Polarization of the Rényi Information Dimension with Applications to Compressed Sensing

Saeid Haghighatshoar, *Member, IEEE,* Emmanuel Abbe, *Member, IEEE*

***Abstract*—In this paper, we show that the Hadamard matrix acts as an extractor over the reals of the Rényi information dimension (RID), in an analogous way to how it acts as an extractor of the discrete entropy over finite fields. More precisely, we prove that the RID of an i.i.d. sequence of mixture random variables polarizes to the extremal values of 0 and 1 (corresponding to discrete and continuous distributions) when transformed by a Hadamard matrix. Further, we prove that the polarization pattern of the RID admits a closed form expression and follows exactly the *Binary Erasure Channel* (BEC) polarization pattern in the discrete setting. We also extend the results from the single- to the multi-terminal setting, obtaining a Slepian-Wolf counterpart of the RID polarization. We discuss applications of the RID polarization to Compressed Sensing of i.i.d. sources. In particular, we use the RID polarization to construct a family of deterministic ± 1-valued sensing matrices for Compressed Sensing. We run numerical simulations to compare the performance of the resulting matrices with that of random Gaussian and random Hadamard matrices. The results indicate that the proposed matrices afford competitive performances while being explicitly constructed.**

***Index Terms*—Rényi Information Dimension, Polarization Theory, Slepian-Wolf coding, Compressed Sensing.**

I. Introduction

Let X be a real-valued random variable. We denote the q-ary quantization of X by $\langle X \rangle_q = \frac{\lfloor qX \rfloor}{q}$, where for a real number r, we denote by $\lfloor r \rfloor$ the largest integer less than or equal to r. The upper and the lower *Rényi Information Dimension* (RID) of X are defined by

$$\overline{d}(X) = \limsup_{q \to \infty} \frac{H(\langle X \rangle_q)}{\log_2(q)}, \tag{1}$$

$$\underline{d}(X) = \liminf_{q \to \infty} \frac{H(\langle X \rangle_q)}{\log_2(q)}, \tag{2}$$

where $H(\langle X \rangle_q)$ denotes the Shannon entropy of the discrete random variable $\langle X \rangle_q$ obtained from the quantization. If the limits coincide, we define $d(X) := \overline{d}(X) = \underline{d}(X)$. Rényi in his paper [3] proved that if the random variable X is discrete, continuos, or a mixture thereof, the upper and the lower RID are equal, thus, $d(X)$ is well-defined. He also provided an example of a singular random variable for which these two limits do not coincide. Apart from being an information measure, the RID appears as the fundamental operational limit in diverse areas in probability theory and signal processing such as signal quantization [4], rate-distortion theory [5], and fractal geometry [6]. More recently, the operational aspect of RID has reappeared in applications as varied as lossless analog compression [7–9], Compressed Sensing of sparse signals [10–12], and characterization of the degrees-of-freedom (DoFs) of vector interference channels [13, 14], which has recently been of significant importance in wireless communication.

In this paper, motivated by [3], we first extend the definition of RID as an information measure from scalar random variables to a family of vector random variables over which the RID is well-defined. We also extend the definition to the joint and the conditional RIDs and provide a closed-form expression for computing them. Using these, we investigate the high-dimensional behavior of the RID of i.i.d. mixture random variables when transformed by a Hadamard matrix. We prove that the conditional RIDs of almost all the resulting random variables polarize to the extremal values of 0 and 1. We also obtain a formula for computing those conditional RIDs and their polarization pattern using the *Binary Erasure Channel* (BEC) polarization in the discrete case [15]. This gives a natural extension of the polarization phenomenon for the entropy over the finite fields to the RID over the reals.

We study some of the potential applications of the new polarization result in Compressed Sensing [16–19]. In particular, motived by the recent results on the operational aspect of RID in Compressed Sensing [7, 10] and inspired by the success of polar codes in achieving information theoretic limits [15, 20], we exploit the RID polarization to design deterministic partial Hadamard matrices for Compressed Sensing of i.i.d. sparse signals. We compare the performance of the resulting matrices with that of other traditional matrices in Compressed Sensing such as random Gaussian and random Hadamard matrices. Numerical simulations provide evidence that the constructed matrices together with recovery algorithms such as l_1-norm minimization provide a low-complexity Compressed Sensing and recovery procedure for the sparse signals. The use of polarization techniques for Compressed Sensing was also investigated independently in [11], approaching noiseless Compressed Sensing via a duality with analog channel coding.

