf{E}', \mathbf{S}'$ and $\mathbf{E}''$ in Table I. In our case, $P = \chi$ and Q is the rounded Gaussian Ψ_σ. The security loss will be minimized by optimizing over the order α.

VI. Performance Comparison

In this section we show the performance of the proposed modification of FrodoKEM. We propose three sets of parameters: the first aims at improving the security level, the second at reducing the bandwidth and the third at increasing the key size. Note that for all sets of parameters, n and $\bar{n}$ will remain unchanged. The performance comparison is shown in Table II. The security level refers to the primal and dual attack via the FrodoKEM script `pqsec.py` with parameters n, σ, q.

a) Parameter set 1 - Reducing the bandwidth: For the first set of parameters, we aim at reducing the bandwidth while keeping the same security level. This is achieved by reducing the modulus q by half, which in turn requires a reduction in standard deviation σ in order to preserve a low error probability[1]. Overall, the modulus to noise ratio of the protocol is increased. Compared to the original FrodoKEM, this allows to reduce the bandwidth by approximately 7% [2].

b) Parameter set 2 - Improving the security level: For the second parameter set, we aim at increasing the plausible security level[3] while keeping the same bandwidth and a similar error probability level as in the original FrodoKEM protocol. To do so, we increase the variance σ while keeping q unchanged. Note that we can increase σ because of the higher error correction capability provided by our modified encoder.

c) Parameter set 3 - Increasing the key size: For the last set of parameters, we aim to increase the key size for Frodo-640 and Frodo-976 with comparable security and error probability. We generate 192 bits from Frodo-640 instead of 128 bits, as well as 256-bit key instead of 192 bits in Frodo-976. The modulus q remains unchanged.

Original FrodoKEM								
	σ	q	Key	Concrete Security			Bandwidth	P_e
			size	C	Q	P	(bytes)	
Frodo-640	2.80	2^{15}	128	145	132	104	9720	2^{-138}
Frodo-976	2.30	2^{16}	192	210	191	150	15744	2^{-199}
Frodo-1344	1.40	2^{16}	256	275	250	197	21632	2^{-252}
Reduce Bandwidth - Parameter set 1								
Modified Frodo-640	2.30	2^{14}	128	152	138	109	9072	2^{-152}
Modified Frodo-976	1.80	2^{15}	192	215	197	155	14760	2^{-203}
Modified Frodo-1344	1.14	2^{15}	256	283	257	203	20280	2^{-271}
Security Improvements - Parameter set 2								
Modified Frodo-640	3.90	2^{15}	128	158	144	113	9720	2^{-149}
Modified Frodo-976	2.75	2^{16}	192	220	200	158	15744	2^{-204}
Modified Frodo-1344	1.68	2^{16}	256	287	261	205	21632	2^{-255}
Increasing the key size - Parameter set 3								
Modified Frodo-640	2.30	2^{15}	192	139	126	100	9720	2^{-149}
Modified Frodo-976	1.80	2^{16}	256	200	184	144	15744	2^{-203}

TABLE II
Modified parameters for improving the bandwidth, the security level and/or increasing the key size for FrodoKEM.

Acknowledgment

The work of C. Saliba and L. Luzzi is supported by the CY Cergy Paris Université INEX Project AAP 2017.

References

[1] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," *Journal of the ACM (JACM)*, vol. 56, no. 6, p. 34, 2009.

[2] V. Lyubashevsky, C. Peikert, and O. Regev, "On ideal lattices and learning with errors over rings," in *EUROCRYPT 2010*, pp. 1–23.

[3] A. Langlois and D. Stehlé, "Worst-case to average-case reductions for module lattices," *Designs, Codes and Cryptography*, vol. 75, no. 3, pp. 565–599, 2015.

[4] R. Cramer, L. Ducas, C. Peikert, and O. Regev, "Recovering short generators of principal ideals in cyclotomic rings," in *EUROCRYPT 2016*. Springer, pp. 559–585.

[5] A. Pellet-Mary, G. Hanrot, and D. Stehlé, "Approx-SVP in ideal lattices with pre-processing," in *EUROCRYPT 2019*. Springer, pp. 685–716.

[6] O. Bernard and A. Roux-Langlois, "Twisted-PHS: Using the Product Formula to Solve Approx-SVP in Ideal Lattices," in *ASIACRYPT 2020*. Springer, 2020, pp. 349–380.

[7] R. Avanzi *et al.*, "Crystals-Kyber algorithm specifications and supporting documentation," *NIST PQC Round 3*, 2020. [Online]. Available: https://pq-crystals.org/kyber/data/kyber-specification-round3.pdf

[8] M. Naehrig *et al.*, "FrodoKEM. tech. rep." in *NIST PQC Round3*, 2020. [Online]. Available: https://csrc.nist.gov/projects/post-quantum-cryptography/round-3-submissions

[9] E. Lee *et al.*, "Modification of FrodoKEM using Gray and error-correcting codes," *IEEE Access*, vol. 7, 2019.

[10] J.-P. D'Anvers, F. Vercauteren, and I. Verbauwhede, "The impact of error dependencies on Ring/Mod-LWE/LWR based schemes," in *International Conference on Post-Quantum Cryptography*. Springer, 2019.

[11] E. Alkim, L. Ducas, T. Pöppelmann, and P. Schwabe, "Post-quantum key exchange-a new hope." in *USENIX Security Symposium*, 2016.

[12] Y. Zhao, Z. Jin, B. Gong, and G. Sui, "A modular and systematic approach to key establishment and public-key encryption based on LWE and its variants," *NIST PQC Round*, vol. 1, p. 4, 2017.

[13] C. Saliba, L. Luzzi, and C. Ling, "A reconciliation approach to key generation based on Module-LWE," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[14] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. Springer Science & Business Media, 2013, vol. 290.

[15] B. Applebaum, D. Cash, C. Peikert, and A. Sahai, "Fast cryptographic primitives and circular-secure encryption based on hard learning problems," in *Annual International Cryptology Conference*. Springer, 2009.

[16] D. Hofheinz, K. Hövelmanns, and E. Kiltz, "A modular analysis of the Fujisaki-Okamoto transformation," in *Theory of Cryptography Conference*. Springer, 2017.

[17] A. van Poppelen, "Cryptographic decoding of the Leech lattice," Master's thesis, Utrecht University, 2016.

[18] J. Conway and N. Sloane, "Fast quantizing and decoding algorithms for lattice quantizers," *IEEE Trans Inform Theory*, vol. 28, no. 2, 1982.

[19] S. Dharmadhikari and K. Joag-Dev, *Unimodality, convexity and applications*. Elsevier, 1988.

[20] R. Lindner and C. Peikert, "Better key sizes (and attacks) for LWE-based encryption," in *Cryptographers' Track at the RSA Conference*. Springer, 2011, pp. 319–339.

[21] Y. Chen and P. Q. Nguyen, "BKZ 2.0: Better lattice security estimates," in *ASIACRYPT 2011*. Springer, pp. 1–20.

[1] The condition $\sigma \geq 2.12$ is imposed in [8] to allow the reduction from the bounded distance decoding with discrete Gaussian sampling (BDDwDGS) to the decision LWE problem. Note that for efficiency reasons, σ is equal to 1.4 in Frodo-1344, while still guaranteeing a large number N of discrete Gaussian samples, namely $N \approx 2^{111}$. For Frodo-1344 we take $\sigma = 1.15$, which still leads to a large number of discrete Gaussian samples, namely $N \approx 2^{75}$.

[2] The communication requirements of the protocol are computed using the functions `Frodo.Pack` and `Frodo.Unpack` presented in [8]. In our case we pack both $\mathbf{U} \in \mathbb{Z}^{\bar{n} \times n}$ and $\mathbf{C} \in \mathbb{Z}^{\bar{n} \times \bar{n}}$. Those two vectors, concatenated together, carry $(\log_2(q) \times n + \log_2(q) \times 8)$ bytes.

[3] The security level of FrodoKEM with respect to primal/dual attacks is already higher than the brute force security level, but this might change due to improvements in the best known attacks. So this choice of parameters represents an even more conservative option for long-term security.

Optimal Simulation of Quantum Measurements via the Likelihood POVMs

Arun Padakandla

Abstract—**We provide a new and simplified proof of Winter's measurement compression [1] via likelihood POVMs. Secondly, we provide an alternate proof of the central tool at the heart of this theorem - the Quantum covering lemma - that does not rely on the Ahlswede Winter's operator Chernoff bound [2], thereby requires only pairwise independence of the involved random operators. We leverage these results to design structured POVMs and prove their optimality in regards to communication rates.**

I. Introduction

The design and analysis of quantum measurements play a central role in both quantum information processing and quantum physics. The outcome of a quantum measurement being inherently random, a question of fundamental interest [3], [4] is to quantify the amount of *information* it contains. A series of works [5], [6] aimed at addressing this question culminated in Winter's measurement compression theorem [1]. Adopting a traditional information-theoretic modeling, Winter provided an elegant and precise solution in the context of a generic Positive Operator Valued Measurement (POVM).

Our motivation is to address the above question in a generic scenario involving multiple centralized/distributed POVMs. Suppose $\lambda_1, \lambda_2, \lambda_3$ are three stochastically compatible POVMs [7, Sec. 2.1.2]. Provided with the outcomes of (λ_1, λ_2), how much additional information does the outcome of λ_3 contain? A second question of interest is to quantify the amount of information contained in a distributed POVM $\lambda_1 \otimes \lambda_2$ operated on a pair of distributed particles that are, in general entangled. In this article, we take our first step towards addressing these problems by providing a newer and much simplified proof of Winter's findings (Sec. III). We also demonstrate the generality of these proof techniques in our study of 'structured' POVMs.

Firstly, we propose simulation of the original POVM with a canonical class of likelihood POVMs that are much easier to describe. While being the most natural for the problem at hand, its performance analysis has remained elusive, leading Winter and subsequent works to design and analyze more involved POVMs. In fact, even the excellent tutorial-style exposition of Wilde et. al. [4] and the more recent works [8] resort to the latter involved POVMs. Our analysis of the likelihood POVMs is based on the following crucial idea. Recognizing that the outcome of the likelihood POVM on the original state has an involved characterization that is not amenable for analysis, we design a specific mixture of quantum states (Sec. III-B) as a proxy for the original state. This mixture is so designed such that the outcome of the likelihood POVM on it takes a simple form (Sec. III-D). We are thus left with two questions. Does the latter outcome closely approximate the outcome of the original POVM on the original state? and does the proxy, i.e., the designed mixture, closely approximate the original state? Identifying the right set of mathematical tools (Sec. III-C, III-E), we reduce the above two conditions to instances of the quantum covering lemma (QCL) [9, Chap. 17].

Our second contribution is a new proof of the QCL. Known proofs of QCL are based on the Ahlswede and Winter's [2] operator Chernoff bound (OCB). The OCB requires that the random operators be *mutually* independent. Its use for the problem at hand precludes the simulation POVM to have any additional structure. For example, relying on the OCB simulation precludes proving optimality of a simulation POVM with a 'algebraic closure structure' (Sec. IV). This is because, if one picks a random POVM with an algebraic structure, its operators are *not* mutually independent. Our second contribution is a new proof (Lem. 1, Sec. III-E) of the QCL that does not rely on the OCB and the underlying concentration only requires pairwise independence. Building on this we design - as an application of all our findings - structured likelihood POVMs (Sec. IV) that simulate POVMs with optimal communication costs (Thm. 2).

Winter's finding [1] and the associated tools remain to be a subject of continued interest. In addition to providing an excellent exposition, Wilde et. al. [4] study single POVM generalization of [1]. Anshu, Jain and Warsi [10] build on their novel convex-split lemma (CSL) [11] to study the POVM compression with side-information in the one-shot setting. The recent work [12] provides a good account of the CSL and QCL. More recently, Pradhan et. al. [8] study the problem of distributed POVM compression using the Winter's approach. Our findings here has the potential to simplify the proofs of the above works in the asymptotic IID setting. Enlarged descriptions, detailed steps and complete proofs of our results can be found in [13].

II. Preliminaries and Problem Statement

Notation : We supplement standard quantum information theory notation with the following. For a positive integer n, we let $[n] \triangleq \{1, \cdots, n\}$, $[\overline{n}] \triangleq \{0\} \cup [n]$. All Hilbert spaces are assumed to be finite dimensional. $\mathcal{L}(\mathcal{H})$, $\mathcal{R}(\mathcal{H})$, $\mathcal{P}(\mathcal{H})$, $\mathcal{D}(\mathcal{H})$ denote the collection of linear, Hermitian, positive and density operators acting on Hilbert space $\mathcal{H}$ respectively. For $s \in \mathcal{D}(\mathcal{H}_A)$, $|\phi_s\rangle \in \mathcal{H}_X \otimes \mathcal{H}_A$, with $\mathcal{H}_X = \mathcal{H}_A$, denotes a purification of s, i.e., $\langle \phi_s | \phi_s \rangle = 1$ and $\text{tr}_X(|\phi_s\rangle\langle\phi_s|) = s$.

POVMs will play a central role in this article. $\mathscr{M}(\mathcal{H}, \mathcal{Y})$ denotes the set of all POVMs acting on $\mathcal{H}$ with outcomes in $\mathcal{Y}$. Often times, we denote a POVM $\lambda = \{\lambda_y \in \mathcal{P}(\mathcal{H}) :$

$y \in \mathcal{Y}\} \in \mathscr{M}(\mathcal{H},\mathcal{Y})$ by adding the outcome set as a subscript, as in $\lambda_{\mathcal{Y}} = \lambda$. For POVM $\lambda = \{\lambda_y : y \in \mathcal{Y}\}$, $\lambda^{\otimes n} \triangleq \{\lambda_{y^n} \triangleq \lambda_{y_1} \otimes \cdots \otimes \lambda_{y_n} : y^n \in \mathcal{Y}^n\}$. To reduce clutter, we let $\lambda^n = \lambda_{\mathcal{Y}^n} = \lambda^{\otimes n}$ denote the same object. Associated with a POVM $\lambda = \{\lambda_y : y \in \mathcal{Y}\}$ is a Hilbert space $\mathcal{H}_{\mathcal{Y}} \triangleq \operatorname{span}\{|y\rangle : y \in \mathcal{Y}\}$ with $\langle \hat{y}|y\rangle = \delta_{\hat{y}y}$ and the CPTP map $\mathscr{E}^{\lambda} : \mathcal{L}(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H}_{\mathcal{Y}})$, defined as $\mathscr{E}^{\lambda}(s) = \sum_{y\in\mathcal{Y}} \operatorname{tr}(s\lambda_y)\, |y\rangle\langle y|$. For a stochastic matrix $(p_{Y|W}(y|w) : (w,y) \in \mathcal{W} \times \mathcal{Y})$, we let $\mathscr{E}_p^{Y|W} : \mathcal{L}(\mathcal{H}_{\mathcal{W}}) \rightarrow \mathcal{L}(\mathcal{H}_{\mathcal{Y}})$ denote the CPTP map $\mathscr{E}_p^{Y|W}(a) \triangleq \sum_{(w,y)\in\mathcal{W}\times\mathcal{Y}} p_{Y|W}(y|w)\, |y\rangle\langle w|\, a\, |w\rangle\langle y|$. The composition of CPTP maps $\mathcal{L}(\mathcal{H}_A) \overset{\mathscr{E}_1}{\rightarrow} \mathcal{H}_B$, $\mathcal{L}(\mathcal{H}_B) \overset{\mathscr{E}_2}{\rightarrow} \mathcal{H}_C$ is denoted $\mathscr{E}_2 \circ \mathscr{E}_1$. For an ensemble $\rho_w \in \mathcal{D}(\mathcal{H}) : w \in \mathcal{W}$ with PMF $p_W(\cdot)$ on $\mathcal{W}$, $\chi(\rho_w; p_W(w) : \mathcal{W}) \triangleq S(\sum_{w\in\mathcal{W}} p_W(w)\rho_w) - \sum_{w\in\mathcal{W}} p_W(w) S(\rho_w)$ denotes Holevo information. SCD, WHP abbreviate spectral decomposition and with high probability.

Problem Description: Let Hilbert space $\mathcal{H}_A$ have dimension d_A. Let $\rho \in \mathcal{D}(\mathcal{H}_A)$ model the behaviour of a given sub-atomic particle and $\lambda \triangleq \lambda_{\mathcal{Y}} \triangleq \{\lambda_y \in \mathcal{P}(\mathcal{H}_A) : y \in \mathcal{Y}\}$ denote a given POVM. We follow [1], [4] in modelling the following question. Suppose a measurement modeled via POVM $\lambda_{\mathcal{Y}}$ is performed on the particle $\rho \in \mathcal{D}(\mathcal{H}_A)$, what fraction of the randomness in the outcome is 'intrinsic' to the particle ρ, and what fraction is 'extrinsic', or unrelated to ρ? To quantify this, we design an 'alternate $n-$letter measurement' - a simulated POVM - that is supplemented with an independent source of *common randomness* of C bits/letter (Fig. 1) available at both terminals. These C bits of common randomness are statistically independent of the particle and are available to (i) design this simulated POVM and (ii) postprocess its outcome to simulate the outcome of the original POVM $\lambda_{\mathcal{Y}}$ on ρ. We require the outcome of the simulated POVM - both the post measurement particle and the observed outcome $Y^n \in \mathcal{Y}^n$ - to be statistically indistinguishable from that of the original POVM $\lambda_{\mathcal{Y}}$. Enforcing this, we aim to quantify the minimum rate R bits/letter that enables Bob reconstruct the classical outcome. Characterizing all possible (R, C) pairs enables us quantify the trade-off between intrinsic information and extrinsic randomness contained in the outcome of POVM $\lambda_{\mathcal{Y}}$.

The above stated requirement is specified in terms of demanding that the combined operators of the reference and outcome post measurement of both the original POVM $\lambda_{\mathcal{Y}}^{\otimes n}$ and the simulated one are statistically 'close'. Adopting trace distance to quantify 'closeness' we are led to the following.

Defn. 1. *Suppose $\rho \in \mathcal{D}(\mathcal{H}_A)$, $\mathcal{H}_X = \mathcal{H}_A$ and $\lambda \in \mathscr{M}(\mathcal{H}_A,\mathcal{Y})$ is a POVM. A sequence $\Xi^{(n)} \in \mathscr{M}(\mathcal{H}_A^{\otimes n},\mathcal{Y}^n) : n \geq 1$ of POVMs simulates λ on ρ if for all $\eta > 0$, $\exists$ $N_\eta \in \mathbb{N}$ such that for all $n \geq N_\eta$, we have $||\alpha_o - \alpha_{sp}||_1 \leq \eta$, where $\alpha_o \triangleq (i_X^{\otimes n} \otimes \mathscr{E}^{\lambda^{\otimes n}})(\left|\phi_{\rho^{\otimes n}}\right\rangle\left\langle\phi_{\rho^{\otimes n}}\right|)$ and $\alpha_{sp} = (i_X^{\otimes n} \otimes \mathscr{E}^{\Xi^{(n)}})(\left|\phi_{\rho^{\otimes n}}\right\rangle\left\langle\phi_{\rho^{\otimes n}}\right|)$.*

Since the simulated POVM can utilize independent randomness at both terminals, we let (i) $\rho_K^{\otimes n} \triangleq \frac{1}{K}\sum_{k\in[K]} \rho^{\otimes n} \otimes |k\rangle\langle k|$ model its input state, (ii) design the simulated measurement to be of the form $\theta \triangleq \{\theta_{k,m} \otimes |k\rangle\langle k| : (k,m) \in [K] \otimes [M]\} \in \mathscr{M}(\mathcal{H}_{A^nK}, [K] \times [M])$, where $\mathcal{H}_{A^nK} \triangleq \mathcal{H}_A^{\otimes n} \otimes \mathcal{H}_K$, $\mathcal{H}_K = \operatorname{span}\{|k\rangle : k \in [K]\}$ and $\left\langle \hat{k} \middle| k \right\rangle = \delta_{\hat{k}k}$. Essentially, the simulated POVM θ observes the common randomness k and chooses to perform the POVM $\{\theta_{k,m} : m \in [M]\} \in \mathscr{M}(\mathcal{H}_A^{\otimes n}, [M])$ and hands over the nature provided common randomness k and the POVM outcome m to the Dec. Denoting $\mathscr{S}(\mathcal{H}_{A^nK}, [KM]) \subseteq \mathscr{M}(\mathcal{H}_{A^nK}, [K] \times [M])$ as the set of POVMs of the above form, we define the quantity of interest.

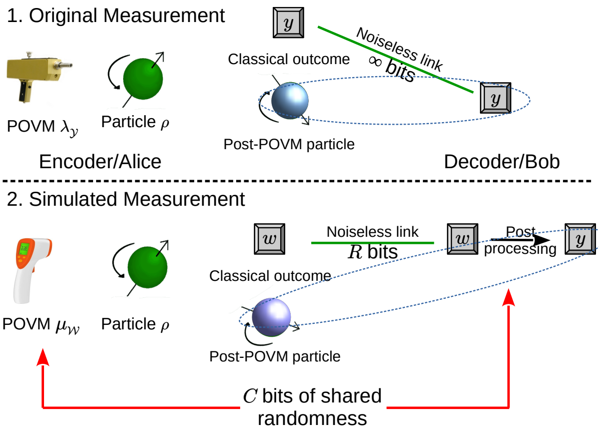


Fig. 1. Illustrates original and simulated POVMs. The components in the two blue ellipses must be statistically indistinguishable.

Defn. 2. *The communication cost of a simulation POVM $\theta \in \mathscr{S}(\mathcal{H}_{A^nK}, [KM])$ is $(\frac{\log K}{n}, \frac{\log M}{n})$. POVM $\lambda_{\mathcal{Y}}$ on ρ can be simulated at a cost (C, R) if $\forall \eta > 0$, $\exists N_\eta \in \mathbb{N}$, such that $\forall n \geq N_\eta$, there exists a $\theta \in \mathscr{S}(\mathcal{H}_{A^nK}, [KM])$ and a POVM $\Delta_{\mathcal{Y}^n} \in \mathscr{M}(\mathcal{H}_K \otimes \mathcal{H}_M, \mathcal{Y}^n)$ such that $\frac{\log K}{n} \leq C + \eta$, $\frac{\log M}{n} \leq R + \eta$ and $||\alpha_o - \alpha_s||_1 \leq \eta$, where*

$$\alpha_o \triangleq (i_X^n \otimes \mathscr{E}^{\lambda^n})(\left|\phi_{\rho^{\otimes n}}\right\rangle\left\langle\phi_{\rho^{\otimes n}}\right|), \alpha_s \triangleq \mathfrak{E}_{sim}(\left|\phi_{\rho_K^{\otimes n}}\right\rangle\left\langle\phi_{\rho_K^{\otimes n}}\right|), \quad (1)$$

$$\mathfrak{E}_{sim} \triangleq (i_X^n \otimes \operatorname{tr}_K \otimes i_{\mathcal{Y}^n}) \circ (i_X^n \otimes i_{\mathcal{H}_K} \otimes \mathscr{E}^{\Delta}) \circ (i_X^n \otimes i_{\mathcal{H}_K} \otimes \mathscr{E}^{\theta}), \quad (2)$$

$\mathcal{H}_X = \mathcal{H}_A$, i_X^n abbreviates the identity map $i_X^{\otimes n}$ on $\mathcal{L}(\mathcal{H}_X^{\otimes n})$.

A. Communication Cost of Simulating $\lambda_{\mathcal{Y}}$ on ρ

Winter [1] derived an elegant computable characterization for the communication cost of simulating $\lambda_{\mathcal{Y}}$ on ρ for the case when both terminals wish to possess outcome of the simulated POVM. We state below Wilde et. al.'s [4] generalization for the case when only Bob wishes to possess the simulated outcome.

Defn. 3. *For $\rho \in \mathcal{D}(\mathcal{H}_A)$ and $\lambda_{\mathcal{Y}} \in \mathscr{M}(\mathcal{H}_A, \mathcal{Y})$, let $\mathcal{C}(\rho, \lambda_{\mathcal{Y}})$ be a collection of triples $(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W})$ wherein (i) $\mathcal{W}$ is a finite set, (ii) $\mu_{\mathcal{W}} \in \mathscr{M}(\mathcal{H}_A, \mathcal{W})$ is a POVM, and (iii) $(p_{Y|W}(y|w) : (w,y) \in \mathcal{W} \times \mathcal{Y})$ is a stochastic matrix such that*

$$\operatorname{tr}_W\{(i_X \otimes \mathscr{E}^{p_{Y|W}}) \circ (i_X \otimes \mathscr{E}^{\mu})(|\phi_\rho\rangle\langle\phi_\rho|)\} = (i_X \otimes \mathscr{E}^{\lambda})(|\phi_\rho\rangle\langle\phi_\rho|). \quad (3)$$

For $(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \in \mathcal{C}(\rho, \lambda_{\mathcal{Y}})$, let $p_W(w) \triangleq \operatorname{tr}(\rho\mu_w)$,

$$\beta_w \triangleq \frac{\sqrt{\rho}\mu_w\sqrt{\rho}}{p_W(w)}, \gamma_w \triangleq \sum_{y\in\mathcal{Y}} p_{Y|W}(y|w)\beta_w \otimes |y\rangle\langle y|\,, \quad (4)$$

Let $\mathscr{RC}(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \triangleq \chi(\gamma_w, p_W(w) : \mathcal{W})$ *and* $\mathscr{R}(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \triangleq \chi(\beta_w, p_W(w) : \mathcal{W})$.

Remark 1. *For* $(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \in \mathcal{C}(\rho, \lambda_{\mathcal{Y}})$, *we note* $\mathrm{tr}_W\{(i_X \otimes \mathscr{E}^{p_{Y|W}}) \circ (i_X \otimes \mathscr{E}^{\mu})(|\phi_\rho\rangle\langle\phi_\rho|)\} = \sum_{w\in\mathcal{W}} p_W(w)\gamma_w$,

$$(i_X \otimes \mathscr{E}^{\lambda}) = \sum_{y\in\mathcal{Y}} \sqrt{\rho}\lambda_y\sqrt{\rho} \otimes |y\rangle\langle y| \,, \text{ and hence} \quad (5)$$

$$\rho = \sum_{w\in\mathcal{W}} p_W(w)\beta_w, \sum_{y\in\mathcal{Y}} \sqrt{\rho}\lambda_y\sqrt{\rho} \otimes |y\rangle\langle y| = \sum_{w\in\mathcal{W}} p_W(w)\gamma_w. \quad (6)$$

Theorem 1. *POVM* $\lambda_{\mathcal{Y}}$ *on* ρ *can be simulated at a cost* (C, R) *iff there exists* $(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \in \mathcal{C}(\rho, \lambda_{\mathcal{Y}})$ *for which* $R > \mathscr{R}(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W})$ *and* $R + C > \mathscr{RC}(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W})$.

III. Proof Via Unstructured Likelihood POVMs

Our first contribution is a simplified proof of achievability of Thm. 1 via unstructured POVMs.

A. Proof Setup : Notations, Definitions and Likelihood POVM

Our aim is to design a simulation POVM and characterize conditions under which $||\alpha_{\text{o}} - \alpha_{\text{s}}||_1$ can be made arbitrarily small. Choose $(\mathcal{W}, \mu_{\mathcal{W}}, p_{Y|W}) \in \mathcal{C}(\rho, \lambda_{\mathcal{Y}})$. Let $c : [K] \times [M] \to \mathcal{W}^n$ be a map and let $w^n(k, m) \triangleq (w(k, m)_1, \cdots, w(k, m)_n) \triangleq (c(k, m)_1, \cdots, c(k, m)_n)$. We let $\mu_{k,m} \triangleq \mu_{c(k,m)_1} \otimes \cdots \otimes \mu_{c(k,m)_n} \in \mathcal{P}(\mathcal{H}_A^{\otimes n})$, $\omega \triangleq \rho^{\otimes n}$ and

$$S_k \triangleq \sum_{m=1}^{M} \frac{\sqrt{\omega}\mu_{k,m}\sqrt{\omega}}{\mathrm{tr}(\omega\mu_{k,m})}, \theta_{k,m} \triangleq \frac{S_k^{-\frac{1}{2}}\sqrt{\omega}\mu_{k,m}\sqrt{\omega}S_k^{-\frac{1}{2}}}{\mathrm{tr}(\omega\mu_{k,m})}, \quad (7)$$

where $S_k^{-\frac{1}{2}}$, as is standard, is the square root of the generalized inverse of S_k. We let $\theta_{k,0} \triangleq I_{H_A}^{\otimes n} - \sum_{m=1}^{M} \theta_{k,m}$ for all $k \in [K]$. Since $\mu_{\mathcal{W}}$ is a POVM, we have $0 \leq \mu_{c(k,m)_i} \leq I_{H_A}$ for $i \in [n]$, and hence $0 \leq \mu_{k,m} \leq I_{H_A}^{\otimes n}$ implying $0 \leq \theta_{k,m} \leq I_{H_A}^{\otimes n}$. By definition, we have ensured $\sum_{m=0}^{M} \theta_{k,m} = I_{H_A}^{\otimes n}$ for all $k \in [K]$. We let $\theta \triangleq \{\overline{\theta}_{k,m} \triangleq \theta_{k,m} \otimes |k\rangle\langle k| : (k, m) \in [K] \times [M]\} \in \mathscr{S}(\mathcal{H}_{A^nK}, [KM])$ be our chosen POVM.

We choose $\Delta \triangleq \Delta_{\mathcal{Y}^n} \triangleq$

$$\{\Delta_{y^n} \triangleq \sum_{(k,m)\in[K]\times[M]} |k\ m\rangle\langle k\ m|\, p_{Y|W}^n(y^n|c(k, m)) : y^n \in \mathcal{Y}^n\} \quad (8)$$

as the post-processing POVM employed by the Dec. It is straight forward to verify $\Delta \in \mathscr{M}(\mathcal{H}_K \otimes \mathcal{H}_M, \mathcal{Y}^n)$ is a POVM.

B. Key Steps Outlining the Proof

The non-commutativity of quantum operations has obfuscated the analysis of the above defined POVM, leading studies [1], [4] including the more recent ones [8] to adopt an alternate simulation POVM leading to much complexity. The crucial idea is to study the outcome α of the likelihood POVM on a mixture state σ_{A^nK} instead of the original state $\rho_K^{\otimes n}$. Towards that end, for $a \in [K]$, let

$$T_a \triangleq \tfrac{S_a}{M} \text{ have SCD } T_a = \textstyle\sum_{t=1}^{d_A^n} \nu_{ta} |x_{ta}\rangle\langle x_{ta}| \,, \text{ i.e.,} \quad (9)$$

$$\langle x_{ta}|x_{\hat{t}a}\rangle = \delta_{\hat{t}t} \text{ and } \sigma_{A^nK} \triangleq \tfrac{1}{K} \textstyle\sum_{a\in[K]} T_a \otimes |a\rangle\langle a| \,. (10)$$

In order to bound $||\alpha_{\text{o}} - \alpha_{\text{s}}||_1$ we define

$$\alpha \triangleq \mathfrak{E}_{\text{sim}}(|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}|). \quad (11)$$

Recognizing that α above and α_{s} in (1) are the result of applying the same CPTP map - $\mathfrak{E}_{\text{sim}}$ in (2)- on two different states provides us with the right clue. Indeed, we have

$$||\alpha_{\text{o}} - \alpha_{\text{s}}||_1 \leq ||\alpha_{\text{o}} - \alpha||_1 + ||\alpha - \alpha_{\text{s}}||_1$$
$$\leq ||\alpha_{\text{o}} - \alpha||_1 + ||\,|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}| - \left|\phi_{\rho_K^{\otimes n}}\right\rangle\left\langle\phi_{\rho_K^{\otimes n}}\right|\,||_1 \quad (12)$$

from the Triangular inequality and the fact that CPTP maps shrink the trace distance [9, Eq. 9.69]. Through the rest of the proof, we analyze each of the terms in the RHS of (12). In Sec. III-C, we analyze the second term where we leverage an elegant bound that relates the distance between states and their 'canonical purifications'. In Sec. III-D, we evaluate the RHS of (11) and thereby knock off the inconvenient outer normalizing factors $S_k^{-\frac{1}{2}}$ in $\theta_{k,m}$ defined in (7)!

C. Relating distance between states and their purifications

Since tracing over components decreases the trace distance, $||\sigma_{A^nK} - \rho_K^{\otimes n}||_1 \leq ||\,|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}| - \left|\phi_{\rho_K^{\otimes n}}\right\rangle\left\langle\phi_{\rho_K^{\otimes n}}\right|\,||_1$. However, we need an inequality in the reverse. The choice of the 'canonical purification' [1], [9, Pg. 166] enables us suitably reverse the above inequality. Specifically, by leveraging the relationship between fidelity and trace distance [9, Thm. 9.3.1] and the specific form of the 'canonical purification', we have

$$||\,|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}| - \left|\phi_{\rho_K^{\otimes n}}\right\rangle\left\langle\phi_{\rho_K^{\otimes n}}\right|\,||_1 \leq 4\sqrt[4]{||\sigma_{A^nK} - \rho_K^{\otimes n}||_1} \quad (13)$$

from [1, App. A, Lem. 14]. RHS of (13) is dealt in Sec. III-E.

D. Characterizing $\alpha = \mathfrak{E}_{\textit{sim}}(|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}|)$ and α_o

In characterizing α, we ought to evolve $|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}|$ through three CPTP maps that define $\mathfrak{E}_{\text{sim}}$ in (2). Referring to (9), (10), we consider the purification $|\phi_{\sigma_{A^nK}}\rangle \triangleq \sum_{t=1}^{d_A^n} \sum_{a\in[K]} \sqrt{K^{-1}\nu_{ta}}\, |x_{ta}\ a\ x_{ta}\ a\rangle$. From definition of $\mathscr{E}^{\theta}$, we have $(i_X^n \otimes i_{\mathcal{H}_K} \otimes \mathscr{E}^{\theta})(|\phi_{\sigma_{A^nK}}\rangle\langle\phi_{\sigma_{A^nK}}|) =$

$$\sum_{t=1}^{d_A^n}\sum_{v=1}^{d_A^n}\sum_{a\in[K]}\sum_{b\in[K]}\sum_{k\in[K]}\sum_{m\in[M]} K^{-1}\sqrt{\nu_{ta}\nu_{vb}}\, |x_{ta}\ a\rangle\langle x_{vb}\ b|$$
$$\mathrm{tr}(\theta_{k,m}\, |x_{ta}\rangle\langle x_{vb}|)\, \mathrm{tr}(|k\rangle\langle k|\, |a\rangle\langle b|)\, |k\ m\rangle\langle k\ m|$$
$$= \sum_{t=1}^{d_A^n}\sum_{v=1}^{d_A^n}\sum_{k\in[K]}\sum_{m\in[M]} K^{-1}\sqrt{\nu_{tk}\nu_{vk}}\, |x_{tk}\ k\rangle\langle x_{vk}\ k|$$
$$\langle x_{vk}|\theta_{m,k}|x_{tk}\rangle\, |k\ m\rangle\langle k\ m| \quad (14)$$
$$= \sum_{t=1}^{d_A^n}\sum_{v=1}^{d_A^n}\sum_{k\in[K]}\sum_{m\in[M]} K^{-1}\, |x_{tk}\ k\rangle\langle x_{vk}\ k|$$
$$\langle x_{vk}\sqrt{\nu_{vk}}|\theta_{m,k}|\sqrt{\nu_{tk}}x_{tk}\rangle\, |k\ m\rangle\langle k\ m| \quad (15)$$
$$= \sum_{t=1}^{d_A^n}\sum_{v=1}^{d_A^n}\sum_{k\in[K]}\sum_{m\in[M]} K^{-1}\, |x_{tk}\ k\rangle\langle x_{vk}\ k|$$
$$\left\langle x_{vk}\sqrt{T_k}\middle|\theta_{m,k}|\sqrt{T_k}x_{tk}\right\rangle |k\ m\rangle\langle k\ m| \quad (16)$$
$$= \sum_{k\in[K]}\sum_{m\in[M]} \frac{\left(\sqrt{T_k}\theta_{m,k}\sqrt{T_k}\right)^t}{K} \otimes |k\rangle\langle k| \otimes |k\ m\rangle\langle k\ m|$$
$$= \frac{1}{KM}\sum_{k\in[K]}\sum_{m\in[M]} \frac{(\sqrt{\omega}\mu_{k,m}\sqrt{\omega})^t}{\mathrm{tr}(\omega\mu_{k,m})} \otimes |k\rangle\langle k| \otimes |k\ m\rangle\langle k\ m| \quad (17)$$

where (15) follows from shifting scalars $\sqrt{\nu_{vk}}, \sqrt{\nu_{tk}}$, (16) follows from spectral decomposition in (9), (10), $(\cdot)^t$ denotes operator transpose and (17) follows from definition of $\theta_{m,k}$ and T_k in (7), (10) respectively. Next, we evolve the state in RHS of (17) through the CPTP map $(i_X^n \otimes i_{\mathcal{H}_K} \otimes \mathscr{E}^{\Delta})(\cdot)$ to yield the state

$$\frac{1}{KM}\sum_{\substack{k\in[K]\\ m\in[M]}} \sum_{\substack{y^n\in\\ \mathcal{Y}^n}} p_{Y|W}^n(y^n|c(k,m))\frac{\left(\sqrt{\omega}\mu_{k,m}\sqrt{\omega}\right)^t}{\mathrm{tr}(\omega\mu_{k,m})}\otimes |k\ y^n\rangle\langle k\ y^n|,$$

where the above follows from defn. (8). Finally, evolving above state through CPTP map $(i_X^n \otimes \mathrm{tr}_K \otimes \mathscr{E}^{\Delta})(\cdot)$ yields

$$\alpha = \frac{1}{KM}\sum_{\substack{(k,m,y^n)\in\\ [K]\times[M]\times\mathcal{Y}^n}} p_{Y|W}^n(y^n|c(k,m))\frac{\left(\sqrt{\omega}\mu_{k,m}\sqrt{\omega}\right)^t}{\mathrm{tr}(\omega\mu_{k,m})}\otimes |y^n\rangle\langle y^n|. \quad (18)$$

An identical sequence of steps yields[1]

$$\alpha_{\mathrm{o}} = \sum_{y^n\in\mathcal{Y}^n} \left(\sqrt{\omega}\lambda_{y^n}\sqrt{\omega}\right)^t \otimes |y^n\rangle\langle y^n|. \quad (19)$$

E. A new proof of the quantum covering lemma

Substituting (18), (19) and (13) in the RHS of (12), we have

$$||\alpha_{\mathrm{o}} - \alpha_{\mathrm{s}}||_1 \le ||\alpha_{\mathrm{o}} - \alpha||_1 + 4\sqrt[4]{||\sigma_{A^nK} - \rho_K^{\otimes n}||_1}. \quad (20)$$

Collating definitions of S_k from (7), T_k from (9) into σ_{A^nK} in (10), recalling $\rho_K^{\otimes n}$ (found after Defn. 1), recognizing the block diagonal structure of $\rho_K^{\otimes n}$ and σ_{A^nK}, we have

$$\begin{aligned}||\sigma_{A^nK} - \rho_K^{\otimes n}||_1 &= \frac{1}{K}\sum_{k=1}^{K}||\rho^{\otimes n} - \frac{1}{M}\sum_{m=1}^{M}\frac{\sqrt{\omega}\mu_{k,m}\sqrt{\omega}}{\mathrm{tr}(\omega\mu_{k,m})}||_1\\ &= \frac{1}{K}\sum_{k=1}^{K}||\rho^{\otimes n} - \frac{1}{M}\sum_{m=1}^{M}\beta_{c(k,m)}||_1 \quad (21)\end{aligned}$$

from the definition of β_w in (4) and $\beta_{c(k,m)} \triangleq \beta_{c(k,m)_1}\otimes\cdots\otimes\beta_{c(k,m)_n}$. Analogously defining $\gamma_{c(k,m)} \triangleq \gamma_{c(k,m)_1}\otimes\cdots\otimes\gamma_{c(k,m)_n}$, recognizing $\alpha = \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}\gamma_{c(k,m)}$ from (18) and $\alpha_{\mathrm{o}} = \left(\sum_{w\in\mathcal{W}}p_W(w)\gamma_w\right)^{\otimes n}$ from (6), (19), we have

$$||\alpha_{\mathrm{o}} - \alpha||_1 = ||\gamma^{\otimes n} - \frac{1}{KM}\sum_{k,m}\gamma_{c(k,m)}||_1. \quad (22)$$

where we have let $\gamma = \sum_{w\in\mathcal{W}}p_W(w)\gamma_w$. Noting $\rho = \sum_w p_W(w)\beta_w$, one recognizes similarity in the RHSs (21) and (22). Indeed, they are instances of the following QCL [9, Sec. 17.4]. In this article, we outline a new proof and refer the reader to [13] for details.

Lemma 1. *Suppose $p_X(\cdot)$ is a PMF on a finite set $\mathcal{X}$, $s_x \in \mathcal{D}(\mathcal{H}) : x\in\mathcal{X}$ and $s = \sum_x p_X(x)s_x \in \mathcal{D}(\mathcal{H})$. Suppose the 2^{nR} elements of $A = (X^n(1),\cdots,X^n(2^{nR}))$ are identically distributed according to $\mathcal{P}(X^n(i) = x^n) = p_X(n)(x^n)\ \forall x^n \in \mathcal{X}^n, \forall i\in[2^{nR}]$, and* ***pairwise independent****, then*

$$\mathbb{E}_{\mathcal{P}}\left\{||s^{\otimes n} - s(A)||_1\right\} \le \exp\left\{-\frac{n}{2}(R - \chi(s_x;p_X:\mathcal{X}))\right\} \quad (23)$$

where $s(A) \triangleq \frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}}s_{X^n(m)}$. In particular, there exists a map $c : [2^{nR_0}]\times[2^{nR_1}] \to \mathcal{X}^n$ such that

$$\frac{1}{2^{nR_0}}\sum_{k=1}^{2^{nR_0}}||s^{\otimes n} - \frac{1}{2^{nR_1}}\sum_{m=1}^{2^{nR_1}}s_{c(k,m)}||_1 \le 2^{-\frac{n\eta}{8}} \quad (24)$$

if $R_1 > \chi(s_x;p_X:\mathcal{X}) + 2\eta$.

Remark 2. Lem. 1 yields an achievability of Thm. 1 if one chooses (i) $R_0 = \frac{\log K}{n}, R_1 = \frac{\log M}{n}$ to bound (21) and (ii) $R_0 = 0$ and $R_1 = \frac{\log KM}{n}$ to bound (22).

Outline of a Proof: Let $s_x = \sum_y \gamma_{y|x}\left|e_{y|x}\right\rangle\left\langle e_{y|x}\right| : x\in\mathcal{X}, s = \sum_y p_Y(y)\left|f_y\right\rangle\left\langle f_y\right|$ be SCDs, $\pi_{x^n}^{\eta} \triangleq \pi_{x^n,p_X\gamma_{Y|X}}^{\eta}\mathbb{1}_{x^n\in T_\delta(p_X)}$ be a conditional typical projector of s_{x^n} and[2] $\pi^\eta \triangleq \pi_{p_Y}^{\eta}$ the (unconditional) typical projector of s. For $a = (x^n(m) : m\in[2^{nR}])$, let

$$s(a) \triangleq \sum_{m=1}^{2^{nR}}\frac{s_{x^n(m)}}{2^{nR}},\ \ w(a) \triangleq \sum_{m=1}^{2^{nR}}\frac{\pi^\eta\pi_{x^n(m)}^{\eta}s_{x^n(m)}\pi_{x^n(m)}^{\eta}\pi^\eta}{2^{nR}}$$

$$w \triangleq \sum_{x^n}p_X^n(x^n)\pi^\eta\pi_{x^n}^{\eta}s_{x^n}\pi_{x^n}^{\eta}\pi^\eta \text{ and note } w = \mathbb{E}_{\mathcal{P}}\{w(A)\}. \quad (25)$$

Let $s(A), w(A)$ be corresponding random quantities. The quantity of interest$\mathbb{E}_{\mathcal{P}}\{||s(A) - s^{\otimes n}||_1\} \le T_1 + T_2 + T_3$ where $T_1 = \mathbb{E}_{\mathcal{P}}\{||s(A) - w(A)||_1\}, T_2 = \mathbb{E}_{\mathcal{P}}\{||w(A) - w||_1\}, T_3 = \mathbb{E}_{\mathcal{P}}\{||w - s^{\otimes n}||_1\}$. T_1, T_3 are handled in a straight forward sequence of arguments leveraging (i) properties of typicality projectors, gentle operator lemma and the bound $||AB||_1 \le ||A||_1||B||_\infty$. These steps can be verified at [9, Sec. 17.4.3] or [13]. Analysis of the crucial term T_2 is where our proof differs. Taking a clue from Cuff's [14, Lem. 19] proof of classical covering, letting $v(A) \triangleq w(A) - \mathbb{E}_{\mathcal{P}}\{w(A)\}$, we have

$$\begin{aligned}T_2 &= \mathbb{E}_{\mathcal{P}}[||v(A)||_1] = \mathbb{E}_{\mathcal{P}}[\mathrm{tr}\{\sqrt{(v(A)^\dagger v(A)}\}] \quad (26)\\ &= \mathrm{tr}\{\mathbb{E}_{\mathcal{P}}[\sqrt{(v(A)^\dagger v(A)}]\} \le \mathrm{tr}\{\sqrt{\mathbb{E}_{\mathcal{P}}[(v(A)^\dagger v(A)]}\} \quad (27)\end{aligned}$$

where (26) follows from definition of trace, (27) from linearity of trace and the operator concavity [15, Thm. 2.6] of the square root function (a consequence of the Lowner-Heinz theorem [15, Thm. 2.6]). As fleshed out in [13], the RHS of (27) can be upper bounded as

$$\mathrm{tr}\{\sqrt{\mathbb{E}_{\mathcal{P}}[(v(A)^\dagger v(A)]}\} \le 2^{-\frac{n}{2}(R-\chi(p_X;s_x:\mathcal{X}))}, \quad (28)$$

thus completing the outline of our proof here.

IV. Simulation Via Algebraically Closed POVMs

We briefly revisit the simulated POVM in Sec. III. Alice possesses POVM operators $\underline{\theta} \triangleq \{\theta_{k,m} : 1 \le K, 0 \le m \le M\}$ and Bob has a corresponding table $\underline{c} \triangleq \{w^n(k,m) : 1 \le K, 0 \le m \le M\}$. On observing (common) random bits k^*, Alice performs POVM $\{\theta_{k^*,m} : 0 \le m \le M\}$. Bob chooses $w^n(k^*,m^*)$, where m^* is the POVM outcome, and evolves this through the stochastic matrix $p_{Y|W}^n(\cdot|w^n(k^*,m^*))$. The

[1] This is a standard computation and can be verified in [4, Proof of Lem. 4]

[2] Note that $\pi_{x^n}^{\eta} = 0$ if x^n is *not* typical

question we ask in this section is whether the table $\underline{c}$, and the corresponding POVMs $\underline{\theta}$ be endowed with certain *algebraic* structure? Specifically, suppose $\mathcal{W}$ is a finite field or a group, can the table $\underline{c}$ be chosen to be a coset of a linear code?

Optimal compression requires that each outcome $w^n(k,m)$ is an equally likely POVM outcome, forcing the entries of the table to be p_W-typical, where, we recall $p_W(w) = \mathrm{tr}(\rho\mu_w)$. Requiring table entries to be algebraically closed forces us to choose entries $w^n(k,m)$ that are *not* p_W-typical. Non-p_w-typical outcomes are extremely rare lending the simulation protocol sub-optimal (in terms of communication rates).

Does this imply that the table cannot have algebraic properties? There is one way to get around this obstacle. Can we add redundant operators and corresponding table entries in a controlled manner to guarantee algebraic closure, yet not suffer on the communication and common randomness rate? The idea is to enlarge the table into a third dimension. Let $\hat{c} : [K] \times [M] \times [B] \rightarrow \mathcal{W}^n$ be a map,

$$\hat{S}_k \triangleq \sum_{\substack{(b,m)\in\\ [B]\times[M]}} \frac{\sqrt{\omega}\mu_{k,m,b}\sqrt{\omega}}{\mathrm{tr}(\omega\mu_{k,m,b})}, \hat{\theta}_{k,m,b} \triangleq \frac{\hat{S}_k^{-\frac{1}{2}}\sqrt{\omega}\mu_{k,m,b}\sqrt{\omega}\hat{S}_k^{-\frac{1}{2}}}{\mathrm{tr}(\omega\mu_{k,m,b})}, \tag{29}$$

and $\hat{\theta} \triangleq \{\hat{\theta}_{k,m,b} \otimes |k\rangle\langle k| : 1 \leq k \leq K, 1 \leq m \leq M, 1 \leq b \leq B\}$ be a POVM. On observing random bits k^*, Alice performs POVM $\hat{\theta}_{k^*} \triangleq \{\theta_{k^*,m,b} : 1 \leq m \leq M, 1 \leq b \leq B\}$. Only the component m^* of the outcome (m^*, b^*) is communicated to Bob. If the table has the desired property that for each $(k,m) \in [K] \times [M]$, there is a *unique* index $b^*(k,m) \in [B]$ such that $w^n(k,m,b^*(k,m))$ is p_W-typical, then Bob can evolve $w^n(k^*, m^*, b^*(k^*, m^*))$ through the stochastic matrix $p^n_{Y|W}(\cdot|w^n(k^*, m^*, b^*(k^*, m^*)))$ and simulate the POVM outcome. Since the POVM outcome is p_W-typical WHP, Bob's choice would indeed be the correct POVM outcome WHP.

This provides us with the clue. For simplicity, let us assume $\mathcal{W} = \mathbb{F}_2$ to be the binary field. In order to prove achievability in Lem. 1, we picked entries of table $\underline{c}$ independently and randomly with distribution p^n_W. Instead, suppose we let table $\hat{C}$ to be range of a generator matrix $G \in \mathbb{F}_2^{(c+r+\beta)\times n}$ whose entries are picked uniformly independently from $\mathbb{F}_2 = \{0, 1\}$, then its range is a random linear code with uniform *pairwise independent* codewords. Common randomness specifies c bits. For each choice of these c bits, we build a POVM with $2^{r+\beta}$ operators. Only r of the $(r+\beta)$ outcome bits is communicated to Bob. Having been provided $c + r$ bits, Bob looks for a unique collection of β bits for which the corresponding entry in $\hat{C}$ that is p_W-typical. Since the entries of $\hat{C}$ are uniformly distributed, the expected number of p_W-typical codewords in any collection of 2^β entries is $\frac{2^\beta|T_\eta(p_W)|}{2^n} = 2^{-n(1-H(p_W)-\beta)}$. Therefore, so long as $\beta < 1 - H(p_W)$, it is natural to expect that Bob will find just one p_W-typical entry whose index agrees with the $c+r$ bits he has been provided. This suggests that, if we can enlarge our table by a factor not greater than $2^{n(1-H(p_W))}$ and prove a QCL analogous to Lem. 1, but with entries of the table A uniformly chosen from $\mathbb{F}_2^n$, instead of p^n_W, then we can perform POVM simulation with an '*algebraically closed POVM*'. This is indeed true. A proof of Lem. 2 is similar to proof of Lem. 1 and is provided in [13]. We follow this up with a final statement on the existence of structured POVMs for simulation. See [13] for a proof.

Lemma 2. *Suppose $p_X(\cdot)$ is a PMF on a finite field $\mathcal{X} = \mathcal{F}_q$ of size q, $s_x \in \mathcal{D}(\mathcal{H}) : x \in \mathcal{X}$ and $s = \sum_x p_X(x)s_x \in \mathcal{D}(\mathcal{H})$. Suppose the q^{nR} elements of $A = (X^n(1), \cdots, X^n(q^{nR}))$ are uniformly distributed and pairwise independent, then* $\mathbb{E}_{\mathcal{U}}\{||s^{\otimes n} - s(A)||_1\} \leq \exp\left\{-\frac{n\left[R-\chi(s_x;p_X:\mathcal{X})-\log q+H(p_W)\right]}{2}\right\}$

*where $s(A) \triangleq \frac{1}{q^{nR}}\sum_{m=1}^{q^{nR}} s_{X^n(m)}$. In particular, $\exists$ a map $c : [q^{nR_0}] \times [q^{nR_1}] \rightarrow \mathcal{X}^n$ whose **range is a coset** such that*

$$\frac{1}{q^{nR_0}}\sum_{k=1}^{q^{nR_0}} ||s^{\otimes n} - \frac{1}{q^{nR_1}}\sum_{m=1}^{q^{nR_1}} s_{c(k,m)}||_1 \leq 2^{-\frac{n\eta}{8}} \tag{30}$$

if $R_1 > \chi(s_x; p_X : \mathcal{X}) + \log q - H(p_W) + 2\eta$.

Theorem 2. *Let $\rho \in \mathcal{D}(\mathcal{H}_A), \lambda_\mathcal{Y} \in \mathscr{M}(\mathcal{H}_A, \mathcal{Y})$ and $(\mathcal{W} = \mathcal{F}_q, \mu_\mathcal{W}, p_{Y|W}) \in \mathcal{C}(\rho, \lambda_\mathcal{Y})$, where $\mathcal{W} = \mathcal{F}_q$ is a finite field with q elements. Suppose $c + r + \beta > \mathscr{RC}(\mathcal{W} = \mathcal{F}_q, \mu_\mathcal{W}, p_{Y|W}) + \log q - H(p_W)$ and $r + \beta > \mathscr{R}(\mathcal{W} = \mathcal{F}_q, \mu_\mathcal{W}, p_{Y|W}) + \log q - H(p_W)$, where $p_W(w) = \mathrm{tr}(\rho\mu_w) : w \in \mathcal{W}$, then there exists a $\hat{c} : [q^{nc}] \times [q^{nr}] \times [q^{n\beta}] \rightarrow \mathcal{W}^n$ whose range is a coset for which the POVM $\hat{\theta} \triangleq \{\hat{\theta}_{k,m,b} \otimes |k\rangle\langle k| : 1 \leq k \leq q^{nc}, 1 \leq m \leq q^{nr}, 1 \leq b \leq q^{n\beta}\}$ defined through (29) simulates POVM $\lambda_\mathcal{Y}$ on ρ with communication cost $(\mathscr{RC}(\mathcal{W}, \mu_\mathcal{W}, p_{Y|W}), \mathscr{R}(\mathcal{W}, \mu_\mathcal{W}, p_{Y|W}))$.*

References

[1] A. Winter, "“Extrinsic”and “Intrinsic” Data in Quantum Measurements: Asymptotic Convex Decomposition of Positive Operator Valued Measures," Commn. in Math. Phy., vol. 244, no. 1, pp. 157–185, 2004.

[2] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," IEEE Trans. on Info. Th., vol. 48, no. 3, 2002.

[3] H. J. Groenewold, "A Problem of Information Gain by Quantal Measurements," Intl. Jrnl of Theoretical Physics, pp. 327–338, Sept. 1971.

[4] M. M. Wilde, P. Hayden, F. Buscemi, and M.-H. Hsieh, "The information-theoretic costs of simulating quantum measurements," Jrnl of Phy. A: Math. and Theoretical, vol. 45, no. 45, 2012.

[5] S. Massar and S. Popescu, "Amount of information obtained by a quantum measurement," Phys. Rev. A, vol. 61, p. 062303, May 2000.

[6] A. Winter and S. Massar, "Compression of quantum-measurement operations," Phys. Rev. A, vol. 64, p. 012311, Jun 2001.

[7] A. S. Holevo, Quantum Systems, Channels, Information, 2nd ed. De Gruyter, 2019.

[8] M. Heidari, T. A. Atif, and S. Sandeep Pradhan, "Faithful simulation of distributed quantum measurements with applications in distributed rate-distortion theory," in 2019 IEEE International Symposium on Information Theory (ISIT), 2019, pp. 1162–1166.

[9] M. M. Wilde, Quantum Information Theory, 2nd ed. Cambridge University Press, 2017.

[10] A. Anshu, R. Jain, and N. A. Warsi, "Convex-split and hypothesis testing approach to one-shot quantum measurement compression and randomness extraction," IEEE Trans. on Info. Th., vol. 65, no. 9, 2019.

[11] A. Anshu, V. K. Devabathini, and R. Jain, "Quantum communication using coherent rejection sampling," Phy. Rev. Let., vol. 119, 2017.