A. Notation

We use $\mathbb{R}$ for the reals, $\mathbb{R}_+$ for the positive reals, $\mathbb{Z}$ for the integers, $\mathbb{Z}_+$ for the set of positive integers, and $\mathbb{N}$ for the set of strictly positive integers. We denote sets by calligraphic letters such as $\mathcal{A}$ and their cardinality by $|\mathcal{A}|$. We use capital letters for random variables and small letters for their realizations, e.g., x is a realization of the random variable X. We denote

This paper is the updated version of the paper [1], which was partially presented in ISIT 2013, Istanbul, Turkey [2].

Saeid Haghighatshoar is with the Communications and Information Theory Group, Technische Universität Berlin (saeid.haghighatshoar@tu-berlin.de). Emmanuel Abbe has a joint position in Applied and Computational Mathematics and Electrical Engineering at Princeton University, New Jersey, USA (eabbe@princeton.edu). This work was started when both the authors were with the Information Processing Group (IPG), EPFL, Switzerland.

the distribution of a random variable X by p_X. For $N = 2^n$, we denote by $\mathbf{H}_N$ the standard Hadamard matrix of order N. We use $[n]$ for the set of integers $\{1, 2, \ldots, n\}$. We denote by X_i^j the column vector $[X_i, X_{i+1}, \ldots, X_j]^\mathsf{T}$, where the vector is empty when $i > j$. Vectors are denoted by boldface letters, e.g., $\mathbf{X} = X_1^n$ and $\mathbf{x} = x_1^n$ denote an n-dim vector of random variables and its realization. For two sequences $f, h : \mathbb{N} \to \mathbb{R}$, we say $f_q \doteq h_q$ if and only if

$$\lim_{q\to\infty} \frac{f_q - h_q}{\log_2(q)} = 0. \tag{3}$$

We denote matrices by capital letters, e.g., A. For an $m \times n$ matrix A, and a subset of its columns $\mathcal{C} \subseteq [n]$, we denote by $\mathsf{A}_\mathcal{C}$ the $m \times |\mathcal{C}|$ submatrix of A obtained by selecting those columns of A belonging to $\mathcal{C}$. In a similar way, we denote by $\mathsf{A}[\mathcal{R}]$ the $|\mathcal{R}| \times n$ submatrix obtained by selecting the rows of A belonging to $\mathcal{R} \subseteq [m]$. For two matrices A and B, we denote by $[\mathsf{A}; \mathsf{B}]$ a matrix obtained by putting the rows of A on top of the rows of B and by $[\mathsf{A}, \mathsf{B}]$ the matrix obtained by putting the columns of B to the right of the columns of A, provided that the resulting matrices are well-defined.

B. Reminder on Polar Codes

Polar codes were introduced by Arikan in his seminal paper [15]. They are the first class of efficient codes that provably achieve channel capacity on all binary input symmetric channels. Recent research on polar codes has illustrated their theoretical optimality for other classical problems in information theory such as lossless and lossy source coding [20, 21], coding over multiple-access channels (MAC) [22], Wyner-Ziv and Gelfand-Pinsker problem [23], and coding for secrecy over the wiretap channel [24, 25].

The underlying structure behind all these applications of polar codes is the polarization phenomenon. To explain briefly, we will mainly focus on the source coding aspect, which is more relevant to our work. Let $\mathbf{X} = X_1^N$, for $N = 2^n$ a power of two, be a sequence of N i.i.d. Bernoulli(p), $p \in (0, \frac{1}{2})$, random variables and let $\mathbf{Y} = \mathbf{G}_N\mathbf{X}$, with the arithmetic over the binary field $\mathbb{F}_2$, where

$$\mathbf{G}_N = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{\otimes n}.$$

The polarization phenomenon states that after applying this linear transformation, every element of the set of conditional entropies $\{H(Y_i|Y_1^{i-1})\}_{i=1}^N$ tends to be either very close to 0 (fully deterministic) or very close to 1 (fully random). Moreover, the fraction of those fully random (informative) variables turns out to be equal to the entropy of the source $h_2(p)$ asymptotically as $N \to \infty$. This allows to build optimal linear source encoders achieving the fundamental information theoretic limit by simply keeping only those rows of $\mathbf{G}_N$ corresponding to the fully random variables.