[12] S. Chakraborty, A. Nema, and P. Sen, "One-shot inner bounds for sending private classical information over a quantum mac," 2021.

[13] A. Padakandla, "Optimal Simulation of Quantum Measurements via the Likelihood POVMs," arXiv preprint arXiv:2109.12586, 2021.

[14] P. W. Cuff, "Communication in networks for coordinating behavior," Ph.D. dissertation, Stanford, CA, USA, 2009.

[15] E. Carlen, "Trace inequalities and quantum entropy: An introductory course," vol. 529, 01 2010.

Classical State Masking over a Quantum Channel

Uzi Pereg
Institute for Communications Engineering
Technical University of Munich & MCQST
uzi.pereg@tum.de

Christian Deppe
Institute for Communications Engineering
Technical University of Munich
christian.deppe@tum.de

Holger Boche
Theoretical Information Technology
Technical University of Munich
MCQST & CASA
boche@tum.de

***Abstract*—Transmission of classical information over a quantum state-dependent channel is considered, when the encoder can measure channel side information (CSI) and is required to mask information on the quantum channel state from the decoder. In this quantum setting, it is essential to conceal the CSI measurement as well. A regularized formula is derived for the masking equivocation region, and a full characterization is established for a class of measurement channels.**

I. Introduction

Security and privacy are critical aspects in modern communication systems [1]. The classical wiretap channel was first introduced by Wyner [2] to model communication in the presence of a passive eavesdropper. On the other hand, Merhav and Shamai [3] introduced a different communication system with the privacy requirement of masking. In this setting, the sender transmits a sequence X^n over a memoryless state-dependent channel $p_{Y|X,S}$, where the state sequence S^n has a fixed memoryless distribution and is not affected by the transmission. The transmitter of X^n is informed of S^n and is required to send information to the receiver while limiting the amount of information that the receiver can learn about S^n. The masking setting can also be viewed as communication with an untrusted party, where Alice wishes to send Bob a limited amount of information, and keep the source hidden [4, 5]. Related settings are also considered in [6–8].

The field of quantum information is rapidly evolving in both practice and theory [9]. Communication through quantum channels can be separated into different categories. For classical communication, the Holevo-Schumacher-Westmoreland (HSW) Theorem provides a regularized ("multi-letter") formula for the capacity of a quantum channel [10, 11]. Although calculation of such a formula is intractable in general, it provides computable lower bounds, and there are special cases where the capacity can be computed exactly.

Another scenario of interest is when Alice and Bob share entanglement resources. While entanglement can be used to produce shared randomness, it is a much more powerful aid [12]. E.g., using super-dense coding, entanglement assistance doubles the transmission rate of classical messages over a noiseless qubit channel. The entanglement-assisted capacity of a noisy quantum channel was fully characterized by Bennett *et al.* [13] in terms of the quantum mutual information.

Boche *et al.* [14] addressed the classical-quantum channel with channel state information (CSI) at the encoder. The capacity was determined given causal CSI, and a regularized formula was provided given non-causal CSI [14]. The first author [15] extended the results to a quantum-quantum channel with random parameters, and further considered communication over quantum channels with parameter estimation at the receiver, given either strictly-causal, causal, or non-causal CSI at the encoder, and without CSI as well. The entanglement-assisted capacity of a quantum channel with non-causal CSI was determined by Dupuis in [16], and with causal CSI in [17]. Considering secure communication over the quantum wiretap channel, Cai *et al.* [18] established a regularized characterization of the secrecy capacity.

In quantum channel state masking, analogously to the classical model, the channel state system C store undesired quantum information which leaks to the receiver [3]. This could model a leakage in the system of secret information, or could stand for another transmission to another receiver (Charlie), with a product state, out of our control, and which is not intended to our receiver (Bob), and is therefore to be concealed from him. Thus, the goal of the transmitter (Alice) now is to try and mask this undesired information as much as possible on the one hand, and to transmit reliable independent information rate on the other. Masking can also be viewed as a building block for cryptographic problems of oblivious transfer of information and secure computation by untrusting parties. In a recent paper by the authors [19], we have considered a quantum state-dependent channel, when the encoder has CSI and is required to mask information on the quantum channel state from the decoder. We have established a full characterization for the entanglement-assisted masking equivocation region with maximally correlated channel state systems, and a regularized formula for the quantum masking region without assistance.

In this paper, we consider a similar model of a quantum state-dependent channel $\mathcal{N}_{EA\to B}$, when the encoder has CSI and is required to mask information on the quantum channel state from the decoder. We derive a regularized formula for the classical masking region and establish full characterization for a class of measurement channels. Here, however, the communication task is to send classical information, while there are no entanglement resources available to Alice and Bob. Specifically, the channel state systems are in an entangled state $|\phi_{E_0EC}\rangle^{\otimes n}$. Alice wishes to send a classical message m. To this end, she measures the CSI systems E_0^n and obtains an outcome V. Based on the measurement outcome, Alice encodes the quantum state of the channel input systems A^n in

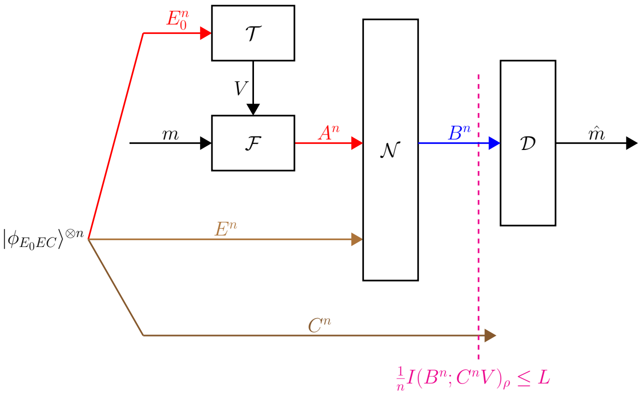


Fig. 1. Coding for a quantum state-dependent channel $\mathcal{N}_{EA\to B}$ given state information at the encoder and masking from the decoder. The quantum systems of Alice and Bob are marked in red and blue, respectively. The channel state systems E^n and C^n are marked in brown.

such a manner that limits the leakage-rate of Bob's information on C^n from B^n, while the systems E_0^n and C^n are entangled with the channel state systems E^n (see Figure 1).

The quantum model involves three channel state systems, E^n, E_0^n, and C^n, as opposed to the classical case [3] of a single random parameter. The systems E_0^n can be thought of as part of the environment of both our transmitter and the source of C^n, possibly entangled if they had previous interaction, while E^n belong to the channel's environment. Another significant distinction from the classical case is that the measurement can cause a collapse of the wave function, hence correlations can be lost. Thereby, it is essential to conceal the CSI observation as well. In the present model, the leakage requirement involves both the masked system C^n and the measurement outcome V. Those subtleties do not exist in the classical problem.

The full manuscript with proofs can be found in [20].

II. Definitions and Related Work

A. Definitions

The quantum state of a system A is a density operator ρ_A on the Hilbert space $\mathcal{H}_A$. A pure state $|\psi_{AB}\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ is called *entangled* if it cannot be expressed as a product of states. Define the quantum entropy, conditional entropy, and mutual information, by $H(A)_\rho \triangleq -\text{Tr}[\rho_A \log(\rho_A)]$, $H(A|B)_\rho = H(AB)_\rho - H(B)_\rho$, and $I(A;B)_\rho = H(A)_\rho + H(B)_\rho - H(AB)_\rho$, respectively.

A quantum state-dependent channel is defined as a linear cptp map $\mathcal{N}_{EA\to B}$. Both the channel state systems and the quantum channel have a product form, $|\phi_{EE_0C}\rangle^{\otimes n}$ and $\mathcal{N}_{EA\to B}^{\otimes n}$. Given CSI, the transmitter has access to E_0^n. We will consider a secrecy requirement that limits the information that the receiver can obtain on C^n. We will also be interested in the special case of a measurement channel, $\mathcal{M}_{EA\to Y}$, *i.e.* a channel with a classical output.

A $(2^{nR}, n)$ classical masking code with CSI at the encoder consists of the following: A message set $[1:2^{nR}]$, assuming 2^{nR} is an integer, an encoding measurement $\mathcal{T} \equiv \{T_{E_0^n}^v\}$, on the CSI system E_0^n, an encoding map $\mathcal{F} : (m,v) \mapsto \rho_{A^n}$, and a decoding measurement $\mathcal{D} \equiv \{D_{B^n}^{\hat{m}}\}$.

Alice chooses a message $m \in [1:2^{nR}]$ uniformly at random. She measures the CSI systems E_0^n, which are entangled with the channel state systems, using the measurement set $\mathcal{T}$, and obtains a measurement outcome v. Then, Alice prepares the input state $\rho_{A^n}^{m,v} = \mathcal{F}(m,v)$, and transmits the systems A^n over n channel uses of $\mathcal{N}_{EA\to B}$. See Figure 1. The average post-measurement input state is $\bar{\rho}_{C^nE^nVA^n}^m = \sum_v \text{Tr}_{E_0^n}\left(T_{E_0^n}^v \phi_{CEE_0}^{\otimes n}\right) \otimes |v\rangle\langle v| \otimes \rho_{A^n}^{m,v}$, and the output is $\rho_{C^nVB^n}^m = \mathcal{N}_{EA\to B}^{\otimes n}(\bar{\rho}_{C^nVE^nA^n}^m)$. Bob receives B^n and measures his estimate $\hat{m}$ for the message. The average probability of decoding error is

$$P_e^{(n)}(\mathcal{F},\mathcal{T},\mathcal{D}) = \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \left[1 - \text{Tr}\left(D_{B^n}^m \rho_{B^n}^m\right)\right] \tag{1}$$

and the masking leakage rate is defined as

$$\ell^{(n)}(\mathcal{F},\mathcal{T},\mathcal{D}) \triangleq \frac{1}{n} I(C^nV;B^n)_\rho. \tag{2}$$

A $(2^{nR}, n, \varepsilon, L)$ masking code satisfies $P_e^{(n)}(\mathcal{F},\mathcal{T},\mathcal{D}) \leq \varepsilon$ and $\ell^{(n)}(\mathcal{F},\mathcal{T},\mathcal{D}) \leq L$. A rate-leakage pair (R,L) is called achievable if for every $\varepsilon, \delta > 0$ and large n, there exists a $(2^{nR}, n, \varepsilon, L+\delta)$ masking code. The classical masking region $\mathbb{R}_{\text{CL}}(\mathcal{N})$ is defined as the set of achievable pairs (R,L).

Notice that the leakage rate (2) includes Alice's observation V of the CSI systems.

B. Related Work

We briefly review known results for the case where there is no masking requirement. First, consider a quantum channel $\mathcal{P}_{A\to B}$ without a channel state. Define

$$\chi(\mathcal{P}) \triangleq \max_{p_X(x), |\phi_A^x\rangle} I(X;B)_\rho \tag{3}$$

with $\rho_{XB} \equiv \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x| \otimes \mathcal{P}(|\phi_A^x\rangle\langle\phi_A^x|)$ and $|\mathcal{X}| \leq |\mathcal{H}_A|^2$. The formula above is sometimes referred to as the Holevo information of the channel [21].

Theorem 1 (see [10, 11]). The classical capacity of a quantum channel $\mathcal{P}_{A\to B}$ that does not depend on a channel state, without a masking requirement, is given by

$$\mathbb{C}_{\text{Cl}}(\mathcal{P}, \infty) = \lim_{n\to\infty} \frac{1}{n} \chi\left(\mathcal{P}^{\otimes n}\right).$$

A single-letter characterization is an open problem for a general quantum channel. Although calculation of a regularized formula is intractable in general, it provides a computable lower bound, and there are special cases where the capacity can be computed exactly [22].

Next, we move to a quantum state-dependent channel $\mathcal{N}_{EA\to B}$ with CSI at the encoder, in the special case where the state is a classical random parameter $S \sim q(s)$. Let

$$\mathsf{R}(\mathcal{N}, \infty) \triangleq \sup_{p_{X|S}(x|s), \varphi_A^x} \left[I(X;B)_\rho - I(X;S)\right] \tag{4}$$

where the supremum is over the ensembles $\{p_{X|S}, \varphi_A^x\}$, with $\rho_{SXB} \equiv \sum_{s,x} q(s)p_{X|S}(x|s)|s,x\rangle\langle s,x| \otimes \mathcal{N}_{SA\to B}(|s\rangle\langle s| \otimes \varphi_A^x)$.

Theorem 2 (see [23]). The classical capacity of a random-parameter quantum channel $(\mathcal{N}_{SA\to B}, S \sim q(s))$, with CSI at the encoder and without a masking requirement, is given by

$$\mathbb{C}_{\text{Cl}}(\mathcal{N}, \infty) = \lim_{n\to\infty} \frac{1}{n} \mathsf{R}\left(\mathcal{N}^{\otimes n}, \infty\right).$$

III. Main Results

We state our results on channel state masking for the quantum state-dependent channel $\mathcal{N}_{EA\to B}$. We establish a regularized formula for the classical masking region and capacity-leakage function for the transmission of classical information over $\mathcal{N}_{EA\to B}$. For the special class of measurement channels, we obtain a single-letter formula.

Define

$$\mathcal{R}_{\text{Cl}}(\mathcal{N}) = \bigcup \left\{ \begin{array}{ll} (R,L):\ 0 \le R & \le I(X;B)_\rho - I(X;S) \\ \qquad\qquad\quad L & \ge I(CS;XB)_\rho \end{array} \right\} \tag{5}$$

where the union is over the POVMs $\{\Lambda_{E_0}^s\}$, the conditional distributions $p_{X|S}$, and the collections of input states φ_A^x, with

$$\rho_{ECSXA} = \sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} p_{X|S}(x|s) \cdot \\ \text{Tr}_{E_0}(\Lambda_{E_0}^s \phi_{E_0EC}) \otimes |s,x\rangle\langle s,x| \otimes \varphi_A^x \tag{6}$$

and $\rho_{BCSX} = \mathcal{N}_{EA\to B}(\rho_{EACSX})$. Our main result on channel state masking is given below.

Theorem 3.

1) The classical masking region of a quantum state-dependent channel $(\mathcal{N}_{EA\to B}, |\phi_{EE_0C}\rangle)$ with CSI at the encoder is given by

$$\mathbb{R}_{\text{Cl}}(\mathcal{N}) = \bigcup_{n=1}^{\infty} \frac{1}{n} \mathcal{R}_{\text{Cl}}(\mathcal{N}^{\otimes n}).$$

2) For a measurement channel $\mathcal{M}_{EA\to Y}$ with a classical CSI system $E_0 \equiv S$,

$$\mathbb{R}_{\text{Cl}}(\mathcal{M}) = \\ \bigcup_{p_{X|S},\, \varphi_A^x} \left\{ \begin{array}{ll} (R,L):\ 0 \le R & \le I(X;Y) - I(X;S) \\ \qquad\qquad\quad L & \ge I(CS;XY)_\rho \end{array} \right\}.$$

We only give the proof outline for the direct part in Section IV, while the full proof can be found in [20].

Remark 1. In [20, Appendix A], we show that the union can be exhausted with cardinality $|\mathcal{X}| \le (|\mathcal{H}_A|^2 + 1)|\mathcal{H}_E|$, using Fenchel-Eggleston-Carathéodory lemma and similar arguments as in [23]. Hence, in principle, the region $\mathcal{R}_{\text{Cl}}(\mathcal{N})$ is computable. Nevertheless, for a general quantum channel, we have only obtained a regularized characterization. As mentioned in Section II-B, a single-letter capacity formula is an open problem, even without a channel state.

Equivalently, we can characterize the capacity-leakage function. The following corollary is an immediate consequence.

Corollary 4.

1) The classical capacity-leakage function of a quantum state-dependent channel $(\mathcal{N}_{EA\to B}, |\phi_{EE_0C}\rangle)$ with CSI at the encoder is given by

$$\mathbb{C}_{\text{Cl}}(\mathcal{N}, L) = \\ \lim_{n\to\infty} \frac{1}{n} \sup_{\substack{\Lambda_{E_0^n}^s,\, p_{X|S},\, \varphi_{A^n}^x : \\ I(C^nS;XB^n)_\rho \le L}} \left[I(X;B^n)_\rho - I(X;S)\right].$$

2) For a measurement channel $\mathcal{M}_{EA\to B}$ with a classical CSI system $E_0 \equiv S$,

$$\mathbb{C}_{\text{Cl}}(\mathcal{M}, L) = \\ \sup_{p_{X|S},\, \varphi_A^x : I(CS;XY)_\rho \le L} \left[I(X;Y) - I(X;S)\right].$$

To illustrate our results, we give a simple example.

Example 1. Consider a qubit channel $\mathcal{N}_{SA\to B}$ that depends on a classical random parameter $S \sim \text{Bernoulli}(\varepsilon)$, hence $E_0 \equiv E \equiv C \equiv S$. Such a random-parameter quantum channel can be viewed as a random selection from a set of quantum channels, $\{\mathcal{N}_{A\to B}^s\}_{s=0,1}$. Let

$$\mathcal{N}^{(0)}(\rho) = \rho \tag{7}$$

$$\mathcal{N}^{(1)}(\rho) = |\psi\rangle\langle\psi| \tag{8}$$

where $|\psi\rangle$ is a given state in the same qubit space. In other words, the parameter S_i chooses whether the ith input system is projected onto $|\psi\rangle$. Ignoring the CSI at the encoder, the model resembles the quantum erasure channel [21], except that $|\psi\rangle$ here is in the qubit space, whereas the "erasure state" is orthogonal to it. Nonetheless, we note that if the decoder knows the locations where the state is projected, then this model is equivalent to the quantum erasure channel. Without this knowledge at the decoder, it is less obvious.

By Theorem 3, the following rate-leakage region is achievable for the random-parameter channel above,

$$\mathbb{C}_{\text{Cl}}(\mathcal{N}) \supseteq \bigcup_{0\le\alpha\le\frac{1}{2}} \left\{ \begin{array}{l} (R,L)\ :\ R \le (1-\varepsilon)h(\alpha), \\ L \ge h\left((1-\varepsilon)\alpha\right) - (1-\varepsilon)h(\alpha) \end{array} \right\} \tag{9}$$

where $h(x)$ is the binary entropy function. We can see the tradeoff between the communication rate and the leakage. Clearly, if the encoder constantly transmits $|\psi\rangle$, then there is no leakage, as the output is $|\psi\rangle \otimes \cdots \otimes |\psi\rangle$. Yet, the rate is zero as well. Indeed, for $\alpha = 0$, we achieve $(R,L) = (0,0)$. On the other hand, taking $\alpha = \frac{1}{2}$, we obtain the maximal rate $R = 1 - \varepsilon$, which is also the capacity of the quantum erasure channel. However, the leakage is positive.

To show achievability, note that the bound on the rate on the RHS of (5) can also be expressed as $R \le H(X|S) - H(X|B)_\rho$. Let the input ensemble be the basis $\{|\psi\rangle, |\psi_\perp\rangle\}$, where $|\psi_\perp\rangle$ is orthogonal to $|\psi\rangle$. The input distribution is chosen as follows. Let $V \sim \text{Bernoulli}(\alpha)$ be independent of S. If $S = 0$, set $X = V$. Otherwise, if $S = 1$, then $X = 0$.

IV. Proof Outline

To derive our results, we use the quantum method of types. Let $\rho_A = \sum_x p_X(x)|x\rangle\langle x|$. Define the δ-typical projector as $\Pi^\delta(\rho_A) \equiv \sum_{x^n \in \mathcal{A}^\delta(p_X)} |x^n\rangle\langle x^n|$, where $\mathcal{A}^\delta(p_X)$ is the classical δ-typical set [21]. For every $\varepsilon, \delta > 0$ and large n,

$$\Pi^\delta(\rho_A)\, \rho_A^{\otimes n}\, \Pi^\delta(\rho_A) \preceq 2^{-n(H(A)_\rho - c\delta)} \tag{10}$$

where $c > 0$ is a constant. The conditional δ-typical projector $\Pi^\delta(\sigma_B|x^n)$ of an average state $\sigma = \sum_{x\in\mathcal{Y}} p_X(x)\rho_B^x$ is defined as in [21, Definition 15.2.3]. For every $\varepsilon', \delta > 0$ and sufficiently large n [21],

$$\begin{aligned} \mathrm{Tr}(\Pi^\delta(\sigma_B|x^n)\rho_{B^n}^{x^n}) &\geq 1 - \varepsilon' && (11)\\ \mathrm{Tr}(\Pi^\delta(\sigma_B|x^n)) &\leq 2^{n(H(B|X')_\sigma + c'\delta)} && (12)\\ \mathrm{Tr}(\Pi^\delta(\sigma_B)\rho_{B^n}^{x^n}) &\geq 1 - \varepsilon'\,. && (13) \end{aligned}$$

for all $x^n \in \mathcal{A}^\delta(p_X)$, where $\rho_{B^n}^{x^n} = \bigotimes_{i=1}^n \rho_{B_i}^{x_i}$, and X' is distributed according to the type of x^n.

To show achievability, we extend the classical binning technique to the quantum setting, and then apply the quantum packing lemma and the classical covering lemma.

1) Classical Code Construction: Let $\delta > 0$, and let $\widetilde{R} > R$ be chosen later. For every $m \in [1 : 2^{nR}]$, select a sub-codebook of $2^{n(\widetilde{R}-R)}$ independent sequences, $\mathscr{B}(m) = \{x^n(k) \,:\, k \in [(m-1)2^{n(\widetilde{R}-R)} + 1 : m2^{n(\widetilde{R}-R)}]\}$, each according to $\prod_{i=1}^n p_X(x_i)$.

2) Encoding: To send a message m,

(i) Measure the CSI systems $E_{0,i}$ using the POVM $\Lambda_{E_0}^s$, for $i \in [1 : n]$. Since the CSI systems are in a product state, the measurement outcome is an i.i.d. sequence $\sim q(s)$, where $q(s) = \mathrm{Tr}(\Lambda_{E_0}^s \sigma_{E_0})$.

(ii) Given a measurement outcome s^n, find a sequence $x^n(k) \in \mathscr{B}(m)$ such that $(s^n, x^n(k)) \in \mathcal{A}^\delta(p_{S,X})$, where $p_{S,X}(s,u) = q(s)p_{X|S}(u|s)$. If there is more than one, choose the first, and if none $x^n(1)$.

(iii) Transmit $\rho_{A^n}^m = \bigotimes_{i=1}^n \varphi_A^{x_i(k)}$.

3) Decoding: Decode $\hat{k}$ by applying a POVM $\{\Lambda_k\}_{k\in[1:2^{n\tilde{R}}]}$, which will be specified later. Declare the estimate $\hat{m}$ such that $x^n(\hat{k}) \in \mathscr{B}(\hat{m})$.

Analysis of Probability of Error and Leakage: First, consider the error probability. By symmetry, we may assume w.l.o.g. that Alice sends the message $M = 1$ using K. Consider the following events,

$$\begin{aligned} \mathscr{E}_1 &= \{(S^n, X^n(k')) \notin \mathcal{A}^{\delta_1}(p_{S,X})\,,\ \text{for all } k' \in \mathscr{B}(1)\} && (14)\\ \mathscr{E}_2 &= \{\hat{K} \neq K\} && (15) \end{aligned}$$

with $\delta_1 \equiv \delta/|\mathcal{S}|$. By the union of events bound,

$$P_e^{(n)}(\mathcal{T}, \mathcal{F}, \mathcal{D}) \leq \Pr(\mathscr{E}_1) + \Pr(\mathscr{E}_2 \mid \mathscr{E}_1^c)\,. \tag{16}$$

By the classical covering lemma [24, Lemma 3.3], the first term tends to zero as $n \to \infty$, if $\widetilde{R} - R > I(X;S) + \varepsilon_1(\delta)$. Hence, we choose

$$\widetilde{R} = R + I(X;S) + 2\varepsilon_1(\delta)\,. \tag{17}$$

Next, we use the quantum packing lemma. Given $\mathscr{E}_1^c$, we have $X^n(K) \in \mathcal{A}^\delta(p_X)$. Now,

$$\begin{aligned} \Pi^\delta(\rho_B)\rho_{B^n}\Pi^\delta(\rho_B) &\preceq 2^{-n(H(B)_\rho - \varepsilon_2(\delta))}\Pi^\delta(\rho_B) && (18)\\ \mathrm{Tr}\left[\Pi^\delta(\rho_B|x^n)\rho_{B^n}^{x^n}\right] &\geq 1 - \varepsilon_2(\delta) && (19)\\ \mathrm{Tr}\left[\Pi^\delta(\rho_B|x^n)\right] &\leq 2^{n(H(B|X)_\rho + \varepsilon_2(\delta))} && (20)\\ \mathrm{Tr}\left[\Pi^\delta(\rho_B)\rho_{B^n}^{x^n}\right] &\geq 1 - \varepsilon_2(\delta) && (21) \end{aligned}$$

for all $x^n \in \mathcal{A}^{\delta_1}(p_X)$, by (10), (11), (12), and (13), respectively. Thus, by the quantum packing lemma [21, Lemma 16.3.1], there exists a POVM D_k such that $\Pr(\mathscr{E}_2 \mid \mathscr{E}_1^c) \leq 2^{-n(I(X;B)_\rho - \widetilde{R} - \varepsilon_3(\delta))}$, which tends to zero as $n \to \infty$, if $\widetilde{R} < I(X;B)_\rho - \varepsilon_3(\delta)$. Hence, by (17), the probability of decoding error tends to zero, provided that $R < I(X;B)_\rho - I(X;S) - \varepsilon_3(\delta) - 2\varepsilon_2(\delta)$.

As for the leakage rate, observe that

$$\begin{aligned} I(C^n;B^n)_\rho &\leq I(C^n;X^n(K),B^n)_\rho\\ &= I(C^n;X^n(K))_\rho + I(C^n;B^n|X^n(K))_\rho\,. && (22) \end{aligned}$$

Then, the first term is bounded by

$$\begin{aligned} I(C^n;M,X^n(K))_\rho &\overset{(a)}{=} I(C^n;X^n(K)|M)_\rho\\ &\leq H(X^n(K)|M)_\rho\\ &\overset{(b)}{\leq} n(\widetilde{R} - R)\\ &\overset{(c)}{=} n(I(X;S) + 2\varepsilon_1(\delta))\\ &\leq n(I(X;C,S) + 2\varepsilon_1(\delta)) && (23) \end{aligned}$$

where (a) holds since there is no correlation between M and C^n, (b) follows as $|\mathscr{B}(M)| = 2^{n(\widetilde{R}-R)}$, and (c) is due to (17). Moving to the second term in the RHS of (22),

$$\begin{aligned} I(C^n;B^n|X^n(K))_\rho &\leq I(C^n,S^n;B^n|X^n(K))_\rho\\ &= H(B^n|X^n(K))_\rho - H(B^n|C^n,S^n,X^n(K))_\rho\\ &\leq nH(B|X)_\rho - H(B^n|C^n,S^n,X^n(K))_\rho && (24) \end{aligned}$$

where the last inequality holds as $H(B^n|X^n(K))_\rho \leq \sum_{i=1}^n H(B_i|X_i(K))_\rho = nH(B|X)_\rho$, since conditioning does not increase the quantum entropy. Furthermore, given $X^n(K) = x^n$ and $S^n = s^n$, we have a product output state $\bigotimes_{i=1}^n \mathcal{N}_{EA\to B}(\sigma_{EC}^{s_i} \otimes \varphi_A^{x_i,s_i})$, where σ_{EC}^s denotes the post-measurement state. Thus, it follows from (22)-(24) that

$$\begin{aligned} \frac{1}{n}I(B^n;C^n) &\leq I(C,S;X) + I(C,S;B|X) + 2\varepsilon_1(\delta)\\ &= I(C,S;X,B) + 2\varepsilon_1(\delta)\,. && (25) \end{aligned}$$

Thereby, the leakage requirement holds if $I(C,S;X,B) \leq L - 2\varepsilon_1(\delta)$. To show that $\frac{1}{\kappa}\mathcal{R}_{\mathrm{Cl}}(\mathcal{N}^{\otimes\kappa})$ is achievable as well, employ the coding scheme above for the product channel $\mathcal{N}^{\otimes\kappa}$. This completes the proof of the direct part.

V. Summary and Concluding Remarks

We consider classical communication over a quantum state-dependent channel $\mathcal{N}_{EA\to B}$, when the encoder measures side information and is required to mask information from the decoder. This could model a leakage in the system of secret information or a transmission to another receiver. In [19], we have considered a similar model with entanglement assistance, and derived a regularized formula for the quantum masking region without assistance. Here, we have removed the entanglement assistance, and considered the transmission of *classical* information.

Masking can also be viewed as a building block for cryptographic problems of oblivious transfer of information, such as bit commitment or secure computation. Suppose that Alice is a server that receives a query to perform a task on a quantum computer, while also using a private source $E_0^n C^n$. To this end, Alice uses E_0^n to encode A^n, including a reference number m (metadata). Next, she performs the computation map $\mathcal{N}_{EA\to B}^{\otimes n}$ on the systems $E^n A^n$, which are entangled with the private source. The quantum output system B^n is delivered to the agent Bob, who performs a measurement to view the metadata m, and then use B^n as he wishes. The masking requirement is to prevent Bob from recovering the server's private source.

Acknowledgments

Uzi Pereg, Christian Deppe, and Holger Boche were supported by the the DFG n. EXC-2111 390814868 (Pereg, Boche), BO 1734/20-1 (Boche), and the BMBF n. 16KISQ028 (Pereg, Deppe), 16KIS0858 (Boche). This work of U. Pereg was also supported in part by the Israel CHE Fellowship for Quantum Science and Technology.

References

[1] J. Lopez, R. Rios, F. Bao, and G. Wang, "Evolving privacy: From sensors to the internet of things," *Future Gen. Comp. Syst.*, vol. 75, pp. 46–57, 2017.

[2] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J*, vol. 54(8), pp. 1355–1387, 1975.

[3] N. Merhav and S. Shamai, "Information rates subject to state masking," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2254–2261, June 2007.

[4] M. Naor, "Bit commitment using pseudorandomness," *J. Cryptology*, vol. 4, no. 2, pp. 151–158, 1991.

[5] C. S. Jensen, H. Lu, and M. L. Yiu, "Location privacy techniques in client-server architectures," in *Privacy in location-based applications*. Springer, 2009, pp. 31–58.

[6] O. O. Koyluoglu, R. Soundararajan, and S. Vishwanath, "State amplification subject to masking constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6233–6250, Nov 2016.

[7] M. Dikshtein, A. Somekh-Baruch, and S. Shamai, "Broadcasting information subject to state masking over a mimo state dependent gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2019)*, July 2019, pp. 275–279.

[8] M. Le Treust and M. R. Bloch, "State leakage and coordination with causal state knowledge at the encoder," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 805–823, 2021.

[9] S. X. Ng, A. Conti, G. L. Long, P. Muller, A. Sayeed, J. Yuan, and L. Hanzo, "Guest editorial advances in quantum communications, computing, cryptography, and sensing," *IEEE J. Selected Areas in Commun.*, vol. 38, no. 3, pp. 405–412, 2020.

[10] A. S. Holevo, "The capacity of the quantum channel with general signal states," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 269–273, Jan 1998.

[11] B. Schumacher and M. D. Westmoreland, "Sending classical information via noisy quantum channels," *Phys. Rev. A*, vol. 56, no. 1, p. 131, July 1997.

[12] R. Bassoli, H. Boche, C. Deppe, R. Ferrara, F. H. P. Fitzek, G. Janßen, and S. Saeedinaeeni, "Quantum communication networks," in *Ser. Found. Signal Proc. Commun. Netw.* Springer, 2021, vol. 23.

[13] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2637–2655, Oct 2002.

[14] H. Boche, N. Cai, and J. Nötzel, "The classical-quantum channel with random state parameters known to the sender," *J. Physics A: Math. and Theor.*, vol. 49, no. 19, p. 195302, April 2016.

[15] U. Pereg, "Communication over quantum channels with parameter estimation," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, pp. 359–383, 2022.

[16] F. Dupuis, "The capacity of quantum channels with side information at the transmitter," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2009)*, June 2009, pp. 948–952.

[17] U. Pereg, "Entanglement-assisted capacity of quantum channels with side information," in *Int. Zürich Seminar Inf. Commun. (IZS'2020)*, Zürich, Switzerland, Feb 2020, pp. 106–110.

[18] N. Cai, A. Winter, and R. W. Yeung, "Quantum privacy and quantum wiretap channels," *Probl. Info. Transm.*, vol. 40, no. 4, pp. 318–336, 2004.

[19] U. Pereg, C. Deppe, and H. Boche, "Quantum channel state masking," *IEEE Trans. Inf. Theory*, vol. 67, no. 4, pp. 2245–2268, 2021.

[20] ——, "Classical state masking over a quantum channel," arXiv:2109.12647 [quant-ph], 2021. [Online]. Available: https://arxiv.org/pdf/2109.12647.pdf

[21] M. M. Wilde, *Quantum information theory*, 2nd ed. Cambridge University Press, 2017.

[22] I. Devetak and P. W. Shor, "The capacity of a quantum channel for simultaneous transmission of classical and quantum information," *Commun. in Math. Phys.*, vol. 256, no. 2, pp. 287–303, June 2005.

[23] U. Pereg, "Communication over quantum channels with parameter estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2020)*, June 2020, pp. 1818–1823.

[24] A. El Gamal and Y. Kim, *Network Information Theory*. Cambridge University Press, 2011.

The Secrecy Gain of Formally Unimodular Lattices on the Gaussian Wiretap Channel

Maiara F. Bollauf, Hsuan-Yin Lin, and Øyvind Ytrehus
Simula UiB, N–5008 Bergen, Norway
Emails: {maiara, lin, oyvindy}@simula.no

***Abstract*—We consider lattice coding for the Gaussian wiretap channel, where the challenge is to ensure reliable communication between two authorized parties while preventing an eavesdropper from learning the transmitted messages. Recently, a measure called the *secrecy function* of a lattice coding scheme was proposed as a design criterion to characterize the eavesdropper's probability of correct decision. In this paper, the family of *formally unimodular lattices* is presented and shown to possess the same secrecy function behavior as unimodular and isodual lattices. Based on Construction A, we provide a universal approach to determine the *secrecy gain*, i.e., the maximum value of the secrecy function, for formally unimodular lattices obtained from formally self-dual codes. Furthermore, we show that formally unimodular lattices can achieve higher secrecy gain than the best-known unimodular lattices from the literature.**

I. Introduction

In recent years, *physical layer security* based on information theory has attracted a great deal of attention for secure applications in wireless communications in 5G and beyond (see [1] and references therein). This line of research has evolved from the classical *wiretap channel (WTC) model* introduced by Aaron Wyner in his landmark work [2], which showed that reliable and secure communication can be achieved simultaneously without the need of an additional cryptographic layer on top of the communication protocol.

Since then, substantial research efforts have been devoted to developing practical codes for reliable and secure data transmission over WTCs. Among the potential candidates are *lattices*, where in [3], [4] it was shown that a lattice-based coset encoding approach can provide secure and reliable communication on the Gaussian WTC. In particular, it was shown that for Gaussian WTC, the so-called *secrecy function* expressed in terms of the *theta series* of a lattice (see the precise definition in Section III) can be considered as a quality criterion of good wiretap lattice codes: to minimize the eavesdropper's probability of correct decision, one needs to maximize the secrecy function, and the corresponding maximum value is referred to as *(strong) secrecy gain*.

Belfiore and Solé [5] studied *unimodular* lattices and showed that their secrecy functions have a symmetry point. The value of the secrecy function at this point is called the *weak secrecy gain*. Based on this, the authors of [5] conjectured that for unimodular lattices, the secrecy gain is achieved at the symmetry point of its secrecy function. I.e., the secrecy gain of a unimodular lattice is equivalent to its weak secrecy gain. Finding good unimodular lattices that attain large secrecy gain is of practical importance. In [6], a novel technique was proposed to verify or disprove the Belfiore and Solé conjecture for a given unimodular lattice. Using this method, the conjecture is validated for all known even extremal unimodular lattices in dimensions less than 80. In another work [7], the authors use a similar method as [6] to classify the best unimodular lattices in dimensions from dimensions 8 to 23. For unimodular lattices obtained by Construction A from binary doubly even self-dual codes up to dimensions 40, their secrecy gains are also shown to be achieved at their symmetry points [8].

This work first introduces a new and wider family of lattices, referred to as *formally unimodular lattices*, that consists of lattices having the same theta series as their dual. We then prove that formally unimodular lattices have the same symmetry point as unimodular or isodual lattices. Similar to the feature of formally self-dual codes defined in coding theory, it is expected that such a broader class of lattices can achieve higher secrecy gain than the unimodular lattices. We pursue this expectation via Construction A lattices obtained from formally self-dual codes and give a universal approach to determine their secrecy gain. For formally unimodular lattices obtained by Construction A from even formally self-dual codes, we also provide a sufficient condition to verify Belfiore and Solé's conjecture on the secrecy gain. (A code is called *even* if all of its codewords have even weight, otherwise the code is *odd*.)

Furthermore, we present numerical evidence supporting the conjecture of secrecy gain also for Construction A lattices obtained from *odd* formally self-dual codes. For dimensions up to 70, we note that formally unimodular lattices have better secrecy gain than the best known unimodular lattices described in the literature, e.g., [7]. Apart from finding good formally self-dual codes from the literature, using the code construction by tailbiting the rate $1/2$ convolution codes [9, App. C], we also obtain several formally self-dual codes resulting in high secrecy gains. Due to page limitations, some proofs and detailed discussions are omitted and can be found in the extended version [9].

II. Definitions and Preliminaries

A. Notation

We denote by $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ the set of integers, rationals, and reals, respectively. Vectors are boldfaced, e.g., $\boldsymbol{x}$. Matrices and sets are represented by capital sans serif letters and calligraphic uppercase letters, respectively, e.g., X and $\mathcal{X}$. We use the customary code parameters $[n, k]$ or $[n, k, d]$ to denote a linear code $\mathscr{C}$ of length n, dimension k, and minimum Hamming distance d. Throughout this paper, we will focus on binary codes only.

B. On Codes and Lattices

Let $\mathscr{C}$ be an $[n, k]$ code and $\mathscr{C}^{\perp} \triangleq \{\boldsymbol{u}\colon \langle \boldsymbol{u}, \boldsymbol{v}\rangle = 0, \forall\, \boldsymbol{v} \in \mathscr{C}\}$. The *weight enumerator* of a code $\mathscr{C}$ is given by

$$W_{\mathscr{C}}(x, y) = \sum_{w=0}^{n} A_w x^{n-w} y^w,$$

where $A_w \triangleq |\{\boldsymbol{c} \in \mathscr{C}\colon w_{\mathrm{H}}(\boldsymbol{c}) = w\}|$. The relation between $W_{\mathscr{C}}(x, y)$ and $W_{\mathscr{C}^{\perp}}(x, y)$ is characterized by the well-known *MacWilliams identity* (see, e.g., [10, Th. 1, Ch. 5]):

$$W_{\mathscr{C}}(x, y) = \frac{1}{2^{n-k}} W_{\mathscr{C}^{\perp}}(x + y, x - y). \tag{1}$$

We have the following families of codes.

Definition 1 (Self-dual, isodual, formally self-dual codes):

- A code $\mathscr{C}$ is said to be *self-dual* if $\mathscr{C} = \mathscr{C}^{\perp}$.
- If there is a permutation π of coordinates such that $\mathscr{C} = \pi(\mathscr{C}^{\perp})$, $\mathscr{C}$ is called *isodual.*
- A code $\mathscr{C}$ is *formally self-dual* if $\mathscr{C}$ and $\mathscr{C}^{\perp}$ have the same weight enumerator, i.e., $W_{\mathscr{C}}(x, y) = W_{\mathscr{C}^{\perp}}(x, y)$.

Clearly, a self-dual code is also isodual, and an isodual code is formally self-dual. Any code in these classes is an $[n, {}^{n}\!/_{2}]$ code and, by (1), its weight enumerator $W_{\mathscr{C}}(x, y)$ satisfies [10, eq. (7), p. 599]

$$W_{\mathscr{C}}(x, y) = W_{\mathscr{C}}\Bigl(\frac{x + y}{\sqrt{2}}, \frac{x - y}{\sqrt{2}}\Bigr). \tag{2}$$

A (full rank) *lattice* Λ is a discrete additive subgroup of $\mathbb{R}^n$, which is generated as $\Lambda = \{\boldsymbol{\lambda} = \boldsymbol{u}\mathsf{G}_{n\times n}\colon \boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{Z}^n\}$, where the n rows of G form a lattice basis. The *volume* of Λ is $\mathrm{vol}(\Lambda) = |\det(\mathsf{G})|$.

If a lattice Λ have generator matrix G, then the lattice $\Lambda^{\star} \subset \mathbb{R}^n$ generated by $\bigl(\mathsf{G}^{-1}\bigr)^{\mathsf{T}}$ is called the *dual lattice* of Λ.

Remark 1: $\mathrm{vol}(\Lambda^{\star}) = \mathrm{vol}(\Lambda)^{-1}$.

For lattices, the analogue of the weight enumerator of a code is the *theta series*.

Definition 2 (Theta series): Let $\Lambda \subset \mathbb{R}^n$ be a lattice, its *theta series* is given by

$$\Theta_{\Lambda}(z) = \sum_{\boldsymbol{\lambda}\in\Lambda} q^{\|\boldsymbol{\lambda}\|^2},$$

where $q \triangleq e^{i\pi z}$ and $\mathrm{Im}\{z\} > 0$.

Analogously, the spirit of the MacWilliams identity can be captured by the *Jacobi's formula* [11, eq. (19), Ch. 4]

$$\Theta_{\Lambda}(z) = \mathrm{vol}(\Lambda^{\star})\Bigl(\frac{i}{z}\Bigr)^{\frac{n}{2}} \Theta_{\Lambda^{\star}}\Bigl(-\frac{1}{z}\Bigr). \tag{3}$$

Note that sometimes the theta series of a lattice can be expressed in terms of the *Jacobi theta functions* defined as follows.

$$\vartheta_2(z) \triangleq \sum_{m\in\mathbb{Z}} q^{\left(m+\frac{1}{2}\right)^2} = \Theta_{\mathbb{Z}+\frac{1}{2}}(z),$$

$$\vartheta_3(z) \triangleq \sum_{m\in\mathbb{Z}} q^{m^2} = \Theta_{\mathbb{Z}}(z), \quad \vartheta_4(z) \triangleq \sum_{m\in\mathbb{Z}} (-q)^{m^2}.$$

In lattice theory, we have similar concepts to self-dual and isodual dual codes. Here, we also introduce *formally unimodular* lattices.

Definition 3 (Unimodular, isodual, formally unimodular lattices): A lattice $\Lambda \subset \mathbb{R}^n$ is said to be *integral* if the inner product of any two lattice vectors is an integer.

- An integral lattice such that $\Lambda = \Lambda^{\star}$ is called *unimodular* lattice.
- A lattice Λ is called *isodual* if it can be obtained from its dual $\Lambda^{\star}$ by (possibly) a rotation or reflection.
- A lattice Λ is *formally unimodular* if it has the same theta series as its dual, i.e., $\Theta_{\Lambda}(z) = \Theta_{\Lambda^{\star}}(z)$.

Remark 2: The relations among unimodular, isodual, and formally unimodular lattices are given as follows.

$$\bigl\{\Lambda_{\text{unimodular}}\bigr\} \subset \bigl\{\Lambda_{\text{isodual}}\bigr\} \subset \bigl\{\Lambda_{\text{formally unimodular}}\bigr\}.$$

Proposition 1: If Λ is formally unimodular, then $\mathrm{vol}(\Lambda) = 1$.

Consequently, unimodular, isodual, and formally unimodular lattices satisfy

$$\Theta_{\Lambda}(z) = \Bigl(\frac{i}{z}\Bigr)^{\frac{n}{2}} \Theta_{\Lambda}\Bigl(-\frac{1}{z}\Bigr). \tag{4}$$

Lattices can be constructed from linear codes through the so called *Construction A*.

Definition 4 (Construction A): Let $\mathscr{C}$ be an $[n, k]$ code, then

$$\Lambda_{\mathrm{A}}(\mathscr{C}) \triangleq \tfrac{1}{\sqrt{2}}(\phi(\mathscr{C}) + 2\mathbb{Z}^n),$$

is a lattice, where $\phi : \mathbb{F}_2^n \to \mathbb{R}^n$ is the natural embedding.

About Construction A lattices obtained from codes over $\mathbb{F}_2$, it is known from [11, p. 183] that

- The volume is $\mathrm{vol}(\Lambda_{\mathrm{A}}(\mathscr{C})) = \frac{2^{n/2}}{|\mathscr{C}|} = 2^{(n-2k)/2}$.
- $\Lambda_{\mathrm{A}}(\mathscr{C}^{\perp}) = \Lambda_{\mathrm{A}}(\mathscr{C})^{\star}$.

A connection between the weight enumerator $W_{\mathscr{C}}(x, y)$ of a code $\mathscr{C}$ and a lattice $\Lambda_{\mathrm{A}}(\mathscr{C})$ can be established.

Lemma 1 ([11, Th. 3, Ch. 7]): Consider an $[n, k]$ code $\mathscr{C}$ with $W_{\mathscr{C}}(x, y)$, then the theta series of $\Lambda_{\mathrm{A}}(\mathscr{C})$ is given by

$$\Theta_{\Lambda_{\mathrm{A}}(\mathscr{C})}(z) = W_{\mathscr{C}}(\vartheta_3(2z), \vartheta_2(2z)).$$

Remark 3: It follows immediately from Lemma 1 that if an $[n, {}^{n}\!/_{2}]$ code $\mathscr{C}$ is formally self-dual then $\Lambda_{\mathrm{A}}(\mathscr{C})$ is a formally unimodular lattice.

III. Secrecy Function of a Lattice

In the Gaussian WTC, the same coset encoding idea proposed in Wyner's seminal paper [2] for linear codes can be implemented in a lattice scenario, and here we follow the lattice coding scheme proposed in [4], [5].

In practice, two lattices $\Lambda_{\mathrm{e}} \subset \Lambda_{\mathrm{b}}$ are considered. Λ_{b} is designed to ensure reliability for a legitimate receiver Bob and required to have a good *Hermite parameter* (that measures the highest attainable coding gain of an n-dimensional lattice) [11]. On the other hand, Λ_{e} is aimed to increase the eavesdropper confusion, so it should be chosen such that $P_{c,\mathrm{e}}$, the eavesdropper's success probability of correctly guessing the transmitted message, is minimized. The performance of the lattice Λ_e is measured in terms of the secrecy gain [4], [5]; to be explained next.

Denote by σ_{e}^2 the variance of the additive Gaussian noise at the eavesdropper's side. Minimizing $P_{c,\mathrm{e}}$ is equivalent to [4] minimizing

$$\sum_{\boldsymbol{r}\in\Lambda_{\mathrm{e}}} e^{-\|\boldsymbol{r}\|^2/2\sigma_{\mathrm{e}}^2} = \Theta_{\Lambda_{\mathrm{e}}}\biggl(z \triangleq \frac{i}{2\pi\sigma_{\mathrm{e}}^2}\biggr),$$

subject to $\log_2|\Lambda_{\mathrm{b}}/\Lambda_{\mathrm{e}}| = k$. Note that $\operatorname{Im}\{i/2\pi\sigma_{\mathrm{e}}^2\} = \operatorname{Im}\{z\} > 0$, thus we consider only the positive values of $\tau \triangleq -iz = 1/2\pi\sigma_{\mathrm{e}}^2 > 0$ for $\Theta_{\Lambda_{\mathrm{e}}}(z)$. Hence, the scheme is aimed at finding a lattice Λ_{e} such that $\Theta_{\Lambda_{\mathrm{e}}}(z)$ is minimized, which motivates the definition of *secrecy function* below. Note that in [12], it is also argued that minimizing the theta series of Λ_e leads to a small *flatness factor*, a criterion that directly relates to the mutual information leakage to the eavesdropper, instead of the success probability. Therefore, the optimization of $\Theta_{\Lambda_{\mathrm{e}}}(z)$ is of interest in both scenarios.

Definition 5 (Secrecy function and secrecy gain [4, Def. 1 and 2]): Let Λ be a lattice with volume $\operatorname{vol}(\Lambda) = \nu^n$. The secrecy function of Λ is defined by

$$\Xi_\Lambda(\tau) \triangleq \frac{\Theta_{\nu\mathbb{Z}^n}(i\tau)}{\Theta_\Lambda(i\tau)},$$

for $\tau \triangleq -iz > 0$. As maximizing $\Xi_\Lambda(\tau)$ is equivalent to minimizing $\Theta_\Lambda(z)$, the *(strong) secrecy gain* of a lattice is given by $\xi_\Lambda \triangleq \sup_{\tau>0} \Xi_\Lambda(\tau)$.

Ideally, the goal is to determine ξ_Λ. However, since the global maximum of a secrecy function is in general not always easy to calculate, a weaker definition is useful. We start by defining the *symmetry point*.

Definition 6 (Symmetry point): A point $\tau_0 \in \mathbb{R}$ is said to be a *symmetry point* if for all $\tau > 0$,

$$\Xi(\tau_0 \cdot \tau) = \Xi\Bigl(\frac{\tau_0}{\tau}\Bigr). \qquad (5)$$

Definition 7 (Weak secrecy gain [4, Def. 3]): If the secrecy function of a lattice Λ has a symmetry point τ_0, then the weak secrecy gain χ_Λ is defined as $\chi_\Lambda = \Xi_\Lambda(\tau_0)$.

IV. Weak Secrecy Gain of Formally Unimodular Lattices

This section shows that formally unimodular lattices also hold the same secrecy function properties as unimodular and isodual lattices [4].

Lemma 2: Consider a lattice Λ and its dual $\Lambda^\star$. Then,

$$\Xi_\Lambda(\tau) = \Xi_{\Lambda^\star}\Bigl(\frac{1}{\tau}\Bigr). \qquad (6)$$

A necessary and sufficient condition for a lattice Λ to achieve the weak secrecy gain at $\tau = 1$ is given as follows.

Theorem 1: Consider a lattice Λ with $\operatorname{vol}(\Lambda) = 1$ and its dual $\Lambda^\star$. Then, Λ achieves the weak secrecy gain at $\tau = 1$, if and only if Λ is formally unimodular.

Proof: By definition, we have

$$\Xi_\Lambda(\tau) = \Xi_\Lambda\Bigl(\frac{1}{\tau}\Bigr). \qquad (7)$$

Using Lemma 2, it follows from (7) and (6) that

$$\Xi_\Lambda\Bigl(\frac{1}{\tau}\Bigr) = \Xi_\Lambda(\tau) = \Xi_{\Lambda^\star}\Bigl(\frac{1}{\tau}\Bigr).$$

By Def. 5, this implies that $\Theta_\Lambda(z) = \Theta_{\Lambda^\star}(z)$ for $\operatorname{vol}(\Lambda) = 1$. Conversely, from Def. 3, we see that (6) implies (7). ■

Note that Theorem 1 holds for isodual lattices as well, which yields to [4, Prop. 1].

Corollary 1: Consider a lattice Λ with $\operatorname{vol}(\Lambda) = \nu^n$ and its dual $\Lambda^\star$. Then, Λ achieves the weak secrecy gain at $\tau = \nu^{-2}$, if and only if $\nu^{-1}\Lambda$ is a formally unimodular lattice.

Equation (5) with $\tau_0 = \nu^{-2}$ holds for a lattice equivalent to its dual. See [4, Prop. 2].

V. Secrecy Gain of Formally Unimodular Lattices

Our goal in this section is to investigate the following conjecture.

Conjecture 1: The secrecy function of a formally unimodular lattice Λ achieves its maximum at $\tau = 1$, i.e., $\xi_\Lambda = \Xi_\Lambda(1)$.

Although we cannot completely prove Conjecture 1, we proceed to study the secrecy gain for formally unimodular lattices obtained from formally self-dual codes via Construction A (see Remark 3). Note that for linear codes, it is known that formally self-dual codes that are not self-dual can outperform self-dual codes in some cases, as they comprise a wider class and hence may allow a better minimum Hamming distance or an overall more favorable weight enumerator. This leads us to look for improved results on the secrecy gain compared to unimodular lattices [6]–[8].

Lemma 3: Consider a Construction A lattice $\Lambda_{\mathrm{A}}(\mathscr{C})$ obtained from a formally self-dual code $\mathscr{C}$. Then, its theta series is equal to

$$\Theta_{\Lambda_{\mathrm{A}}(\mathscr{C})} = \frac{W_{\mathscr{C}}\Bigl(\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}, \sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}\Bigr)}{2^{\frac{n}{2}}}.$$

Proof: Using Lemma 1 and the useful identities given in [11, eq. (26), Ch. 4], the theta series $\Theta_{\Lambda_{\mathrm{A}}(\mathscr{C})}$ becomes

$$\begin{aligned}
&\Theta_{\Lambda_{\mathrm{A}}(\mathscr{C})}(z)\\
&\quad= W_{\mathscr{C}}(\vartheta_3(2z), \vartheta_2(2z))\\
&\quad\overset{(a)}{=} W_{\mathscr{C}}\biggl(\frac{\vartheta_3(2z)+\vartheta_2(2z)}{\sqrt{2}}, \frac{\vartheta_3(2z)-\vartheta_2(2z)}{\sqrt{2}}\biggr)\\
&\quad= W_{\mathscr{C}}\Biggl(\frac{\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}+\sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}}{\sqrt{2}\sqrt{2}},\\
&\qquad\qquad \frac{\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}-\sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}}{\sqrt{2}\sqrt{2}}\Biggr)\\
&\quad= \frac{1}{2^{\frac{n}{2}}}W_{\mathscr{C}}\Biggl(\frac{\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}+\sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}}{\sqrt{2}},\\
&\qquad\qquad \frac{\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}-\sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}}{\sqrt{2}}\Biggr)\\
&\quad\overset{(b)}{=} \frac{1}{2^{\frac{n}{2}}}W_{\mathscr{C}}\Bigl(\sqrt{\vartheta_3^2(z)+\vartheta_4^2(z)}, \sqrt{\vartheta_3^2(z)-\vartheta_4^2(z)}\Bigr).
\end{aligned}$$

where (a) and (b) follow from (2). ■

Lemma 4: Let $s(\tau) \triangleq \vartheta_4(i\tau)/\vartheta_3(i\tau)$. Then, $s(\tau)$ is an increasing function for $\tau > 0$, and $0 < s(\tau) < 1$.

Remark 4: Let $t(\tau) \triangleq s(\tau)^2$. Then, $0 < t(\tau) < 1$ and $t(\tau)$ is also an increasing function for $\tau > 0$. Hence, according to Lemma 4, given any $t \in (0,1)$, there always exists a unique $\tau > 0$ such that $t(\tau) = \vartheta_4^2(i\tau)/\vartheta_3^2(i\tau)$. Moreover, we have $t(1) = 1/\sqrt{2}$ by using the identity of $\vartheta_3(i) = 2^{1/4}\vartheta_4(i)$ from [13].

Due to Remark 4, Lemma 3, and the fact that $\Theta_{\mathbb{Z}^n}(z) = \vartheta_3^n(z)$, now we are able to give a new universal approach to derive the strong secrecy gain of a Construction A lattice obtained from formally self-dual codes.

Theorem 2: Let $\mathscr{C}$ be a formally self-dual code. Then

$$\left[\Xi_{\Lambda_{\mathrm{A}}(\mathscr{C})}(\tau)\right]^{-1} = \frac{W_{\mathscr{C}}\big(\sqrt{1+t(\tau)}, \sqrt{1-t(\tau)}\big)}{2^{\frac{n}{2}}},$$

where $0 < t(\tau) = \vartheta_4^2(i\tau)/\vartheta_3^2(i\tau) < 1$. Moreover, define $f_{\mathscr{C}}(t) \triangleq W_{\mathscr{C}}(\sqrt{1+t}, \sqrt{1-t})$ for $0 < t < 1$. Then, maximizing the secrecy function $\Xi_{\Lambda_{\mathrm{A}}(\mathscr{C})}(\tau)$ is equivalent to determining the minimum of $f_{\mathscr{C}}(t)$ on $t \in (0,1)$.

Example 1: Consider a $[6,3,3]$ odd formally self-dual code $\mathscr{C}$ with $W_{\mathscr{C}}(x,y) = x^6 + 4x^3y^3 + 3x^2y^4$ [14]. Thus $f_{\mathscr{C}}(t) = W_{\mathscr{C}}(\sqrt{1+t}, \sqrt{1-t}) = 4[1 + t^3 + (1-t^2)^{3/2}]$ and $f'_{\mathscr{C}}(t) = 12t(t - \sqrt{1-t^2})$. Observe that for $0 < t < 1/\sqrt{2}$, we have $\sqrt{1-t^2} > 1/\sqrt{2}$. Then, $t - \sqrt{1-t^2} < 1/\sqrt{2} - 1/\sqrt{2} = 0$. This indicates that the derivative $f'_{\mathscr{C}}(t) < 0$ on $t \in (0, 1/\sqrt{2})$. Similarly, one can also show that $f'_{\mathscr{C}}(t) > 0$ on $t \in (1/\sqrt{2}, 1)$, and $t = 1/\sqrt{2}$ is the minimum of $f_{\mathscr{C}}(t)$. Hence, Remark 4 and Theorem 2 indicate that the maximum of $\Xi_{\Lambda_{\mathrm{A}}(\mathscr{C})}(\tau)$ is achieved at $\tau = 1$. Also, one can get $\xi_{\Lambda_{\mathrm{A}}(\mathscr{C})} \approx 1.172$. ♢

Using Gleason's Theorem [15, Th. 9.2.1], an expression of $f_{\mathscr{C}}(t)$ can be shown if $\mathscr{C}$ is an even formally self-dual code.

Lemma 5: If $\mathscr{C}$ is an $[n, n/2]$ even formally self-dual codes, then we have

$$f_{\mathscr{C}}(t) = 2^{\frac{n}{2}} \sum_{r=0}^{\lfloor \frac{n}{8} \rfloor} a_r (t^4 - t^2 + 1)^r, \tag{8}$$

where $a_r \in \mathbb{Q}$ and $\sum_{r=0}^{\lfloor \frac{n}{8} \rfloor} a_r = 1$.

Next, we provide a sufficient condition for a Construction A formally unimodular lattice obtained from even formally self-dual codes to achieve the strong secrecy gain at $\tau = 1$, or, equivalently, $t = 1/\sqrt{2}$.

Theorem 3: Consider $n \geq 8$ and an $[n, n/2]$ even formally self-dual code $\mathscr{C}$. If the coefficients a_r of $f_{\mathscr{C}}(t)$ expressed in terms of (8) satisfy

$$\sum_{r=1}^{\lfloor \frac{n}{8} \rfloor} r a_r \left(\frac{3}{4}\right)^{r-1} > 0, \tag{9}$$

then the secrecy gain of $\Lambda_{\mathrm{A}}(\mathscr{C})$ is achieved at $\tau = 1$.

To prove this theorem, it is sufficient to show that the function $f_{\mathscr{C}}(t)$ as in (8) defined for $0 < t < 1$ achieves its minimum at $t = 1/\sqrt{2}$. The detailed proof is given in [9].

Example 2: Consider an $[18, 9, 6]$ even formally self-dual code $\mathscr{C}$ with

$$\begin{aligned} W_{\mathscr{C}}(x,y) = x^{18} &+ 102x^{12}y^6 + 153x^{10}y^8 \\ &+ 153x^8y^{10} + 102x^6y^{12} + y^{18}. \end{aligned}$$

By solving $f_{\mathscr{C}}(t) = W_{\mathscr{C}}(\sqrt{1+t}, \sqrt{1-t})$ with (8) (see the details of derivation provided in [9, App. B]), we find that $a_0 = -29/16, a_1 = 27/8$ and $a_2 = -9/16$. The condition (9) in Theorem 3 for those coefficients is satisfied since $27/8 - 27/32 = 81/32 > 0$. Thus, the secrecy gain conjecture is true for the formally unimodular lattice $\Lambda_{\mathrm{A}}(\mathscr{C})$. ♢

VI. Numerical Results

Even though the result of Theorem 3 is restricted to formally unimodular lattices obtained from even formally self-dual codes, we have numerical evidence showing that Conjecture 1 also holds for formally unimodular lattices obtained from odd formally self-dual codes. The secrecy gains of some formally unimodular Construction A lattices obtained from (even and odd) formally self-dual codes are summarized in Table I. Note that all codes have the parameters $[n, n/2]$ and the superscript "(d)" refers to the minimum Hamming distance d of the code. Their exact weight enumerators can be found in [9, App. D]. The highlighted values represent the best values found in the respective dimensions, when comparing self-dual (sd), even and odd formally self-dual (efsd and ofsd) codes.

Remark 5: We remark the following about Table I:

- "[·]" indicates the reference number.
- We use the sufficient condition (9) in Theorem 3 for the even codes and the numerical derivative analysis with Wolfram Mathematica [25] for the odd codes to confirm the strong secrecy gain in Table I.
- For most dimensions $n > 8$, the secrecy gain of formally unimodular lattices that are not unimodular outperform the unimodular lattices (obtained from self-dual codes), presented in [7, Tables I and II]. In some cases (*e.g.* [12,6], [22,11]) we were unable to find good efsd codes with different secrecy gains form the sd codes. Also, to highlight the comparison with unimodular lattices, the second column refers to the upper bound on the secrecy gain of unimodular lattices obtained from Construction A in [16, Tab. III] and not all of the values are known to be achieved. Gains can be observed in dimensions $10, 12, 14, 20$, and 22.
- It is known that the well-known Barnes-Wall lattice BW_{32} achieves the secrecy gain of $64/9 \approx 7.11$ [4, Sec. IV-C], which is better than all the tabulated values in dimension 32. However, because BW_{32} is not obtained via Construction A, we did not address the details here.
- Observe that for codes of length 40, the self-dual code in the table is a Type I (weights divisible by two), as it presents a higher secrecy gain ($\xi_{\Lambda_{\mathrm{A}}(\mathscr{C}_{\mathrm{sd}})} \approx 12.191$) compared to the Type II (weights divisible by four) ($\xi_{\Lambda_{\mathrm{A}}(\mathscr{C}_{\mathrm{sd}})} \approx 11.977$). The same happens with codes of length 32 and this confirms the advantage of this approach as to the results in [8].
- Formally self-dual (isodual) codes without references in Table I are constructed by tailbiting the rate $1/2$ convolutional codes. Details can be found in [9, App. C].

TABLE I
COMPARISON OF (STRONG) SECRECY GAINS FOR SEVERAL VALUES OF EVEN DIMENSIONS n. CODES WITHOUT REFERENCES ARE OBTAINED BY TAILBITING THE RATE $1/2$ CONVOLUTIONAL CODES.

n	Upper bound [16]	$\mathscr{C}_{\text{sd}}^{(d)}$	$\xi_{\Lambda_{\text{A}}(\mathscr{C}_{\text{sd}})}$	$\mathscr{C}_{\text{efsd}}^{(d)}$	$\xi_{\Lambda_{\text{A}}(\mathscr{C}_{\text{efsd}})}$	$\mathscr{C}_{\text{ofsd}}^{(d)}$	$\xi_{\Lambda_{\text{A}}(\mathscr{C}_{\text{ofsd}})}$
6	1	–	–	$\mathscr{C}_{\text{efsd}}^{(2)}$ [15]	1	$\mathscr{C}_{\text{ofsd}}^{(3)}$ [14]	$\mathbf{1.172}$
8	1.33	$\mathscr{C}_{\text{sd}}^{(4)}$ [15]	$\mathbf{1.333}$	–	–	$\mathscr{C}_{\text{ofsd}}^{(3)}$ [14]	1.282
10	1.45	–	–	$\mathscr{C}_{\text{efsd}}^{(4)}$ [17]	1.455	$\mathscr{C}_{\text{ofsd}}^{4}$ [14]	$\mathbf{1.478}$
12	1.6	$\mathscr{C}_{\text{sd}}^{(4)}$ [7]	1.6	$\mathscr{C}_{\text{efsd}}^{(4)}$ [18]	1.6	$\mathscr{C}_{\text{ofsd}}^{(4)}$ [14]	$\mathbf{1.657}$
14	1.78	$\mathscr{C}_{\text{sd}}^{(4)}$ [7]	1.778	$\mathscr{C}_{\text{efsd}}^{(4)}$ [18]	1.825	$\mathscr{C}_{\text{ofsd}}^{(4)}$ [14]	$\mathbf{1.875}$
16	2.21	$\mathscr{C}_{\text{sd}}^{(4)}$ [7]	2	$\mathscr{C}_{\text{efsd}}^{(4)}$ [19]	2.133	$\mathscr{C}_{\text{ofsd}}^{(5)}$ [14]	$\mathbf{2.141}$
18	2.49	$\mathscr{C}_{\text{sd}}^{(4)}$ [7]	2.286	$\mathscr{C}_{\text{efsd}}^{(6)}$ [20]	$\mathbf{2.485}$	$\mathscr{C}_{\text{ofsd}}^{(5)}$	2.427
20	2.81	$\mathscr{C}_{\text{sd}}^{(4)}$ [7]	2.667	$\mathscr{C}_{\text{efsd}}^{(6)}$ [21]	2.813	$\mathscr{C}_{\text{ofsd}}^{(6)}$ [18]	$\mathbf{2.868}$
22	3.2	$\mathscr{C}_{\text{sd}}^{(6)}$ [7]	3.2	$\mathscr{C}_{\text{efsd}}^{(6)}$	3.2	$\mathscr{C}_{\text{ofsd}}^{(7)}$ [14]	$\mathbf{3.335}$
30	5.84	$\mathscr{C}_{\text{sd}}^{(6)}$ [22]	5.697	$\mathscr{C}_{\text{efsd}}^{(8)}$ [23]	$\mathbf{5.843}$	$\mathscr{C}_{\text{ofsd}}^{(7)}$	5.785
32	7.00	$\mathscr{C}_{\text{sd}}^{(8)}$ [22]	6.737	$\mathscr{C}_{\text{efsd}}^{(8)}$	$\mathbf{6.748}$	$\mathscr{C}_{\text{ofsd}}^{(7)}$	6.628
40	12.81	$\mathscr{C}_{\text{sd}}^{(8)}$ [22]	12.191	$\mathscr{C}_{\text{efsd}}^{(8)}$	12.134	$\mathscr{C}_{\text{ofsd}}^{(9)}$	$\mathbf{12.364}$
70	130.15	$\mathscr{C}_{\text{sd}}^{(12)}$ [24]	127.712	$\mathscr{C}_{\text{efsd}}^{(12)}$	128.073	$\mathscr{C}_{\text{ofsd}}^{(13)}$	$\mathbf{128.368}$

VII. CONCLUSION

This paper introduced the *formally unimodular lattices*, a new class consisting of lattices having the same theta series as their dual. We showed some properties of formally unimodular lattices and their secrecy function behavior in the Gaussian WTC. Furthermore, we investigated Construction A lattices obtained from formally self-dual codes and gave a universal approach to determine their secrecy gain. We found formally unimodular lattices of better secrecy gain than the best known unimodular lattices from the literature.

REFERENCES

[1] Y. Wu, A. Khisti, C. Xiao, G. Caire, K.-K. Wong, and X. Gao, "A survey of physical layer security techniques for 5G wireless networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 679–695, Apr. 2018.

[2] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[3] J.-C. Belfiore and F. Oggier, "Secrecy gain: A wiretap lattice code design," in *Proc. IEEE Int. Symp. Inf. Theory Appl. (ISITA)*, Taichung, Taiwan, Oct. 17–20, 2010.

[4] F. Oggier, P. Solé, and J.-C. Belfiore, "Lattice codes for the wiretap Gaussian channel: Construction and analysis," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5690–5708, Oct. 2016.

[5] J.-C. Belfiore and P. Solé, "Unimodular lattices for the Gaussian wiretap channel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Dublin, Ireland, Aug. 30 – Sep. 3, 2010.

[6] A.-M. Ernvall-Hytonen, "On a conjecture by Belfiore and Solé on some lattices," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5950–5955, Sep. 2012.

[7] F. Lin and F. Oggier, "A classification of unimodular lattice wiretap codes in small dimensions," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3295–3303, Jun. 2013.

[8] J. Pinchak, "Wiretap codes: Families of lattices satisfying the Belfiore-Solé secrecy function conjecture," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 7–12, 2013, pp. 2617–2620.

[9] M. F. Bollauf, H.-Y. Lin, and Ø. Ytrehus, "The secrecy gain of formally unimodular lattices on the Gaussian wiretap channel," Oct. 2021, arXiv:2111.01439v1 [cs.IT]. [Online]. Available: https://arxiv.org/abs/2111.01439

[10] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam, The Netherlands: North-Holland, 1977.

[11] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. New York, NY, USA: Springer, 1999.

[12] C. Ling, L. Luzzi, J.-C. Belfiore, and D. Stehle, "Semantically secure lattice codes for the Gaussian wiretap channel," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6399–6416, Oct. 2014.

[13] E. W. Weisstein, "Jacobi theta functions," From MathWorld—A Wolfram Web Resource. [Online]. Available: https://mathworld.wolfram.com/JacobiThetaFunctions.html

[14] K. Betsumiya and M. Harada, "Binary optimal odd formally self-dual codes," *Des., Codes Cryptography*, vol. 23, no. 1, pp. 11–21, 2001.

[15] W. C. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*. Cambridge, U.K.: Cambridge University Press, jun 2003.

[16] F. Lin and F. Oggier, "Gaussian wiretap lattice codes from binary self-dual codes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 3–7, 2012.

[17] G. T. Kennedy and V. Pless, "On designs and formally self-dual codes," *Des., Codes Cryptography*, vol. 4, no. 1, pp. 43–55, 1994.

[18] K. Betsumiya, T. A. Gulliver, and M. Harada, "Binary optimal linear rate $1/2$ codes," in *Proc. Appl. Algebra, Algebr. Algorithms Error-Correcting Codes (AAECC)*, Honolulu, HI, USA, Nov. 15–19, 1999, pp. 462–471.

[19] K. Betsumiya and M. Harada, "Classification of formally self-dual even codes of lengths up to 16," *Des., Codes Cryptography*, vol. 23, no. 3, pp. 325–332, 2001.

[20] N. J. A. Sloane and N. Heninger, *The On-Line Encyclopedia of Integer Sequences*, OEIS Foundation Inc., Jun. 2006. [Online]. Available: http://oeis.org/A123456

[21] J. E. Fields, P. Gaborit, W. C. Huffman, and V. Pless, "On the classification of extremal even formally self-dual codes of lengths 20 and 22," *Discrete Appl. Math.*, vol. 111, no. 1-2, pp. 75–86, Jul. 2001.

[22] J. H. Conway and N. J. A. Sloane, "A new upper bound on the minimal distance of self-dual codes," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1319–1333, Nov. 1990.

[23] S. Bouyuklieva and I. Bouyukliev, "Classification of the extremal formally self-dual even codes of length 30," *Adv. Math. Commun.*, vol. 4, no. 3, pp. 433–439, 2010.

[24] M. Harada, "The existence of a self-dual $[70, 35, 12]$ code and formally self-dual codes," *Finite Fields Th. App.*, vol. 3, no. 2, pp. 131–139, Apr. 1997.

[25] Wolfram Research, Inc., "Mathematica, Version 12.3.1," champaign, IL, 2021. [Online]. Available: https://www.wolfram.com/mathematica

Reed-Muller Identification

Mattia Spandri, Roberto Ferrara, Christian Deppe
Institute for Communication Engineering, *Technical University of Munich*, Munich, Germany

***Abstract*—Ahlswede and Dueck identification has the potential of exponentially reducing traffic or exponentially increasing rates in applications where a full decoding of the message is not necessary and, instead, a simple verification of the message of interest suffices. However, the proposed constructions can suffer from exponential increase in the computational load at the sender and receiver, rendering these advantages unusable. This has been shown in particular to be the case for a construction achieving identification capacity based on concatenated Reed-Solomon codes. Here, we consider the natural generalization of identification based on Reed-Muller codes and we show that, although without achieving identification capacity, they allow to achieve the exponentially large rates mentioned above without the computational penalty increasing too much the latency with respect to transmission.**

***Index Terms*—identification, verifier, encoder, latency, complexity, Ahlswede, Dueck Reed-Solomon, Reed-Muller**

I. Introduction

Ahlswede and Dueck's identification is a different communication paradigm from Shannon's transmission that promises an exponential larger capacity, or equivalently an exponential reduction in channel uses, at the trade-off of only allowing an hypothesis test at the receiver instead of a full decoding [3]. Identification capacity on a noisy channel can be achieved by concatenating a capacity-achieving identification code for the noiseless channel with a capacity achieving transmission code for the noisy channel [3]. In other words, it is enough to correct the channel first and then apply some pre and post processing, as also done to achieve secrecy in a wiretap channel.

As common for capacity results, the achievability proof ignores the complexity of constructing the code, of the encoder and of the decoder. In particular, since identification promises an exponential increase in the rates, even simply reading the chosen identity (sometimes still called message to make the parallel with transmission) will incur some penalty. In previous works [5, 7], we analyzed the time spent encoding and the noiseless-channel error probability for capacity-achieving noiseless identification codes based on concatenated Reed-Solomon codes [14]. The result from those works was that, with todays transmission speeds, it is generally faster to simply send the unique string defining the identity than spend the time encoding for identification. In order to make noiseless identification competitive in terms of latency, the use of Zech tables was necessary to speed up the computation over finite fields, however this option was limited to codes of small size, leaving the open question of finding similarly fast identification-codes at larger sizes. The codes from [14] are only one of possible identification capacity-achieving constructions, which can generally be obtained via block codes satisfying the Gilbert-Varshamov bound [1, Section III.B]. Other such constructions are the algebraic codes of [10, 6] as pointed out in [14], a construction based on hash functions [11], and the recent construction of [8].

In this work, we naturally generalize to identification codes base on Reed-Muller codes in order to increase the size of the identities without increasing the size of the finite fields we work on. We find that, although the small field sizes also limits how low we can make the error probability, we can circumvent this using multiple encoding [7] and efficiently reduce the error without much impact on the other parameters.

The paper is structured as follows. In Sections II and III, we quickly review identification and Reed-Muller codes. In Section IV, we show that they still cannot achieve identification capacity without concatenation. In Section V, we discuss the implementation and show how it allows to achieve large exponential increase in rates without much latency and false-accept penalty compared to transmission. In the appendices, we describe in detail how we measured the time cost of operations in an attempt to predict the performance of the code.

II. Identification

We use the notation $[n] = \{0, ..., n-1\}$ for any natural number n and the notation $PW = \sum_x P(x)W(\cdot|x)$ for a probability distribution P and a channel W. An (n, I, ε) identification code for W^n is a tuple $\{E_i, V_i\}_{i\in[I]}$ of probability distributions and verifier sets (like stochastic codes for transmission) such that $e_{ij} = |E_i W^n(V_j) - \delta_{ij}| \leq \varepsilon$, where δ_{ij} is the Kronecker delta (notice that e_{ii} are the usual errors in transmission). No disjointness or limit on the intersection is imposed on the verifier sets. The rate is defined as $\frac{1}{n}\log\log I$ rather than $\frac{1}{n}\log I$, and the capacity is then the supremum of achievable rates as usual. As mentioned already, we can focus only on coding for the noiseless channel, like in [7], in which case it is enough to construct the appropriate verifier sets V_i and let E_i be the uniform distributions on these sets [3]. One way to do this, is to construct the verifier sets sets using a function $f_i : [R] \to [T]$ for each identity i, such that $[R] \times [T]$ can be mapped one-to-one to the inputs of the noiseless channel. These sets are then none other than the relation sets $V_i = \{(r, f_i(r))\}_{r\in[R]} \subset [R] \times [T]$ defined by f_i. We call r the *randomness* and $f_i(r)$ the *tag*. By construction, we can then think of the encoder as choosing a random challenge in the form of a randomness-tag pair $(r, f_i(r))$ and sending it through the channel, so that the receiver wanting to verify identity j

We acknowledge support from the German Federal Ministry of Education and Research (BMBF) to C. Deppe and R. Ferrara under Grant 16KIS1005. Contact information: {roberto.ferrara,christian.deppe}@tum.de

will recompute the tag $f_j(r)$ and conclude that $i = j$ if the recomputed tag is equal to the received tag $f_j(r) \overset{?}{=} f_i(r)$ [2, 13, 5, 7]. With such a scheme e_{ii} will always be 0, while e_{ij} is bounded by the fraction of collisions of f_i and f_j, the number of outputs that coincide (see Eq. (10)). To limit e_{ij}, the number of collisions need to be limited, which makes the set of such identification codes in one to one correspondence with error-correction block codes: each codeword (a string of symbols) defines a function from symbol positions to symbol values and the distance of the code gives a bound on the false-accept error probability [5]. For example, using Reed-Solomon codes, the functions corresponding to the codewords are the polynomials used to generate the codewords.

III. Reed-Muller Codes

For our purpose it will make sense to consider q-ary rather than just binary Reed-Muller codes [9, 4, 12]. Let $k, m \in \mathbb{N}$ and $q > k$ a prime power. Because of our application to identification, we define the $\mathrm{RM}_q(k, m)$ Reed-Muller code as the collection of multivariate polynomials with m variables and degree at most k over $\mathbb{F}_q$. For this, we introduce some notation first. We define $|z| := \sum_{j=1}^{m} z_j$ for any $z \in [k]^m$. We then define $r^z := \prod_{j=1}^{m} r_j^{z_j}$ for $r \in \mathbb{F}_q^m$. The Reed-Muller code is then defined as

$$\mathrm{RM}_q(k, m) = \left\{ \begin{matrix} p_w : \mathbb{F}_q^m \to \mathbb{F}_q : \\ p_w(r) = \sum_{z \in [k]^m : |z| \leq k} w_z r^z \end{matrix} \right\}, \tag{1}$$

where $w = \{w_z\}_{z \in [k]^m : |z| \leq k}$ are the $\binom{k+m}{m}$ coefficients in $\mathbb{F}_q$. In case of a Reed-Muller error-correction code, every polynomial constructs a codeword by concatenating polynomial evaluations at different input points. The maximum blocklength is the number of possible inputs, which results in a

$$\left[q^m, \binom{k+m}{m}, (q-k)q^{m-1} \right]_q . \tag{2}$$

block code. In identification, a functional encoding that can compute a single letter of the codewords without computing the whole codeword is preferred [5, 7]. For the Reed-Muller code, this is the polynomial encoding. The size of the Reed-Muller identification code is the number of distinct polynomials, in bits this is

$$\log I = \binom{k+m}{m} \log q. \tag{3}$$

However, only a transmission of

$$\log C = (m+1) \log q \tag{4}$$

bits is needed, since only the challenge, composed of

$$\log R = m \log q \qquad \text{and} \qquad \log T = \log q \tag{5}$$

bits of randomness and tag, is sent through the channel. Thus for a single Reed-Muller code, the increase from the transmission rate r_{T} to the identification rate r_{ID} is

$$\frac{r_{\mathrm{ID}}}{r_{\mathrm{T}}} = \frac{\log I}{\log C} = \frac{\binom{k+m}{m}}{m+1}. \tag{6}$$

compared to $\frac{r_{\mathrm{ID}}}{r_{\mathrm{T}}} = \frac{k}{2}$ of a Reed-Solomon code [13, 5]. If multiple n challenges are sent, this reduces the error but also reduces the rate increase to $\binom{k+m}{m}/n/(m+1)$. The errors e_{ij} are upper bounded by the fractional distance

$$E = 1 - \frac{(q-k)q^{m-1}}{R} = \frac{k}{q}. \tag{7}$$

This is independent of the number of variables m and less than one because $k < q$. The error decreases as $E^n = (\frac{k}{q})^n$ with the number of challenges, because all challenges need to be verified simultaneously.

IV. Capacity

In order to achieve identification capacity, the noiseless identification codes need to satisfy three simple conditions [14][1]:

1) Randomness: asymptotically all the transmission rate is used for randomness:

$$\frac{\log T}{\log R} \to 0 \qquad \Leftrightarrow \qquad \frac{\log R}{\log RT} \to 1 \tag{8}$$

where $C = RT$ is the size of the challenge;

2) Size: asymptotically the identification rate must equal the randomness/transmission rate:

$$\frac{\log \log I}{\log R} \to 1; \tag{9}$$

3) Error: asymptotically the error must go to zero

$$E = \max e_{ij} = 1 - \frac{1}{R} \max_{i \neq j} d(T_i, T_j) \to 0. \tag{10}$$

It will be more convenient to use the equivalent condition

$$\log E \to -\infty, \tag{11}$$

Since Reed-Muller codes contain Reed-Solomon codes as a special case, they are also able to achieve identification capacity using concatenation of multiple codes. The question is whether capacity can be achieved without concatenation, which is not possible with Reed-Solomon codes [14]. Below we prove that the conditions to achieve capacity cannot be satisfied simultaneously, even taking into consideration the use of multiple challenges, and the expansion of extension fields.

A. Without expansion

We first do the analysis without expansion for simplicity. We begin with the randomness, Eq. (8), which simply requires

$$\frac{\log R^n}{\log T^n} = \frac{\log T}{\log R} = \frac{1}{m} \to 0 \qquad \Rightarrow \qquad m \to \infty$$

which is independent of the number of challenges. From the error requirement, Eq. (11), we need to satisfy $n \log \frac{k}{q} = n \log k - n \log q \to -\infty$ which gives

$$n \log q \to \infty \qquad \text{and} \qquad n \log q \gg n \log k \tag{12}$$

[1] In [14], these conditions are called “optimal” for identification in the sense of achieving capacity, not in the sense of being optimal at finite blocklengths.

For the size, Eq. (9), we can use the upper bound on the binomial $\binom{a}{b} \leq \left(\frac{\mathrm{e}a}{b}\right)^b$ with e Euler's number. We thus bound

$$\frac{\log\log I}{n\log R} = \frac{\log\log q + \log\binom{k+m}{m}}{nm\log q} \lesssim \frac{m\log\frac{\mathrm{e}(k+m)}{m}}{nm\log q}$$

where we used that $\log\log q/n\log q \to 0$ by Eq. (12), then for the same reason we can further bound

$$\frac{\log\frac{\mathrm{e}(k+m)}{m}}{n\log q} \approx \frac{\log\left(1+\frac{k}{m}\right)}{n\log q} \lesssim \frac{\log(1+k)}{n\log q}$$

which goes to zero again by Eq. (12).

B. With expansion

A feature of the concatenated Reed-Solomon codes achieving identification capacity [14] is that the first code is defined over $\mathbb{F}_{q^\alpha}$ but then each symbol is considered as a string of α symbols in $\mathbb{F}_q$. This allows to satisfy the randomness and size requirement, while the second Reed-Solomon code takes care of the error requirement. Here, we prove that expanding symbols in $\mathbb{F}_{q^\alpha}$ is still not enough to satisfy the requirements with Reed-Muller codes. We change all parameters to powers of q, the expansion transforms the code as

$$\begin{aligned}&\left[(q^\alpha)^{q^\gamma}, \binom{q^\beta+q^\gamma}{q^\gamma}, (q^\alpha-q^\beta)(q^\alpha)^{q^\gamma-1}\right]_{q^\alpha}\\ &= \left[\alpha(q^\alpha)^{q^\gamma}, \alpha\binom{q^\beta+q^\gamma}{q^\gamma}, (q^\alpha-q^\beta)(q^\alpha)^{q^\gamma-1}\right]_q \qquad (13)\end{aligned}$$

which has error

$$E = 1 - \frac{(q^\alpha-q^\beta)(q^\alpha)^{q^\gamma-1}}{\alpha(q^\alpha)^{q^\gamma}} = \frac{\alpha q^\alpha - q^\alpha + q^\beta}{\alpha q^\alpha}$$

and thus (with n challenges and then taking the log) we need

$$2n\alpha\log q \geq n(\log\alpha + \alpha\log q) \gg n\log\left(\alpha q^\alpha - q^\alpha + q^\beta\right)$$

which gives in particular

$$n\alpha\log q \gg \log\alpha, \log q, \log q^\beta \qquad (14)$$

We use this on the size size requirement and obtain

$$\frac{\log\log I}{\log R} = \frac{\log\log q + \log\alpha + \log\binom{q^\beta+q^\gamma}{q^\gamma}}{n(\log\alpha + \alpha q^\gamma\log q)} \lesssim \frac{\log\binom{q^\beta+q^\gamma}{q^\gamma}}{n\alpha q^\gamma\log q},$$

then with the same upper bound on the binomial get

$$\frac{\log\log I}{\log R} \lesssim \frac{q^\gamma\log\mathrm{e}\frac{q^\beta+q^\gamma}{q^\gamma}}{n\alpha q^\gamma\log q} = \frac{1+\log\frac{q^\beta+q^\gamma}{q^\gamma}}{n\alpha\log q} \lesssim \frac{\log\left(1+q^\beta\right)}{n\alpha\log q}$$

which again goes to zero.

V. Performance

While Reed-Muller codes cannot achieve identification capacity, they were actually successful in our goal of implementing large identification codes with end-to-end time comparable with direct transmission and arbitrarily small error, as shown in figures Figs. 1 and 2. In order to achieve this performance, a combination of field size, computation optimization, and multiple challenges was used.

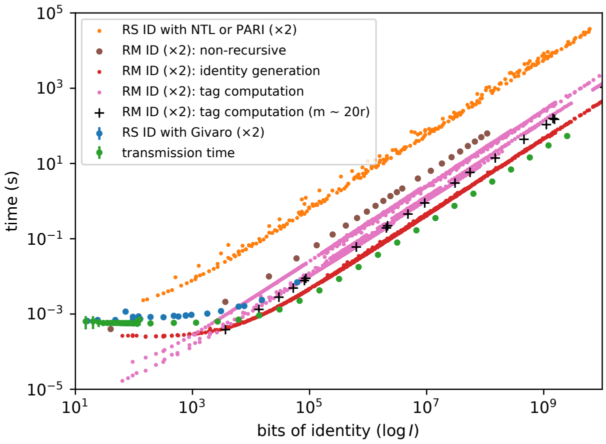


Fig. 1. Time cost of identity generation (red) and non-optimized (brown) and optimized encoding (pink and black) for Reed-Muller identification codes compared to the data from [7] (blue and orange: the cost of generation and encoding for concatenated Reed-Solomon identification codes; green: the cost of direct transmission with an experimental setup).

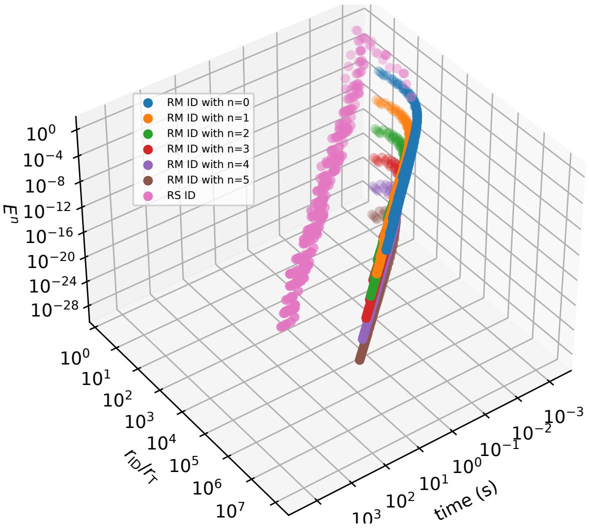


Fig. 2. Trade-off between time (generation, computation of n challenges and transmission) the error and the size of the codes (shown for $n = 1, ..., 6$) compared to the concatenated Reed-Solomon (pink points). The increased number of challenges successfully reduces the error without meaningfully impacting the computation time.

A. Field size

As identified in [7], the largest contribution to the computation time was the actual time of addition and multiplication operations in the Sagemath implementation. Limiting the field size to $q < 2^{16}$, where Zech tables of element logarithms are used, was the first step in achieving faster computation. As shown in Section A, this led us to addition and multiplication times being essentially equal and constant across any field size $q < 2^{16}$. Thus, choosing the largest field within the constraint allows to increase the size (Eq. (3)) and lower the error (Eq. (7)). However, the bound $q < 2^{16}$ also puts a lower bound on the error with a single challenge and thus multiple n challenges need to be used to reduce the error exponentially.

B. Computational optimization

The most efficient way of computing a polynomial is clearly to have it reduced into product of irreducible polynomials. However, the cost of the reduction contributes to the identification encoding. A fully reduced polynomial of degree k is computed in $\sim 2k$ operations. However, the same number of operations is sometimes achieved by computing the non reduced polynomial recursively. From a programming point recursion might introduce noticeable overhead and memory increase. Still, this means that we can optimize the number of operations without reduction. Recursion over degrees turned out to be too expensive and thus we use recursion only over variables as

$$p_w(r) = \sum\nolimits_{k'=[k]} r_1^{k'} p_{w_{k'}}(r_2 ... r_m),$$

where $w_{k'}$ denote a partition of coefficients for polynomials of degree $k - k'$ and $m - 1$ variables. Even then, the recursion turned out to be expensive as explained later below.

The computation time (times two, since the tag must be computed at the sender and at the receiver) is plotted in Fig. 1 in pink and black. For comparison, the brown points are the computation times without recursion. The improvement is larger than the caching optimization available for finite field computation (mentioned in Section A). The pink points actually form a band rather than a line, indicating that that there is room for optimization even in the choice of parameters k, m (q was already optimized as the largest $q < 2^{16}$). Here is where we can see that the recursion still constitutes an expensive contribution for large m; the black point are a heuristic selection of parameters satisfying $k/m \in [10, 50]$ indicating that the fastest computations happen for $k \gg m$. Section B describes a failed attempt to analytically predict this behaviour and extract the optimal parameters.

Finally, the red points represent the time spend randomly generating the identities w. We have timed the generation and the encoding separately since this contribution might not be relevant depending on the application. Since Fig. 1 is in log scale, the generation time is minor compared to the total time.

C. Multiple challenges

Figure 1 only shows size and computational time and thus does not show that the error of the Reed-Muller code increases with k and thus the size (Eq. (7)), as opposed to the Reed-Solomon code where it decreases. Multiple challenges can be used to reduce the error [7] at the cost of increasing computation time and transmission size. The decrease is exponential and thus only a small number of challenges is needed. The trade-off is displayed in Fig. 2 where the Reed-Muller code with a few challenges achieves points toward large size, small computation time and small error, more efficiently than the Reed-Solomon code.

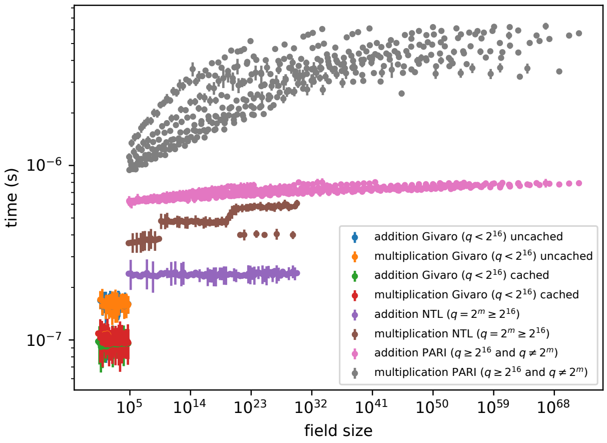


Fig. 3. Above: Time cost of addition and multiplication for all field sizes, depending on the prime power either Givaro, NTL or PARI implementations are used by Sage. With Givaro, there is an additional option to cache elements for faster computation.

VI. Conclusion

We have shown that it is possible to implement identification with latency comparable to current transmission speeds and arbitrarily small error. Better codes might even be able to be strictly faster than transmission in end-to-end identification. In particular, Polar codes are a potential candidate as they are characterized, among other things, by fast encoding times. Future work will also focus on verifying the advantage of identification in specific applications. Overall, our work shows that identification could potentially be an important technology in reducing traffic, load, latency in applications where the amount of data eventually grows faster than the capacity of the infrastructure.

Appendix A
Field additions and multiplications

We measured the time spent performing additions and multiplications at various field sizes q. The results are shown in Fig. 3. As expected, for $q \geq 2^{16}$ operation time increases considerably and multiplication time is noticeably larger than addition time. Multiplication and addition time is essentially the same for $q < 2^{16}$ and, maybe unexpectedly, is independent of q. Such result suggests that the optimal choice is to choose the largest field size below 2^{16} in order to reduce the error.

Finally, for $q < 2^{16}$ there is an option to cache field elements, which seems to improve operation times uniformly by a factor ~ 0.6. As seen in Sections B and V, other contributions influence the computation time more than the cache, making this factor not particularly relevant at the moment. We also did not investigate the memory impact of enabling the cache, which may be relevant in systems with limited memory, but is left for future work.

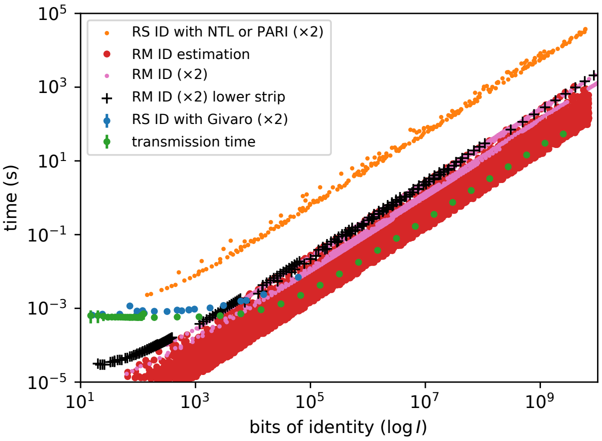


Fig. 4. Measured (black and pink) and estimated (red) time cost of Reed-Muller identification, together with the time cost of Reed-Solomon identification (orange and blue) and transmission (green) from [7] for comparison.

Appendix B
Analytic time complexity

We tried to estimate the time spent by the Reed-Muller identification encoder with the goal of estimating semi-analytically the best parameters in terms of time, size and error. However, the analysis of this estimation did not accurately predict the best measured parameter. This is explained in detail below together with possible further improvements.

The time estimation was done by simply counting the number of additions and multiplications performed by the recursive implementation of the polynomial. Let $t_+(q)$ and $t_*(q)$ be the times of performing one addition or multiplication respectively, and let us assume that exponentiation has the same cost as multiplication. The estimated time has a simple recursive relation given by

$$C(q,m,0)=0 \qquad C(q,1,k)=kt_+(q)+2kt_*(q)$$
$$\begin{aligned} C(q,m,k) &= kt_+(q)+kt_*(q)+\sum\nolimits_{k'\in[k]} C(m-1,k-k') \\ &= kt_+(q)+kt_*(q)+\sum\nolimits_{k'\in[k]} C(m-1,k') \end{aligned}$$

When operation time is constant $t_+(q)=t_*(q)=t$ as for $q<2^{16}$, the above cost function satisfies

$$C(q,m,k)=tC(m,k)$$

where $C(m,k)$ is $C(q,m,k)$ calculated with $t_+=t_*=1$. This suggests to use the largest field size below 2^{16} in order to reduce the error and increase the size of the Reed-Muller identification code, since no penalty is incurred in choosing these fields. The analysis can then focus on finding the best parameters m and k that optimize the estimated encoding time $C(m,k)$. By induction, the highest order term in $C(m,k)$ is $3\frac{k^m}{m!}$, however, since already the exact computation of $C(m,k)$ did not lead to the desired results, we did not investigate further how well $3\frac{k^m}{m!}$ approximates $C(m,k)$.

The estimated time plotted against the size $\log I$ is shown in the red points in Fig. 4 for $r,m=1,\dots,50$. The points form a band with the same slope as measured points (black and pink), suggesting that the bottom of the band could lead to optimized parameters. We divided it in strips and the lowest was used for the parameters measured in the black points, which however lie among the slowest points of the measured parameters. We take this as an indication that $C(m,k)$ is too simple to give accurate predictions. More accurate estimates could be achieved by including the cost of recursion and variables assignment, which is left for future work.

References

[1] R. Ahlswede and Z. Zhang. In: *IEEE Transactions on Information Theory* 41.4 (1995), pp. 1040–1050.

[2] Rudolf Ahlswede and Gunter Dueck. In: *IEEE Transactions on Information Theory* 35.1 (1989), pp. 30–36. DOI: 10.1109/18.42172.

[3] Rudolf Ahlswede and Gunter Dueck. In: *IEEE Transactions on Information Theory* 35.1 (1989), pp. 15–29.

[4] Philippe Delsarte, Jean-Marie Goethals, and F Jessie Mac Williams. In: *Information and control* 16.5 (1970), pp. 403–442.

[5] Sencer Derebeyoğlu, Christian Deppe, and Roberto Ferrara. In: *Entropy* 22.10 (2020), p. 1067. ISSN: 1099-4300. DOI: 10.3390/e22101067.

[6] T. Ericson and V. Zinoviev. In: *IEEE Transactions on Information Theory* 33.5 (1987), pp. 721–723. DOI: 10.1109/TIT.1987.1057344.

[7] Roberto Ferrara, Luis Torres-Figueroa, Holger Boche, Christian Deppe, Wafa Labidi, Ullrich Mönich, and Andrei Vlad-Costin. 2021. arXiv: 2107.06801 [cs.IT].

[8] Onur Günlü, Joerg Kliewer, Rafael F. Schaefer, and Vladimir Sidorenko. 2021. arXiv: 2106.13495 [cs.IT].

[9] T. Kasami, Shu Lin, and W. Peterson. In: *IEEE Transactions on Information Theory* 14.2 (1968), pp. 189–199. DOI: 10.1109/TIT.1968.1054127.

[10] G. Katsman, M. Tsfasman, and S. Vladut. In: *IEEE Transactions on Information Theory* 30.2 (1984), pp. 353–355. DOI: 10.1109/TIT.1984.1056879.

[11] K. Kurosawa and T. Yoshida. In: *IEEE Transactions on Information Theory* 45.6 (1999), pp. 2091–2095. DOI: 10.1109/18.782144.

[12] James L. Massey, Daniel J. Costello, and Jorn Justesen. In: *IEEE Transactions on Information Theory* 19.1 (1973), pp. 101–110. DOI: 10.1109/TIT.1973.1054936.

[13] Pierre Moulin and Ralf Koetter. In: *Security, Steganography, and Watermarking of Multimedia Contents VIII.* Vol. 6072. SPIE, 2006, pp. 565–574. DOI: 10.1117/12.644642.

[14] S. Verdu and V. K. Wei. In: *IEEE Transactions on Information Theory* 39.1 (1993), pp. 30–36. DOI: 10.1109/18.179339.

Learning Maximum Margin Channel Decoders for Additive Noise Channels

Amit Tsvieli and Nir Weinberger
The Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200004, Israel
amit.tsvieli@campus.technion.ac.il, nirwein@technion.ac.il

***Abstract*—The problem of learning a channel decoder for an additive noise channel whose noise distribution is nonparametric is considered. The learner is provided with a fixed codebook and a dataset comprised of independent samples of the noise, and is required to select a precision matrix for a nearest neighbor decoder in terms of the Mahalanobis distance. The objective of maximizing the margin of the decoder is addressed. Accordingly, a regularized loss minimization problem with a codebook-related regularization term and a hinge-like loss function is developed, which is inspired by the support vector machine paradigm for classification problems. Expected generalization error bound for that hinge loss is provided, and shown to scale at a rate of $O(1/(\lambda n))$, where λ is a regularization tradeoff parameter. A theoretical guidance for choosing the training signal-to-noise ratio is proposed based on this bound. A stochastic sub-gradient descent algorithm for solving the regularized loss minimization problem is proposed, and an optimization error bound is stated, which scales at a rate of $\tilde{O}(1/(\lambda T))$. The performance of the proposed algorithm is demonstrated through an example.**

I. Introduction

The choice of a proper channel decoder is a key element in the design of a communication system, and is typically based on rich expert knowledge on the statistical model of the channel operation. In this paper, we address a scenario in which such knowledge is *not available.* Specifically, we consider an additive noise channel, whose noise distribution is unknown, and is also not known to belong to any parametric family. We develop a learning algorithm which selects a proper decoder from the class of nearest neighbor (NN) decoders with respect to (w.r.t.) the Mahalanobis distance (MD), based on noise samples. The class of possible decoders is thus parameterized by the precision matrix defining the MD.

The approach considered here follows the common practice of partitioning the communication epoch to a *training phase* – in which no data is transmitted, and the transmitter sends a known signal which the receiver uses to select a decoder, and a *data phase* – in which the decoder is fixed (or only tracks slight variations in the channel statistics). Typically, a parametric form is assumed for the channel statistics. Then, the training phase is used to estimate the parameter and the decoder is chosen to match the estimated parameter. In various scenarios of interest, e.g. interference in massive multiple-input multiple-output systems [1] or ultra low-latency communication [2], parameter estimation, or even the parametric modeling itself, may be inaccurate. In order to address such cases, in this work we make no assumptions regarding the distribution of the channel. This setting naturally motivates the use of *machine-learning* methods, as they are typically *distribution-free*, that is, do not make any assumptions on the data statistics, which for the additive noise channel corresponds to the noise probability distribution function. The learning process we propose, however, is strongly tailored to the given codebook used by the encoder, and the additive structure of the channel.

Following [3], we consider the class of NN decoders, w.r.t. the MD, that is, decoders parameterized by a precision matrix. The optimality of this class of decoders for the additive Gaussian noise channels motivates their usage when no assumption is made on the noise distribution. Following similar reasoning, for Gaussian noise, the signal-to-noise ratio (SNR) scaling of the error probability is determined by the *margin*, or *minimal distance*, between codewords (w.r.t. the MD). It is thus plausible to use the same performance criterion for general unknown noise distributions, and aim the learner to select a decoder which maximizes this margin. Evidently, this reasoning parallels the approach of support vector machines (SVM), in which a learned classifier is not only required to obtain low classification error on the training data, but also to maximize the *margin* between the classes.

The contributions and the outline of the rest of the paper is as follows. In Section II, we establish notation conventions and formulate the learning problem. In Section III, we formulate a maximum margin regularized risk minimization (RLM) rule for the problem. In Section IV, we prove a $O(1/(\lambda n))$ generalization error bound for the RLM, where λ is the regularization tradeoff parameter, and n is the number of noise samples. This tractable optimization problem suffers from large complexity, mainly due to a $O(n^2)$ dependence. To circumvent this, in Section V, we develop a stochastic sub-gradient descent algorithm for the decoder learning problem, and prove that $\tilde{O}(1/\epsilon)$ iterations suffices in order to obtain a solution of accuracy ϵ. We stress that this bound does not depend on the dimension of the channel or the number of noise samples, which are abundant in many scenarios. In Section VI, we exemplify the operation of the algorithm through an example. All proofs, further discussions, simulations and topics for further research are available in a full version of the paper [4].

II. Formulation of the Decoder Learning Problem

We begin with a few notation conventions. Random variables or vectors are denoted by capital letters and specific values they take are denoted by the corresponding lower case letters. The expectation operator is denoted by $\mathbb{E}_{\mu}[\cdot]$ where μ is the underlying probability measure. The indicator of an event $\mathcal{A}$ is denoted by $\mathbb{I}\{\mathcal{A}\}$. All vectors are taken as column vectors. The standard Euclidean norm for $x \in \mathbb{R}^d$ is denoted by $\|x\|$ and the inner product by $\langle x_1, x_2 \rangle$ or $x_1^T x_2$, interchangeably. The Frobenius norm for a matrix $A \in \mathbb{R}^{d\times d}$ is denoted by $\|A\|_F$. The positive semi-definite (PSD) cone is denoted by $\mathbb{S}_+$. For $n \in \mathbb{N}^+$, the set $\{1, 2, \ldots, n\}$ is denoted by $[n]$. Standard Bachmann-Landau asymptotic notation is used, where specifically, $\tilde{O}(\cdot)$ is such that the logarithmic factors are hidden, namely, $f(n) \in \tilde{O}(h(n)) \iff \exists k : f(n) \in O(h(n)\log^k(h(n)))$.

Consider the problem of communicating over a d-dimensional additive noise channel $Y = X + Z$, where $Y \in \mathbb{R}^d$ is the channel output, $X \in \mathbb{R}^d$ is a codeword that is chosen from a fixed given codebook $C = \{x_j\}_{j\in[m]}$ with a uniform probability, and $Z \in \mathbb{R}^d$ is a noise statistically independent of the input X. The distribution μ of the noise Z is unknown to the designer of the decoder, and is not known to belong to any parametric family. Further consider the class of NN decoders w.r.t. the MD

$$\hat{j}\left(y\right) \in \operatorname*{arg\,min}_{j\in[m]} \|x_j - y\|_S \tag{1}$$

$$\triangleq \operatorname*{arg\,min}_{j\in[m]} \sqrt{\left(x_j - y\right)^T S \left(x_j - y\right)}, \tag{2}$$

where $S \in \mathbb{S}_+^d$ is a precision matrix (the inverse of a covariance matrix). In what follows, a decoder from this class will be identified by its precision matrix S.

For a decoder S, the expected error probability conditioned that the j-th codeword was transmitted is given by

$$\boldsymbol{p}_{\mu}\left(S \mid j\right) \triangleq \mathbb{E}_{\mu}\left[\mathbb{I}\left\{\min_{j'\in[m]\setminus\{j\}} \left\|x_j + Z - x_{j'}\right\|_S < \|Z\|_S\right\}\right], \tag{3}$$

and the expected error probability averaged over all codewords is given by $\boldsymbol{p}_{\mu}\left(S\right) = \frac{1}{m}\sum_{j\in[m]} \boldsymbol{p}_{\mu}\left(S|j\right)$. A learner, which does not know μ, is provided with n noise samples $\boldsymbol{Z} = \{Z_i\}_{i\in[n]}$ drawn i.i.d. from μ (as well as the given codebook C), and is required to find S which minimizes the expected error probability. A common learning approach is empirical risk minimization (ERM), in which the empirical average error probability of the noise samples, given by

$$\boldsymbol{p}_{\boldsymbol{Z}}\left(S\right) \triangleq \frac{1}{m}\sum_{j\in[m]} \frac{1}{n}\sum_{i\in[n]} \left[\mathbb{I}\left\{\min_{j'\in[m]\setminus\{j\}} \left\|x_j + z_i - x_{j'}\right\|_S < \|z_i\|_S\right\}\right], \tag{4}$$

is minimized by the learner. This ERM problem has been studied in [3], and is difficult to solve, mainly due to the loss function being discontinuous in S. Another disadvantage of this approach is that it does not capture the structure of the codebook and therefore its generalization bound depends linearly on m, as was proved and discussed therein. Therefore, in this paper, we take a different approach, and derive learning rules which attempt to maximize the *margin* of the decoder.

The decoder learning problem resembles a multiclass classification problem, in which the decoder is required to classify every channel output as the outcome of one input codeword. Thus, as a starting point, we assume that the learner synthesizes the following dataset of mn labeled samples, in which each of the scaled codewords in the codebook is perturbed by all noise samples $\{z_i\}_{i\in[n]}$, namely $\boldsymbol{D}(\boldsymbol{Z}) \triangleq \{y_k, l_k\}_{k=1}^{mn} = \bigcup_{j\in[m]} \boldsymbol{D}_j(\boldsymbol{Z})$,where $\boldsymbol{D}_j(\boldsymbol{Z}) \triangleq \{\Gamma \cdot x_j + z_i, j\}_{i\in[n]}$, and $\Gamma > 0$ is a scaling constant which determines the *training SNR*. For the sake of brevity, we will omit from now on the explicit dependence of $\boldsymbol{D}$ in $\boldsymbol{Z}$. Let the zero-one loss for the multiclass classification problem be denoted by $\ell^{0-1}(l', l) \triangleq \mathbb{I}[l' \neq l]$ and let the corresponding empirical average loss be denoted by $L_{\boldsymbol{D}}^{0-1}(S) \triangleq \frac{1}{mn}\sum_{k\in[mn]} \mathbb{I}\{\hat{l}\left(y_k\right) \neq l_k\}$. A simple observation is that the empirical average error probability over $\boldsymbol{Z}$ is the same as the empirical average loss over $\boldsymbol{D}$, that is, $\boldsymbol{p}_{\boldsymbol{Z}}(S) = L_{\boldsymbol{D}}^{0-1}(S)$.

Despite the equivalence of the risk in both problems, decoder learning is different from multiclass classification, and standard SVM learning algorithms cannot be directly applied. This, in fact, is apparent from the synthesized dataset: Unlike datasets of regular classification problems, this synthesized dataset has *structure* (multiple translations of the noise samples dataset). Moreover, the *training SNR* is a design parameter that can be chosen by the learner. One consequence of this possibility is in contrast to standard classification problems, in which the margin prevailing in the dataset determines the sample complexity of the problem [5, Thm. 15.4], the margin in the decoder-learning problem is a parameter to be optimized. This can be used in order to achieve the best generalization possible for a given dataset.

III. Maximum Margin Decoder Learning Algorithm

A NN decoder partitions the output space into m decision regions, whose boundaries are d-dimensional hyperplanes. This NN decoding rule maximizes the minimal MD between each pair of codewords in the codebook, w.r.t. the noise covariance matrix. As is well known, for Gaussian additive noise channels at high SNR, this minimal distance $d_{\min}$ is the dominant parameter in determining the decay rate of the error probability w.r.t. the SNR, as evident from the upper bound $P_e \leq (m-1)\cdot\exp(-d_{\min}^2/8\sigma^2)$ [6, Sec. 5.2] (where σ^2 is the Gaussian noise power). Naturally, such a bound does not necessarily hold for non-Gaussian noise distributions. However, such a criterion is plausible to adopt for general noise distributions in the absence of any other knowledge. Thus, in this section, we formulate a maximum margin problem for a decoder, which is partially analogous to the maximum margin problem in SVM. The development of the optimization problem will be made in several steps, which we next describe.

Step 1 – maximization of the minimum margin: We begin with the assumption that the dataset $\boldsymbol{D}$ is *separable* i.e., there exists a precision matrix S that achieves zero loss over $\boldsymbol{D}$. This assumption is analogous to the linear separability assumption made for hard SVM, and will be relaxed in the following steps, similarly to the way it is is relaxed for soft SVM [5, Ch. 15]. The *margin* of a hyperplane w.r.t. a dataset is defined to be the minimal distance between a point in the dataset and the hyperplane [5, Ch. 15]. We denote the difference between a pair of codewords indexed by (p,q) by $\delta_{pq} \triangleq x_p - x_q$. In addition, for a given ordered pair of codeword indices (p,q) and a sample $y_i \in \boldsymbol{D}_p \bigcup \boldsymbol{D}_q$ we denote the following transformation of the sample and codewords $a_{pqi} \triangleq (-1)^{\mathbb{I}(i \in \boldsymbol{D}_q)}(y_i - \frac{1}{2}(x_p + x_q))$. The learner's goal is to find a precision matrix S that maximizes the minimum margin, over all codeword-pairs $x_p, x_q \in C$ and the learning problem is formulated as follows:

Claim 1. The maximum margin induced by a MD NN decoder with precision matrix S is

$$\max_{S \in \mathbb{S}_+} \min_{1 \le p < q \le m} \min_{i \in \boldsymbol{D}_p \cup \boldsymbol{D}_q} \frac{a_{pqi}^T S \delta_{pq}}{\|S\delta_{pq}\|}. \tag{5}$$

Step 2 – a convex lower bound: The problem (5) is not necessarily convex, and therefore we proceed to maximize the following convex lower bound on its value.

Claim 2. The optimization problem

$$\begin{array}{c} \max_{S \in \mathbb{S}_+} \min_{1 \le p < q \le m} \min_{i \in \boldsymbol{D}_p \cup \boldsymbol{D}_q} a_{pqi}^T S \delta_{pq} \\ \text{subject to } \max_{1 \le p < q \le m} \|S\delta_{pq}\| \le 1 \end{array} \tag{6}$$

is a convex lower bound on the value of (5).

Step 3 – minimum norm formulation: With the prospect removal of the separability assumption, we next derive a minimum norm optimization problem, so that every solution to it is a solution to (6). This is similar to the equivalent formulation of hard SVM from [5, Lem. 15.2].

Lemma 3. *Every solution to the following minimum norm problem is a solution to* (6)*:*

$$\begin{array}{c} \min_{S \in \mathbb{S}_+} \max_{1 \le p < q \le m} \|S\delta_{pq}\|^2 \\ \text{subject to } \min_{1 \le p < q \le m} \min_{i \in \boldsymbol{D}_p \cup \boldsymbol{D}_q} a_{pqi} S \delta_{pq} \ge 1 \end{array} . \tag{7}$$

Step 4 – relaxation of the separability assumption: Next, we introduce slack variables in order to relax the assumption that the dataset $\boldsymbol{D}$ is separable. This is similar to the relaxation made for soft SVM [5, Ch. 15.2]. Following a short derivation, the result is a RLM problem for a specific *hinge loss* function over a transformation of the noise samples. Specifically, we first denote $\mathring{\ell}^{\text{hinge}}(S,p,q,i) \triangleq \max\{0, 1 - a_{pqi} S \delta_{pq}\}$, and define the empirical risk $\mathring{L}_{\boldsymbol{D}}^{\text{hinge}}(S)$ as the average *hinge loss* of the induced binary linear classifiers $\{S\delta_{pq}\}$ over the transformed noise samples $\{a_{pqi}\}$, to wit,

$$\mathring{L}_{\boldsymbol{D}}^{\text{hinge}}(S) \triangleq \frac{2}{m(m-1)} \sum_{1 \le p < q < m} \frac{1}{2n} \sum_{i \in \boldsymbol{D}_p \bigcup \boldsymbol{D}_q} \mathring{\ell}^{\text{hinge}}(S,p,q,i). \tag{8}$$

The RLM problem is then given by

$$\min_{S \in \mathbb{S}_+} \mathring{L}_{\boldsymbol{D}}^{\text{hinge}}(S) + \lambda \cdot \max_{1 \le p < q \le m} \|S\delta_{pq}\|^2, \tag{9}$$

where λ is a parameter that controls the tradeoff between the two terms. In [3], a *different* surrogate hinge-type upper bound for the average error probability loss over Z was proposed, which was defined there as

$$\bar{L}_{\boldsymbol{Z}}^{\text{hinge}}(S) \triangleq \frac{1}{n} \sum_{i \in [n]} \frac{1}{m} \sum_{p \in [m]} \max\left\{0, 1 - \min_{q \in [m] \setminus \{p\}} \|x_p + z_i - x_q\|_S - \|z_i\|_S\right\}. \tag{10}$$

We review the differences between these hinge-type losses in light of two possible interpretations for the SVM optimization problem. The first interpretation is that the SVM objective function is a specific convex and continuous upper bound to the non-convex and discontinuous zero-one loss function, chosen so its ERM problem can be efficiently solved. The regularization term is interpreted as Tikhonov regularization. The second interpretation interprets the objective function as a balance between increasing the margin and increasing classification errors. As is well known, for binary classification, both interpretations lead to exactly the same optimization problem [5, Ch. 15]. This is, however, not the case for the decoder learning problem. The hinge loss (10) from [3] follows the first interpretation, as a relaxation of the error probability loss function. This approach leads to an ERM problem with the hinge loss over the noise samples, and an implicit regularization in the form of maximal eigenvalue constraint. However, this hinge loss is not directly related to the margin. In this paper, we follow the second interpretation for maximizing the margin induced by the channel decoder. This approach leads to the RLM problem with the hinge loss over the transformed noise samples, and a codebook related regularization in (9). We argue that for the channel decoding problem the second approach is better since error probability is strongly related to margin, as briefly discussed above.

Step 5 – inducing stability by a generalization of the regularization: Various generalization bounds for SVM are based on the *stability* of its learning rule. However, the problem (9) is, in general, not stable because the regularization term $\max_{1 \le p < q \le m} \|S\delta_{pq}\|^2$ is indifferent to changes in directions orthogonal to the maximizer δ_{pq}. Nonetheless, we next assume, without loss of generality, that $\text{Span}\{\delta_{pq}\}_{1 \le p < q \le m} = \mathbb{R}^d$ (if this is not the case, we can project the codebook and noise samples to a lower dimension spanned by the codebook). We next slightly modify the learning rule to a stable one, and to this end, we consider a partition of the codeword pairs which satisfies the following property.

Definition 4. A partition $P = \bigcup_{j=1}^{d+1} P_j$ of $\{(p,q)\}_{1 \le p < q \le m}$ is *proper* if $\text{Span}[\{\delta_{p_j,q_j}\}_{j=1}^{d+1}] = \mathbb{R}^d$ for any set $\{\delta_{p_j,q_j}\}_{j=1}^{d+1}$ of representatives, such that $(p_j, q_j) \in P_j$ for all $j \in [d+1]$.

A simple way of finding a proper partition is by first finding a basis of $\mathbb{R}^d$: $\{\delta_{p_j,q_j}\}_{j=1}^{d} \subset \{\delta_{pq}\}_{1 \le p < q \le m}$ and then setting $P_j = \{\delta_{p_j,q_j}\} \forall j \in [d]$, $P_{d+1} = \{\delta_{pq}\}_{1\leq p<q\leq m} \setminus \{\delta_{p_j,q_j}\}_{j=1}^{d}$. Nonetheless, our following results hold for any arbitrary proper partition. For notational convenience, we will henceforth occasionally use a single index in $[\frac{1}{2}m(m-1)]$ instead of double indices $\{(p,q)\}_{1\leq p<q\leq m}$. The final RLM rule for finding a maximum minimum margin decoder is defined for a given positive parameters $\{\eta_i\}_{i\in[d+1]}$ which satisfy $\sum_{i=1}^{d+1}\eta_i = 1$, and a proper partition $\{P_j\}_{j\in[d+1]}$, as

$$A\left(\boldsymbol{D}\right) = \operatorname*{arg\,min}_{S\in\mathbb{S}_+} \mathring{L}_{\boldsymbol{D}}^{\text{hinge}}\left(S\right) + \lambda\sum_{i=1}^{d+1}\eta_i \cdot \max_{j\in P_i}\left\|S\delta_j\right\|^2. \quad (11)$$

IV. A Generalization Error Bound

In this section, we state an average generalization error bound for (11), through on-average-replace-one-stability argument, following the proof in [5, Sec. 13.3]. The generalization error is the error inflicted by learning only from noise samples and not the noise distribution itself. This generalization bound can be used to evaluate the expected loss a solution to (11) will achieve, given the empirical error. Furthermore, we will also utilize it to obtain a theoretical guidance for choosing the training SNR.

Theorem 5. *Let A be the RLM rule* (11). *Then,*

$$\mathbb{E}_\mu\left[\mathring{L}_\mu^{\text{hinge}}\left(A\left(\boldsymbol{D}\left(\boldsymbol{Z}\right)\right)\right) - \mathring{L}_{\boldsymbol{D}}^{\text{hinge}}\left(A\left(\boldsymbol{D}\left(\boldsymbol{Z}\right)\right)\right)\right] \leq \frac{1}{\lambda n\eta_{\text{min}}}\mathbb{E}_\mu\left[h\left(C,\boldsymbol{Z}\right)\right], \quad (12)$$

where $h(C,\boldsymbol{Z})$ is a codebook and dataset dependent constant

$$h\left(C,\boldsymbol{Z}\right) \triangleq \max_{1\leq p<q\leq m}\max_{i\in[n]} \frac{\left|\langle z_i,\delta_{pq}\rangle\right|\left\|\delta_{pq}\right\|^2 + \langle z_i,\delta_{pq}\rangle^2 + \left\|z_i\right\|^2\left\|\delta_{pq}\right\|^2 + \frac{1}{2}\left\|\delta_{pq}\right\|^4}{\min_{1\leq p<q\leq m}\left\|\delta_{pq}\right\|^2}. \quad (13)$$

Furthermore, if $r_x \triangleq \max_{x\in C}\|x\|$ and the noise is bounded, i.e., $\|z\| \leq r_z$ w.p. 1, then

$$\mathbb{E}_\mu\left[h\left(C,\boldsymbol{Z}\right)\right] \leq \frac{16r_z r_x^3 + 8r_z^2 r_x^2 + 8r_x^4}{\min_{1\leq p<q\leq m}\left\|\delta_{pq}\right\|^2}. \quad (14)$$

We remark that (14) can be easily generalized for weaker assumptions on the noise distribution tail, e.g. sub-Gaussian. In [3], a $\tilde{O}(m\sqrt{d/n}+\sqrt{\log\left(1/\delta\right)/n})$ high-probability generalization error bound was proved for the error probability loss function, as well as a $\tilde{O}(\sqrt{d\left(d+m\right)/n} + \sqrt{\log\left(1/\delta\right)/n})$ high probability generalization error bound for the surrogate hinge-type upper bound (10). In comparison, here we prove a $O(1/(\lambda n))$ generalization error bound on the regularized hinge loss over the transformed samples. The convergence rate of this bound is much faster, however, this is only an average error bound, and does not have a high probability guarantee.

Previous works (e.g., [7], [8]), discussed the question of how to optimize the training SNR. Intuitively, on one hand, training with a sufficiently high SNR leads to zero empirical error for many decoders, not necessarily the one with the lowest expected error. On the other hand, training with SNR too low may produce a decoder which has high error probability (as most evident from the extreme case of zero SNR), and might be too pessimistic in assessing the error probability. In [8] a rule-of-thumb for choosing the training SNR was proposed, based on the capacity of the Gaussian channel. This rule, however, did not take into account generalization error aspects. Following the generalization error bound of Theorem 5, we propose to tune the training SNR so that the empirical error $\mathring{L}_{\boldsymbol{D}}^{\text{hinge}}(A(\boldsymbol{D}(\boldsymbol{Z})))$ roughly equals to the generalization bound (12). With this training SNR, it is guaranteed that the expected error is on the same order as the empirical error.

The generalization bound from Theorem 5 decreases with increasing λ, whereas the empirical hinge loss in (11) increases with increasing λ. Similarly to [5, Cor. 13.9], λ may be optimized as follows:

Theorem 6. *Let $\mathcal{S}_B \triangleq \{S \in \mathbb{S}_+ \colon \max_{j\in[\frac{1}{2}m(m-1)]}\|S\delta_j\| \leq B\}$, denote $\rho \triangleq \max_{p,q,i}\rho_{p,q,i}$, and let $\lambda = \sqrt{\frac{2\rho^2}{B^2\gamma n}}$. Then, the RLM rule* (11) *with $\mathbb{S}_+$ replaced with $\mathcal{S}_B$ satisfies*

$$\mathbb{E}_\mu\left[\mathring{L}_\mu^{\text{hinge}}\left(A\left(\boldsymbol{D}\left(\boldsymbol{Z}\right)\right)\right)\right] \leq \min_{S\in\mathcal{S}_B}\mathring{L}_\mu^{\text{hinge}}\left(S\right) + \rho B\sqrt{\frac{8}{\gamma n}}. \quad (15)$$

Hence, for every $\epsilon > 0$, if $n \geq \frac{8\rho^2B^2}{\gamma\epsilon^2}$ then for every distribution μ, $\mathbb{E}_\mu[\mathring{L}_\mu^{\text{hinge}}(A(\boldsymbol{D}(\boldsymbol{Z})))] \leq \min_{S\in\mathcal{S}_B}\mathring{L}_\mu^{\text{hinge}}(S) + \epsilon$.

It should be remarked, however, that the bound of Theorem 6 is not readily translated to a bound on the error probability itself (unlike in binary classification).

V. A Stochastic Sub-gradient Descent Algorithm

In this section, we propose a stochastic sub-gradient descent algorithm for solving (11), inspired by an algorithm for classification called PEGASOS [9]. Its $\tilde{O}(1/\epsilon)$ run-time, for an ϵ-accurate solution, is independent of the dataset size n, which makes the algorithm especially suitable for learning from large datasets, just as an offline design of the decoder. Moreover, even if the decoder is learned online and the number of samples is relatively small, a learning rule based on low-complexity iteration is also attractive, since communication devices are typically limited in computational power (as, for example, motivates learning equalizers by the least mean squares algorithm [10, Ch. 9]). In comparison, and as discussed in [11], the computational cost of solving a standard SVM problem grows at least like $O(n^2)$, and even if the solver is efficient in the data-laden regime, in which data is virtually unlimited, it has a worse dependence on ϵ compared to the sub-gradient descent algorithm [12]. The pseudo code of our proposed algorithm is given in Algorithm 1 and requires the following definitions. Let p_a, q_a be the codeword-pair related to a transformed sample a, let $\delta_a \triangleq \delta_{p_a,q_a}$ and let $j_k^{(t)} \triangleq \arg\max_{j\in P_k}\|S_t\delta_j\|^2$. The sub-gradient of the approximate objective at round t is:

$$\nabla_t \triangleq \lambda\sum_{k=1}^{d+1}\eta_k\left(S_t\delta_{j_k^{(t)}}\delta_{j_k^{(t)}}^T + \delta_{j_k^{(t)}}\delta_{j_k^{(t)}}^T S_t\right)$$

Algorithm 1 RLM Sub-gradient Algorithm

```
input $\boldsymbol{D} \in \mathbb{R}^{d\times n} \times \mathbb{N}^n, \lambda \in \mathbb{R}^+, T \in \mathbb{N}, c \in \mathbb{N}$
begin
  Set $S_1 = 0$
  for $t = 1, 2, \ldots, T$
    Choose $A_t \subset [nm(m-1)]$, s.t. $|A_t| = c$, uniformly at random.
    Set $S_{t+1} \leftarrow \Pi_{\mathbb{S}_+}(S_t - \frac{1}{\lambda t}\nabla_t)$
  end for
end
output $S_{T+1}$
```

$$-\frac{1}{|A_t|}\sum_{i\in A_t}\mathbb{I}\left[a_i^T S_t \delta_{a_i} < 1\right]\frac{1}{2}\left(\delta_{a_i} a_i^T + a_i \delta_{a_i}^T\right). \tag{16}$$

Notice that all matrix derivatives are w.r.t symmetric matrices. In general, a gradient step may result in $S_{t+1} \notin \mathbb{S}_+$ hence we include an obligatory projection step to the PSD cone. According to [13, Ch. 8], the projection of a symmetric matrix to the PSD cone w.r.t. the Frobenius norm is, $\Pi_{\mathbb{S}_+}(S) \triangleq \sum_{i=1}^{d} \max\{\lambda_i, 0\} v_i v_i^T$, and this definition is used here. We prove the following optimization error bound.

Theorem 7. *Let S^* be the RLM rule* (11) *and S_t be the hypothesis generated by algorithm 1 at a random round $t \in [T]$. Denote $f(S) \triangleq \mathring{L}_{\boldsymbol{D}}^{\text{hinge}}(S) + \lambda \sum_{i=1}^{d+1} \eta_i \cdot \max_{j\in P_i} \|S\delta_j\|^2$, and assume that for all t, each element in A_t is sampled uniformly at random from the dataset (with or without replacement). Then,*

$$f(S_t) - f(S^*) = O\left(\frac{\ln^3(T) \cdot \ln(1/\delta)}{\lambda T}\right) \tag{17}$$

with probability larger than $\frac{1}{2} - 2\delta \ln(T)$.

Since Theorem 7 guarantees a good solution with probability close to $\frac{1}{2}$ then, as discussed in [9], roughly *two* validation attempts are required to obtain a good solution. The bounds of Theorems 6 and 7 result the following total error bound. First, the choice of the class of NN decoders, which do not necessarily contain the optimal decoder for the true noise distribution, inflicts an *approximation error*, which is difficult to characterize in general. Second, learning the decoder based on noise samples rather than the unknown noise distribution, inflicts a *generalization error*, which was bounded in Theorem 6 with expected generalization error of $O(1/(\lambda n \eta_{\min}))$. Third, solving the optimization problem only approximately, using a finite number T of iterations of Algorithm 1, inflicts an *optimization error*. Theorem 7 establishes that this error is $O(\ln^3(T)/\lambda T)$.

VI. An Example

In this section, we exemplify the empirical performance of the proposed algorithm for a two-dimensional zero-mean Gaussian mixture noise of $l = 4$ Gaussians $Z_i \sim \mathcal{N}(0, K_i)$, and mixture weights $\{\omega_i\}_{i=1}^{l}$. The learner is provided with a codebook of $m = 32$ points, which is a 32-PAM constellation, and a training set with $n = 100$ i.i.d. noise samples. Additionally, the tradeoff parameter is $\lambda = 0.1$, the training SNR is 16[dB] and the proposed Algorithm 1 has run for $T = 100$ iterations with a batch size of 1, and produced hypotheses $\{S_t\}_{t=1}^{T}$. We compare our results with $\hat{S} \triangleq (\frac{1}{l}\sum_{i=1}^{l}\omega_i K_i)^{-1}$ which is a reasonable choice but *not* necessarily optimal. The left panel of Fig. 1 shows the fast convergence of the algorithm to an approximately minimal error solution. The right panel displays the error probability of the final hypothesis decoder, under various SNR values, for each value a validation set with 10^3 noise samples is used. We observe that the error probability scales similarly for both S_T and $\hat{S}$, even for SNR values that are significantly smaller than the training SNR.

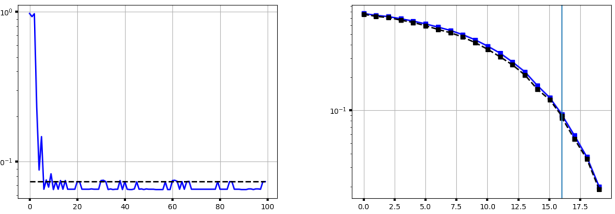


Figure 1. Left: train error probability vs. training iterations. Right: error probability vs. various SNRs [dB]. S_t - blue, $\hat{S}$ - dashed.

References

[1] T. L. Marzetta, "Massive MIMO: An introduction," *Bell Labs Technical Journal*, vol. 20, pp. 11–22, 2015.

[2] M. Sybis, K. Wesolowski, K. Jayasinghe, V. Venkatasubramanian, and V. Vukadinovic, "Channel coding for ultra-reliable low-latency communication in 5G systems," in *2016 IEEE 84th vehicular technology conference (VTC-Fall)*, pp. 1–5, IEEE, 2016.

[3] N. Weinberger, "Learning additive noise channels: Generalization bounds and algorithms," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2586–2591, IEEE, 2020. https://drive.google.com/file/d/11snRohUPiICM62SJJZzZhSpweInEBAaa/view?usp=sharing.

[4] A. Tsvieli and N. Weinberger, "Learning maximum margin channel decoders for additive noise channels." unpublished, available at https://drive.google.com/file/d/1bUNA6R0ybPqTeMFE0i2wfgWj4JcC1FYx/view?usp=sharing, 2021.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[6] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding.* Courier Corporation, 2013.

[7] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2017.

[8] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," *arXiv preprint arXiv:1805.09317*, 2018.

[9] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-Gradient SOlver for SVM," 2007. A fast online algorithm for solving the linear svm in primal using sub-gradients.

[10] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital communication.* Springer Science & Business Media, 2012.

[11] L. Bottou and C.-J. Lin, "Support vector machine solvers," *Large scale kernel machines*, vol. 3, no. 1, pp. 301–320, 2007.

[12] S. Shalev-Shwartz and N. Srebro, "SVM optimization: inverse dependence on training set size," in *Proceedings of the 25th international conference on Machine learning*, pp. 928–935, 2008.

[13] S. P. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge university press, 2004.

Mismatched Estimation of Symmetric Rank-One Matrices Under Gaussian Noise

Farzad Pourkamali and Nicolas Macris
SMILS, EPFL, Lausanne, Switzerland
Emails: {farzad.pourkamali,nicolas.macris}@epfl.ch

***Abstract*—We consider the estimation of an n-dimensional vector s from noisy element-wise measurements of ss^T, a problem that frequently arises in statistics and machine learning. We investigate a mismatched Bayesian inference setting in which the statistician is unaware of some of the parameters. For the particular case of Gaussian priors for the vector s and additive noise, we derive the complete exact analytic expression for the asymptotic mean squared error (MSE) in the large system size limit. Our formulas demonstrate that estimation is still possible in the mismatched case. Also, the minimum MSE (MMSE) can be achieved by selecting a non-trivial set of parameters beyond the matched case. Our technique is based on the asymptotic behavior of spherical integrals and can be used as long as the statistician chooses a rotationally invariant prior.**

I. Introduction

Many problems in machine learning and statistics can be expressed as estimating a low-rank matrix from its noisy observation. Examples are sparse PCA [1], the spiked Wigner model, community detection [2]. For the rank-one symmetric case, the problem is formulated as follows: a vector $\mathrm{s} \in \mathbb{R}^n$ is generated with i.i.d. elements distributed according to $s_i \sim \mathbf{P}^*$, the matrix ss^T is observed through an element-wise additive white gaussian noise channel. The goal is to estimate the vector s upon observing the noisy version of $\mathbf{ss}^T$.

The statistical and computational limits of this problem have been extensively studied. Most works have so far considered the "Bayes-optimal" setting, in which the prior $\mathbf{P}^*$ and possibly other hyper-parameters (e.g., SNR) are known to the statistician. In the Bayes-optimal setting, computing the mutual information enables us to compute the minimum mean squared error (MMSE) and derive the information-theoretical limits of the estimation. The analytical but highly non-rigorous replica and cavity methods rooted in statistical physics have been used to derive expressions for the mutual information between the true signal and the observation matrix [3]. These expressions were already rigorously derived in early work [4] for binary signals using Guerra-Toninelli interpolation [5]. Later the problem has been studied in much detail for general signals, in [6] using approximate message passing (AMP) and spatial coupling, in [7] by Guerra-Toninelli interpolation and Aizenman-Sims-Starr methods. Further, [8], [9] used the adaptive interpolation method to rigorously prove the limiting expressions of mutual information and MMSE. All these methods crucially rely on the assumption that the prior and the parameters of the estimation problem are known to the statistician. The Bayes law then induces remarkable identities that enable the analysis to proceed. In the present case, we lack such identities.

Despite the vast amount of work on this problem in the Bayes-optimal setting, to the best of our knowledge, there is no rigorous result for the *mismatched* case corresponding to the realistic situation where the statistician does not know the true prior or/and hyper-parameters, and can only make assumptions about them. Mismatched inference for the scalar and vector estimation problems has been considered in [10], [11]. In particular, [10] proved a result relating the MSE in the mismatched inference to the relative entropy of the true prior and the statistician's prior. We follow this work and define the MSE similarly (up to natural modification for the matrix case).

The main contribution of this paper is to compute the asymptotic mismatched MSE for the rank-one matrix estimation problem in the large n limit. Our approach uses results on the spherical integrals from the mathematical physics literature [12]. A primary assumption in our method that would be difficult to dispense of, is the rotational invariance of the statistician's prior. Despite this restriction, we can study non-rotation invariant true priors, non-symmetric matrix estimation, higher-ranks (finite w.r.t $n \to +\infty$). In this short note, we limit ourselves to the theoretical limits of mismatched estimation for the case of Gaussian priors (both for the true and the statistician's) and postpone the detailed study of the more general cases to a forthcoming detailed work. As will become clear in section III, already under this limited setting, the phase transitions phenomenology is quite rich.

The rest of the paper is organized as follows. In Section II, we introduce the setting and formulate the problem. Section III describes the main result and discusses it in several special cases, followed by the proof sketch of the main theorem in Section IV. Lastly, we conclude the paper with some remarks and possible future directions for this line of work.

II. Problem Setting

Suppose the ground-truth vector $\mathrm{s} \in \mathbb{R}^n$ is generated with i.i.d. elements from $\mathbf{P}^* = \mathcal{N}(0, \sigma^2)$, the observed matrix is

$$\boldsymbol{Y} = \sqrt{\frac{\lambda}{n}}\mathrm{ss}^T + \boldsymbol{Z} \tag{1}$$

where we call (with an abuse of language) $\lambda \in \mathbb{R}_+$ the signal-to-noise-ratio (SNR), and the noise matrix $\boldsymbol{Z}$ is a symmetric matrix with i.i.d. $\mathcal{N}(0,1)$ off-diagonal and $\mathcal{N}(0,2)$ diagonal entries. This model is called the *Spiked-Wigner model*. The purpose of the scaling factor $\frac{1}{\sqrt{n}}$ is to make the inference problem neither trivially easy nor completely impossible in the large system limit.

The statistician is aware that the channel is additive Gaussian and that the true prior is a centered Gaussian, but he does not know the values λ and σ. He assumes values λ' and σ' as the SNR and the prior variance. Following the Bayesian estimation principle, he chooses the posterior mean as the estimate of the ground-truth. Our goal is to compute the asymptotic MSE for this mismatched estimation problem. Define the *mismatched* matrix-MSE as

$$\mathrm{MSE}_n(\sigma, \sigma', \lambda, \lambda') := \frac{1}{n^2}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\|\mathbf{s}\mathbf{s}^T - \langle \boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda',\sigma'}\big\|_F^2\Big]$$

where $\|.\|_F$ is the Frobenius norm, and $\langle . \rangle_{\lambda',\sigma'}$ denotes the expectation with respect to the posterior distribution from the statistician's point of view, that the SNR is λ' and $\boldsymbol{x} \sim \mathbf{P} = \mathcal{N}(0, \sigma'^2)$. Here we adopt the traditional statistical mechanics notation for the internal (annealed) expectations

$$\langle f(\boldsymbol{x})\rangle_{\lambda',\sigma'} = \frac{\int d\boldsymbol{x}\, \mathbf{P}(\boldsymbol{x}) f(\boldsymbol{x}) e^{-\frac{1}{4}\|\sqrt{\frac{\lambda}{n}}\mathbf{s}\mathbf{s}^T + \boldsymbol{Z} - \sqrt{\frac{\lambda'}{n}}\boldsymbol{x}\boldsymbol{x}^T\|_F^2}}{\int d\boldsymbol{x}\, \mathbf{P}(\boldsymbol{x}) e^{-\frac{1}{4}\|\sqrt{\frac{\lambda}{n}}\mathbf{s}\mathbf{s}^T + \boldsymbol{Z} - \sqrt{\frac{\lambda'}{n}}\boldsymbol{x}\boldsymbol{x}^T\|_F^2}}$$

for any reasonable function $f(\boldsymbol{x})$ such that the integrals are finite.

Note that, when we are in the matched (Bayes optimal) case $\lambda' = \lambda$, $\sigma' = \sigma$, the best achievable error is the matrix-MMSE which is defined as

$$\mathrm{MMSE}_n(\sigma, \lambda) := \frac{1}{n^2}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\|\mathbf{s}\mathbf{s}^T - \langle \boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda,\sigma}\big\|_F^2\Big]$$

We necessarily have $\mathrm{MSE}_n \geq \mathrm{MMSE}_n$.

*Notation:*We often drop the n subscript to denote the asymptotic large n limit. We may also drop the parameter dependency for notational simplicity.

III. Main Result

The main result is the following:

Theorem 1. *Assume that the sequence* $(\mathrm{MSE})_{n\geq 1}$ *converges uniformly in* $(\lambda, \lambda') \in K \subset \mathbb{R}_+^2$*. Then, for all* σ, σ' *(strictly positive) and* $(\lambda, \lambda') \in K$*, the asymptotic mismatched MSE is given by eq. (2).*

Remark 1. *In the matched case, uniform convergence of the sequence* $(\mathrm{MMSE})_{n\geq 1}$ *- except possibly at phase transition points which form a set of measure zero - follows using the concavity of mutual information with respect to* λ*. Then, using the I-MMSE relation [13], this allows to interchange limit and derivative to go from asymptotic mutual information (a.k.a. free energy) to asymptotic MMSE. For the present mismatched MSE, we use a relation similar to I-MMSE but in terms of mismatched free energies, which lack concavity w.r.t.* λ *and* λ'*. Therefore almost everywhere uniform convergence is difficult to establish from general principles. However, we conjecture that it holds and that eq. (2) holds almost everywhere (i.e., except possibly at phase transition lines).*

Remark 2. *One can see that the normalized MSE, i.e.* $\sigma^{-4}\mathrm{MSE}_n$*, can be expressed as a function of the three dimensionless variables* $\lambda\sigma^4$*,* $\lambda'\sigma^4$*,* $\frac{\sigma^2}{\sigma'^2}$*. This allows us to study the problem for the case* $\sigma^2 = 1$ *and easily generalize the analysis to other cases by rescaling the parameters.*

The MSE is illustrated for the case of $\sigma = 1, \lambda = 2$ in Fig. 1. The observed behavior is generic for $\lambda\sigma^4 > 1$. We observe one phase transition line and an intermediate region where estimation better than chance is possible, in the sense that the MSE is smaller than σ^4. We refer to the caption of Fig. 1 for details. In the case $\sigma = 1$ and $\lambda < 1$, or more generally $\lambda\sigma^4 < 1$, it is easy to see from Eq. (2) that the intermediate region disappears and the MSE is always greater or equal to σ^4 (the phase transition line is still present technically speaking).

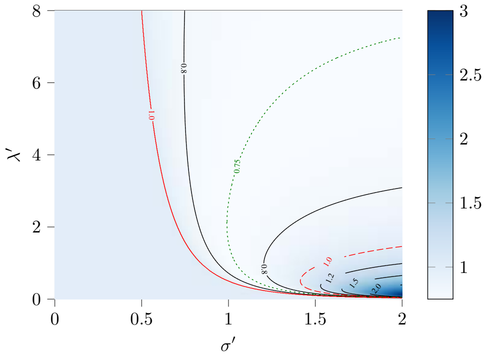


Fig. 1: Plot of MSE according to Eq. (2) for $\sigma = 1, \lambda = 2$. The solid leftmost (red) curve is a *phase transition* line (the MSE is continuous but the derivative is discontinuous). On the left of this curve $\mathrm{MSE} = \sigma^4 = 1$. In the *intermediate region* between the solid leftmost (red) curve and the dashed (red) curve the MSE takes values less than $\sigma^4 = 1$. In this intermediate region estimation better than chance is possible. On the dotted (green) curve the MSE attains the $\mathrm{MMSE}(\sigma, \lambda) = \frac{2}{\lambda} - \frac{1}{\lambda^2\sigma^4} = 0.75$ (even though we do not have $\lambda' = \lambda$, $\sigma' = \sigma$ except for one point with a vertical tangent on the curve). The MSE equals $\sigma^4 = 1$ on the dashed (red) line and takes higher values in the region on the right hand side of this line. Note that this is *not* a phase transition line, and the MSE is a perfectly analytic function there. The analytical expressions of the phase transition line, as well as dotted and dashed lines can easily be written down from eqs. (2) and (3). For $\sigma = 1, \lambda = 2$ the dotted (resp. dashed) curves have horizontal asymptotes $\lambda' = 8$ (resp. $\lambda' = 2$). See also figures 2 and 3 in [14].

$$\lim_{n\to\infty} \mathrm{MSE}_n(\sigma, \sigma', \lambda, \lambda') = \begin{cases} \sigma^4 + \big(\frac{1}{\sqrt{\lambda'}} - \frac{1}{\lambda'\sigma'^2}\big)^2 & \text{if } \lambda\sigma^4 \leq 1, \text{ and } \lambda'\sigma'^4 \geq 1 \\ \sigma^4(1 - \sqrt{\frac{\lambda}{\lambda'}})^2 + \frac{2}{\sqrt{\lambda\lambda'}} + \frac{1}{\lambda'^2\sigma'^4} + \frac{2}{\lambda'}\frac{\sigma^2}{\sigma'^2}(1 - \sqrt{\frac{\lambda}{\lambda'}}) - \frac{2}{\lambda\lambda'\sigma^2\sigma'^2} & \text{if } \lambda\sigma^4 \geq 1, \text{ and } \sqrt{\lambda\lambda'} \geq \frac{1}{\sigma^2\sigma'^2} \\ \sigma^4 & \text{if o.w.} \end{cases} \tag{2}$$

A. Inference with Matched SNR

Suppose that the statistician fully knows the channel and can choose $\lambda' = \lambda$. The asymptotic mismatched MSE in the limit $n \to \infty$ is:

$$\text{if } \sigma' \leq \sigma, \text{MSE} = \begin{cases} \sigma^4 & \text{if } \lambda \leq \frac{1}{\sigma^2\sigma'^2} \\ \frac{2}{\lambda} - \frac{1}{\lambda^2\sigma'^2}(\frac{2}{\sigma^2} - \frac{1}{\sigma'^2}) & \text{if } \lambda \geq \frac{1}{\sigma^2\sigma'^2} \end{cases}$$

$$\text{if } \sigma' \geq \sigma, \text{MSE} = \begin{cases} \sigma^4 & \text{if } \lambda \leq \frac{1}{\sigma'^4} \\ \frac{2}{\lambda} - \frac{1}{\lambda^2\sigma'^2}(\frac{2}{\sigma^2} - \frac{1}{\sigma'^2}) & \text{if } \lambda \geq \frac{1}{\sigma^4} \\ \sigma^4 + \frac{1}{\lambda} - \frac{1}{\lambda^{\frac{3}{2}}\sigma'^2}(2 - \frac{1}{\sqrt{\lambda}\sigma'^2}) & \text{o.w.} \end{cases}$$

For $\sigma = 1$ the MSE is plotted as a function of SNR for various values of σ' in Fig. 2. When $\sigma' > \sigma$, we observe that the MSE increases as the SNR increases (a similar behavior occurs in Fig. 1 in [10] for the scalar case). Although this happens when we are still in the regime of small SNR and estimation is impossible, we find this behavior rather counterintuitive.

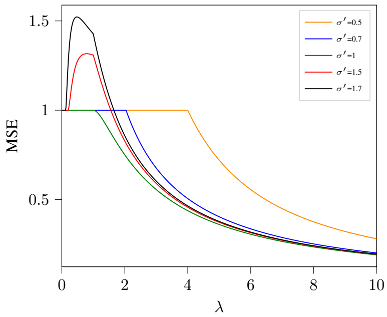


Fig. 2: Behavior of the MSE for matched SNR $\lambda' = \lambda$.

Remark 3. *For $\sigma' = \sigma, \lambda' = \lambda$, we are in the Bayes optimal setting and we recover the minimum MSE (MMSE) in the limit $n \to \infty$.*

$$\text{MMSE} = \begin{cases} \sigma^4 & \text{if} \lambda \leq \frac{1}{\sigma^4} \\ \frac{2}{\lambda} - \frac{1}{\lambda^2\sigma^4} & \text{if} \lambda \geq \frac{1}{\sigma^4} \end{cases}$$

This expression is well known and derived previously by a host of different approaches (see [1], [2], [6]–[8]).

As a sanity check of our result for the matched SNR case, with a bit of work we can check explicitly that

$$\begin{aligned} \int_0^\infty [\text{MSE}(\sigma, \sigma', \lambda, \lambda) - &\text{MMSE}(\sigma, \lambda)]\, d\lambda \\ &= 4D_{KL}(\mathcal{N}(0, \sigma^2), \mathcal{N}(0, \sigma'^2)) \end{aligned} \tag{3}$$

where D_{KL} denotes the Kullback-Leibler divergence. This sum-rule for vector channels is derived in [10] (with a factor of 2 instead of 4 in the vector case).

IV. Analysis

A. Mismatched free Energy and MSE

From the statistician's point of view, the posterior distribution reads up to a normalizing factor

$$\begin{aligned} \mathbf{P}\{\boldsymbol{x}|\boldsymbol{Y}\} &\propto e^{-\frac{1}{4}\|\boldsymbol{Y} - \sqrt{\frac{\lambda'}{n}}\boldsymbol{x}\boldsymbol{x}^T\|_F^2}\mathbf{P}(\boldsymbol{x}) \\ &\propto e^{-\frac{\lambda'}{4n}\|\boldsymbol{x}\|^4 + \frac{1}{2}\sqrt{\frac{\lambda'}{n}}\operatorname{Tr}\boldsymbol{Y}\boldsymbol{x}\boldsymbol{x}^T}\mathbf{P}(\boldsymbol{x}) \end{aligned} \tag{4}$$

where $\mathbf{P}$ is the normal distribution with iid entries and variance σ'. In deriving the second line, we use the fact that $\|\boldsymbol{Y}\|_F$ is a constant (because it is being conditioned on). Note that, $\boldsymbol{Y}$ is symmetric and the upper (or lower) part is distributed as $(Y_{i,j})_{i<j} \sim \mathcal{N}(\sqrt{\frac{\lambda}{n}}s_i s_j, 1)$, and the diagonal $(Y_{i,i}) \sim \mathcal{N}(\sqrt{\frac{\lambda}{n}}s_i s_i, 2)$.

The *partition function* is defined as the normalization factor of the last expression

$$Z(\boldsymbol{Y}) = \int d\boldsymbol{x}\, e^{-\frac{\lambda'}{4n}\|\boldsymbol{x}\|^4 + \frac{1}{2}\sqrt{\frac{\lambda'}{n}}\operatorname{Tr}\boldsymbol{Y}\boldsymbol{x}\boldsymbol{x}^T}\mathbf{P}(\boldsymbol{x}) \tag{5}$$

and the *mismatched free energy* is defined as

$$f_n(\sigma, \sigma', \lambda, \lambda') = -\frac{1}{n}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}[\ln Z(\boldsymbol{Y})] \tag{6}$$

Now we state a lemma relating the mismatched free energy to MSE. This lemma does not require any assumption on priors, and holds as long as the noise is additive Gaussian. Keep in mind that both mismatched free energy and MSE are functions of $\sigma, \sigma', \lambda, \lambda'$, but for simplicity of notation, we drop the arguments.

Lemma 1.

$$\frac{d}{d\lambda'}f_n + (2 - \sqrt{\frac{\lambda}{\lambda'}})\sqrt{\frac{\lambda}{\lambda'}}\frac{d}{d\lambda}f_n + \frac{1}{4}\frac{1}{n^2}\mathbb{E}\big[\|\mathbf{s}\mathbf{s}^T\|_F^2\big] = \frac{1}{4}\text{MSE}_n \tag{7}$$

Remark 4. *Eq.* (7) *generalizes the classical I-MMSE relation. Here the mismatched free energy cannot be related to a mutual information. However, note that, in the special case where $\lambda' = \lambda$ Eq.* (7) *simplifies slightly and combining with the I-MMSE relation, we obtain that the difference of MSE and MMSE is directly related to a derivative of a relative entropy, equivalent to relations discussed in detail in [10] for vector channels.*

Proof of lemma. We have

$$\frac{d}{d\lambda}f_n = -\frac{1}{4}\frac{1}{n^2}\sqrt{\frac{\lambda'}{\lambda}}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\langle(\mathbf{s}^T\boldsymbol{x})^2\big\rangle_{\lambda', \sigma'}\Big]$$

and by using a standard Gaussian integration by parts trick,

$$\frac{d}{d\lambda'}f_n = \frac{1}{4}\frac{1}{n^2}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\|\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda', \sigma'}\big\|_F^2 - \sqrt{\frac{\lambda}{\lambda'}}\big\langle(\mathbf{s}^T\boldsymbol{x})^2\big\rangle_{\lambda', \sigma'}\Big]$$

Putting these two equations together, the left-hand side of eq. (7) is equal to

$$\begin{aligned} &\frac{1}{4}\frac{1}{n^2}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\|\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda', \sigma'}\big\|_F^2 - 2\big\langle(\mathbf{s}^T\boldsymbol{x})^2\big\rangle_{\lambda', \sigma'} + \|\mathbf{s}\|^4\Big] \\ &= \frac{1}{4}\frac{1}{n^2}\mathbb{E}_{\mathbf{P}^*, \mathbf{P}_{\boldsymbol{Z}}}\Big[\big\|\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda', \sigma'}\big\|_F^2 - 2\operatorname{Tr}\mathbf{s}\mathbf{s}^T\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{\lambda', \sigma'} \\ &\qquad + \|\mathbf{s}\mathbf{s}^T\|_F^2\Big] = \frac{1}{4}\text{MSE}_n \qquad \square \end{aligned}$$

Thus, the problem is reduced to computing the (mismatched) free energy. The main idea is to exploit the rotational invariance of the normal distribution that the statistician chooses.

Changing variables $\boldsymbol{x} \to \boldsymbol{U}\boldsymbol{x}$, for an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$, the integral in eq. (5) becomes ($|\mathrm{det}\boldsymbol{U}| = 1$):

$$
\begin{aligned}
Z(\boldsymbol{Y}) &= \int d\boldsymbol{x}\, e^{-\frac{\lambda'}{4n}\|\boldsymbol{U}\boldsymbol{x}\|^4+\frac{1}{2}\sqrt{\frac{\lambda'}{n}}\operatorname{Tr}\boldsymbol{Y}\boldsymbol{U}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}^T}\mathbf{P}(\boldsymbol{U}\boldsymbol{x}) \\
&= \int d\boldsymbol{x}\mathbf{P}(\boldsymbol{x})\, e^{-\frac{\lambda'}{4n}\|\boldsymbol{x}\|^4+\frac{1}{2}\sqrt{\frac{\lambda'}{n}}\operatorname{Tr}\boldsymbol{Y}\boldsymbol{U}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}^T}
\end{aligned}
$$

Since this holds for any orthogonal matrix $\boldsymbol{U}$, we can take the expectation over the *Haar* measure on the group of $n \times n$ orthogonal matrices.

$$
Z(\boldsymbol{Y}) = \int d\boldsymbol{x}\mathbf{P}(\boldsymbol{x})\, e^{\frac{-\lambda'}{4n}\|\boldsymbol{x}\|^4} \int D\boldsymbol{U} e^{\frac{1}{2}\sqrt{\frac{\lambda'}{n}}\operatorname{Tr}\boldsymbol{Y}\boldsymbol{U}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}^T} \tag{8}
$$

where $D\boldsymbol{U}$ denotes the *Haar* measure.

In the next subsection, we will discuss computing the inner integral in eq. (8).

B. Spherical Integrals

The spherical integral is defined as:

$$
I_n(\boldsymbol{A},\boldsymbol{B}) = \int D\boldsymbol{U} e^{n\operatorname{Tr}\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^T} \tag{9}
$$

where $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{n\times n}$, and $D\boldsymbol{U}$ denotes the *Haar* measure over the orthogonal matrices. Note that, this definition can also be extended to the unitary matrices. In the mathematical physics literature, such integrals are often called *Harish Chandra-Itzykson-Zuber (HCIZ)* integrals. The interest for these objects dates to the work of the mathematician Harish Chandra [15], and they have been extensively studied and developed in physics and mathematics. In particular, [12] derived the asymptotics of spherical integrals when the rank of matrix $\boldsymbol{B}$ is $O(1)$ w.r.t n. We will apply this result for the rank-one $\boldsymbol{B}$ to our problem . For simplicity of notation, we denote the integral by $I_n(\theta,\boldsymbol{A})$, where θ is the only non-zero eigenvalue of $\boldsymbol{B}$.

From the definition (9), one may notice that the integral only depends on the eigenvalues of $\boldsymbol{A}$,$\boldsymbol{B}$. So, it is natural to expect that the asymptotic of the integral depends on the limiting spectral measure of the matrix $\boldsymbol{A}$. The result of [12] is based on the hypothesis that the spectral measure $\mu_{\boldsymbol{A}}$ converges weakly towards a compactly supported measure μ, and the minimum and maximum eigenvalues of $\boldsymbol{A}$ converge to the finite values γ_{min}, γ_{max}, respectively.

For a probability measure μ, the *Hilbert* (or *Stieltjes*) transform is the map $H_\mu : \mathbb{R}\backslash\mathrm{supp}(\mu) \to \mathbb{R}$, $H_\mu(z) = \int \frac{1}{z-t}\, d\mu(t)$. This map is invertible, and denoting its inverse by $H_\mu^{-1}(.)$, for z in range of H_μ we define the *R-transform* of a probability measure μ as $R_\mu(z) = H_\mu^{-1}(z) - \frac{1}{z}$.

Theorem 2 (Guionnet and Maïda [12])**.** *Suppose $\mu_{\boldsymbol{A}}$ converges weakly towards μ. Let $H_{min} = \lim_{z\to\gamma_{min}} H_\mu(z)$, $H_{max} = \lim_{z\to\gamma_{max}} H_\mu(z)$. Then:*

$$
\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\ln I_n(\theta,\boldsymbol{A}) \\
= \theta\nu(\theta) - \frac{1}{2}\int \ln(1+2\theta\nu(\theta)-2\theta t)\, d\mu(t)
\end{aligned}
$$

where

$$
\nu(\theta) = \begin{cases} R_\mu(2\theta) & \text{if } H_{min} \le 2\theta \le H_{max} \\ \gamma_{max} - \frac{1}{2\theta} & \text{if } 2\theta > H_{max} \\ \gamma_{min} - \frac{1}{2\theta} & \text{if } 2\theta < H_{min} \end{cases}
$$

C. Computing Free Energy

To apply the result from [12], we can rewrite the spherical integral in eq. (8) as

$$
I_n\Big(\frac{\sqrt{\lambda'}}{2n}\|\boldsymbol{x}\|^2, \frac{\boldsymbol{Y}}{\sqrt{n}}\Big) = \int D\boldsymbol{U} e^{n\operatorname{Tr}\frac{\boldsymbol{Y}}{\sqrt{n}}\boldsymbol{U}\frac{\sqrt{\lambda'}}{2n}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}^T} \tag{10}
$$

$\frac{\boldsymbol{Y}}{\sqrt{n}} = \frac{\sqrt{\lambda}}{n}\mathbf{s}\mathbf{s}^T + \frac{1}{\sqrt{n}}\boldsymbol{Z}$, where $\frac{1}{\sqrt{n}}\boldsymbol{Z}$ is the suitably normalized Wigner matrix whose limiting spectral measure is the renowned *semi-circle law* with density $\mu_{\text{SC}} = \frac{1}{2\pi}\sqrt{4-t^2}$. At the same time, the spectral measure of $\frac{\boldsymbol{Y}}{\sqrt{n}}$ converges almost surely (a.s) as $n \to \infty$ to the *semi-circle law* (see e.g. proposition 1 in [16]). We have $H_{\mu_{SC}}(z) = \frac{1}{2}(z - \sqrt{z^2-4})$ and $R_{\mu_{SC}}(z) = z$.

Let γ_{min} and γ_{max} be the smallest and the largest eigenvalues of $\frac{\boldsymbol{Y}}{\sqrt{n}}$, from the results in [17], as $n \to \infty$ we have (a.s.)

$$
\gamma_{\text{min}} = -2, \gamma_{\text{max}} = \begin{cases} 2 & \text{if } \sqrt{\lambda}\sigma^2 \le 1 \\ \sqrt{\lambda}\sigma^2 + \frac{1}{\sqrt{\lambda}\sigma^2} & \text{if } \sqrt{\lambda}\sigma^2 \ge 1 \end{cases}
$$

So,

$$
H_{\text{min}} = -1, H_{\text{max}} = \begin{cases} 1 & \text{if } \sqrt{\lambda}\sigma^2 \le 1 \\ \frac{1}{\sqrt{\lambda}\sigma^2} & \text{if } \sqrt{\lambda}\sigma^2 \ge 1 \end{cases}
$$

Theorem 3. *For all $\sigma, \sigma', \lambda, \lambda'$ positive, the asymptotic free energy of the mismatched inference model is given in eq. (11).*
Proof sketch. Eq. (8) can be written as

$$
Z(\boldsymbol{Y}) = \int d\boldsymbol{x}\mathbf{P}(\boldsymbol{x})\, e^{\frac{-\lambda'}{4n}\|\boldsymbol{x}\|^4+\ln I_n\left(\frac{\sqrt{\lambda'}}{2n}\|\boldsymbol{x}\|^2, \frac{\boldsymbol{Y}}{\sqrt{n}}\right)} \tag{12}
$$

Since $\mathbf{P}(\boldsymbol{x})$ is Gaussian, the integrand in (12) is a function of $\|\boldsymbol{x}\|$, so we can use spherical coordinates to reduce the integral in (12) to a one-dimensional integral

$$
\begin{aligned}
Z(\boldsymbol{Y}) &= \frac{2^{-\frac{n}{2}+1}}{\Gamma(\frac{n}{2})}\frac{1}{\sigma'^n} \\
&\quad\times \int_0^{+\infty} d\rho\, \rho^{n-1} e^{-\frac{\rho^2}{2\sigma'^2}-\frac{-\lambda'}{4n}\rho^4+\ln I_n\left(\frac{\sqrt{\lambda'}}{2n}\rho^2, \frac{\boldsymbol{Y}}{\sqrt{n}}\right)}
\end{aligned}
$$

where $\rho := \|\boldsymbol{x}\|$, and $\Gamma(.)$ is the *Gamma* function. Changing variable $\frac{\rho^2}{n} \to \rho$, we obtain

$$
\begin{aligned}
Z(\boldsymbol{Y}) &= \frac{2^{-\frac{n}{2}}n^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\frac{1}{\sigma'^n} \\
&\quad\times \int_0^{+\infty} \frac{d\rho}{\rho}\, e^{-n\left(\frac{\rho}{2\sigma'^2}-\frac{1}{2}\ln\rho+\frac{\lambda'}{4}\rho^2-J_n\left(\sqrt{\lambda'}\frac{\rho}{2}, \frac{\boldsymbol{Y}}{\sqrt{n}}\right)\right)}
\end{aligned} \tag{13}
$$

where $J_n\big(\theta, \frac{\boldsymbol{Y}}{\sqrt{n}}\big) \equiv \frac{1}{n}I_n\big(\theta, \frac{\boldsymbol{Y}}{\sqrt{n}}\big)$. By Theorem 2, $J_n\big(\theta, \frac{\boldsymbol{Y}}{\sqrt{n}}\big)$ converges to a deterministic function $J(\theta; \mu_{\text{SC}}, \gamma_{\text{max}})$.

We are interested in $\lim_{n\to\infty} f_n = \lim_{n\to\infty}\mathbb{E}\big[-\frac{1}{n}\ln Z(\boldsymbol{Y})\big]$. The prefactors in (13) are independent of $\boldsymbol{Y}$ and

$$\lim_{n\to\infty} f_n(\sigma,\sigma',\lambda,\lambda') = \begin{cases} -\frac{1}{4\lambda'\sigma'^4} + \frac{1}{\sqrt{\lambda'}\sigma'^2} - \frac{3}{4} + \ln \lambda'^{\frac{1}{4}}\sigma' & \text{if } \lambda\sigma^4 \leq 1, \text{ and } \lambda'\sigma'^4 \geq 1 \\ \frac{1}{2}\ln\sqrt{\lambda\lambda'}\sigma^2\sigma'^2 - \frac{1}{4\lambda'\sigma'^4} - \frac{\lambda\sigma^4}{4} + \sqrt{\frac{\lambda}{\lambda'}}\frac{\sigma^2}{2\sigma'^2} + \frac{1}{2\sqrt{\lambda\lambda'}\sigma^2\sigma'^2} - \frac{1}{2} & \text{if } \lambda\sigma^4 \geq 1, \text{ and } \sqrt{\lambda\lambda'} \geq \frac{1}{\sigma^2\sigma'^2} \\ 0 & \text{if o.w.} \end{cases} \tag{11}$$

the limit $\lim_{n\to\infty} -\frac{1}{n}\ln\{\text{prefactors}\}$ equals $-\frac{1}{2}+\ln\sigma'$. Next, we compute the asymptotic of the integral in (13), denoted from now on by $K(\boldsymbol{Y})$. Let us define the function

$$\psi(\rho) = \frac{\rho}{2\sigma'^2} - \frac{1}{2}\ln\rho + \frac{\lambda'}{4}\rho^2 - J(\sqrt{\lambda'}\frac{\rho}{2}; \mu_{\text{SC}}, \gamma_{\max}) \tag{14}$$

We can show that $\mathbb{E}\big[-\frac{1}{n}\ln K(\boldsymbol{Y})\big]$ is bounded above and below by $\min_\rho \psi(\rho) \pm o_n(1)$. Therefore, we get:

$$\lim_{n\to\infty} f_n = \min_\rho \psi(\rho) - \frac{1}{2} + \ln\sigma'$$

Solving this optimization problem, we find (11). □

Once we have the expression for the free energy, we can compute the MSE using Lemma 1. As explained in remark 1 this step uses the assumption that for $(\lambda,\lambda') \in K \subset \mathbb{R}_+^2$ the sequence $(\text{MSE})_{n\geq 1}$ converges uniformly.

V. Conclusion

Studying inference problems in settings where priors and hyper-parameters are unknown or partially known and deriving fundamental limits of estimation is a problem with practical importance. We derived analytical formulas for asymptotic MSE in estimating a rank-one matrix corrupted by additive Gaussian noise when both the channel and prior are partially known. In this short note, we have shown how to treat one of the most straightforward such situations by using beautiful asymptotic formulas of spherical integrals. The major limitation of our technique is that the statistician assumes a spherically invariant prior. This can be a Gaussian which has the advantage of being factorized, but we can also treat a uniform distribution over a sphere. Given such distributions for the statistician, it is then possible to extend our analysis to a broader class of problems, namely:

- Estimation of finite rank and rectangular matrices can be accomodated (i.e., $\text{rank} = O(1)$ w.r.t $n \to +\infty$).
- The true prior does not need to be rotation invariant. General factorized priors can be accommodated, for example, a Rademacher-Bernoulli mixture modeling sparse signals.
- A temperature parameter can be introduced by the statistician in his mismatched posterior distribution (with minor modifications in the analysis).

These extensions result in a very rich phenomenology with many possible phase transitions. Already in the simplest situation considered here, the MSE displays non-trivial features. Other problems of interest are the construction of more general estimators (non-Bayesian or non-Gibbsian), which can still be analyzed through spherical integrals, as well as comparing the analytical expressions of the MSE to algorithmic predictions, for example, those based on AMP [18], Approximate Survey Propagation [19], and spectral methods applied to mismatched situations.

Acknowledgment

The work of F. P is supported by the SNSF grant no 200021E 175541. N.M is thankful to Jean Barbier and Emanuele Mingione for numerous discussions.

References

[1] Y. Deshpande and A. Montanari, “Information-theoretically optimal sparse pca,” in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 2197–2201.

[2] Y. Deshpande, E. Abbe, and A. Montanari, “Asymptotic mutual information for the two-groups stochastic block model,” *arXiv preprint arXiv:1507.08685*, 2015.

[3] T. Lesieur, F. Krzakala, and L. Zdeborová, “MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel,” in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 680–687.

[4] S. B. Korada and N. Macris, “Exact solution of the gauge symmetric p-spin glass model on a complete graph,” *Journal of Statistical Physics*, vol. 136, no. 2, pp. 205–230, 2009.

[5] F. Guerra and F. L. Toninelli, “Quadratic replica coupling in the sherrington–kirkpatrick mean field spin glass model,” *Journal of Mathematical Physics*, vol. 43, no. 7, pp. 3704–3716, 2002.

[6] M. Dia, N. Macris, F. Krzakala, T. Lesieur, L. Zdeborová *et al.*, “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 424–432, 2016.

[7] M. Lelarge and L. Miolane, “Fundamental limits of symmetric low-rank matrix estimation,” *Probability Theory and Related Fields*, vol. 173, no. 3, pp. 859–929, 2019.

[8] J. Barbier and N. Macris, “The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference,” *Probability theory and related fields*, vol. 174, no. 3, pp. 1133–1185, 2019.

[9] ——, “The adaptive interpolation method for proving replica formulas. applications to the curie–weiss and wigner spike models,” *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 29, p. 294002, 2019.

[10] S. Verdú, “Mismatched estimation and relative entropy,” *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, 2010.

[11] T. Weissman, “The relationship between causal and noncausal mismatched estimation in continuous-time awgn channels,” *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4256–4273, 2010.

[12] A. Guionnet and M. Maïda, “A fourier view on the R-transform and related asymptotics of spherical integrals,” *Journal of functional analysis*, vol. 222, no. 2, pp. 435–490, 2005.

[13] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in gaussian channels,” *IEEE transactions on information theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[14] F. Pourkamali and N. Macris, “Mismatched estimation of rank-one symmetric matrices under gaussian noise,” *arXiv preprint arXiv:2107.08927*, 2021.

[15] Harish-Chandra, “Differential operators on a semisimple lie algebra,” *American Journal of Mathematics*, pp. 87–120, 1957.

[16] M. Capitaine and C. Donati-Martin, “Spectrum of deformed random matrices and free probability,” *arXiv preprint arXiv:1607.05560*, 2016.

[17] F. Benaych-Georges and R. R. Nadakuditi, “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices,” *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011.

[18] S. Rangan and A. K. Fletcher, “Iterative estimation of constrained rank-one matrices in noise,” in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 1246–1250.

[19] F. Antenucci, F. Krzakala, P. Urbani, and L. Zdeborová, “Approximate survey propagation for statistical inference,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 2, p. 023401, 2019.

Author Index