Polar codes have been applied to other problems in communication theory such as as multi-level lattice coding [26], and designing capacity achieving codes over the AWGN channel [27], mainly by extending the finite alphabet results. What is less understood is the similar counterpart of the polarization phenomenon for infinite-alphabet sources. In [12], using a new *Entropy Power Inequality* (EPI) for integer-valued random variables [28], a novel polarization result was proved for integer-valued sources under the conventional arithmetic over $\mathbb{Z}$. The result was used to construct deterministic partial Hadamard matrices for almost lossless encoding of integer-valued signals with a vanishing measurement rate of $o(N)$ for large block-lengths N. The more general case of real-valued sources, however, was left open in [12]. In this paper, we extend this result to real-valued sources and provide a polarization theory for infinite-alphabet signals.

II. Rényi information dimension

Let X be a random variable with a probability distribution p_X over $\mathbb{R}$. The upper and the lower RID of this random variable were defined in (1) and (2) respectively. Let $m \in \mathbb{N}$ and suppose $X \in [0, 1]$ almost surely. It is not difficult to see that if $X = 0.X_1X_2\ldots$ is the m-ary expansion of the random variable X with $X_i \in \{0, 1, \ldots, m-1\}$, then for $q = m^k$, we have $H(\langle X\rangle_q) = H(X_1, X_2, \ldots, X_k)$, where H denotes the discrete entropy in basis m. From (1) and (2), we have

$$\underline{d}(X) \leqslant H_\infty(\{X_i\}_{i=1}^\infty) \leqslant H^\infty(\{X_i\}_{i=1}^\infty) \leqslant \overline{d}(X), \tag{4}$$

where $H_\infty = \liminf_{k\to\infty} \frac{H(X_1,X_2,\ldots,X_k)}{k}$ denotes the lower entropy rate of the stochastic process $\{X_i\}_{i=1}^\infty$ (with a similar expression for H^∞ by replacing $\liminf$ with $\limsup$). As a special case, when X is uniformly distributed over $[0, 1]$, the random variables $\{X_i\}_{i=1}^\infty$ are i.i.d. each having a uniform distribution over $\{0, 1, \ldots, m-1\}$. Thus, the upper and lower RID are equal to $\overline{d}(X) = \underline{d}(X) = 1$. Also, $\overline{d}(X) = \underline{d}(X) = 0$ for any discrete random variable X with $H(X) < \infty$.

By Lebesgue decomposition theorem [29], any probability distribution p_X over $\mathbb{R}$ can be written as a convex combination of a continuous part p_c, a singular part p_s, and a discrete part p_d (with the latter two being singular with respect to Lebesgue measure) as follows

$$p_X = \alpha_c\, p_c + \alpha_s\, p_s + \alpha_d\, p_d, \tag{5}$$

where $\alpha_c, \alpha_s, \alpha_d \geqslant 0$ and $\alpha_c + \alpha_s + \alpha_d = 1$. In this paper, we only consider the case $\alpha_s = 0$, where p_X is the mixture of a continuous and a discrete distribution. In [3], Rényi showed that for such a mixture distribution, the RID is well-defined and is given by the weight of the continuous part α_c. In particular, it is 1 for the continuous and 0 for the discrete distributions. He also defined the RID of a continuous vector random variable X_1^n of dimension n, where he proved that

$$d(X_1^n) = \lim_{q\to\infty} \frac{H(\langle X_1^n\rangle_q)}{\log_2(q)} = n, \tag{6}$$

where the quantization is done component-wise, i.e.,

$$\langle X_1^n\rangle_q = (\langle X_1\rangle_q, \ldots, \langle X_n\rangle_q). \tag{7}$$

III. Summary of the Results

In this section, we briefly explain the results proved in our paper. Let Z_1 and Z_2 be two i.i.d. nonsingular random variables with a mixture distribution $p_Z(z) = (1-\delta)p_d(z) + \delta p_c(z)$, where p_d and p_c denote the discrete and the continuous

part of p_Z. Note that $d(Z_1) = d(Z_2) = \delta$, as explained in Section II. Let us consider $X_1^2 = \mathbf{H}_2 Z_1^2$ where

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

denotes the 2×2 Hadamard matrix. A direct calculation shows that $X_1 = Z_1 + Z_2$ has the following distribution

$$\begin{aligned} p_{X_1}(x) = p_Z(x) \star p_Z(x) = (1-\delta)^2 p_d \star p_d(x) \\ + 2\delta(1-\delta) p_d \star p_c(x) + \delta^2 p_c \star p_c(x), \end{aligned} \tag{8}$$

where $\star$ denotes the convolution operator over $\mathbb{R}$. From (8), it is seen that p_{X_1} is a mixture distribution with a discrete part $(1-\delta)^2 p_d \star p_d$, and has the RID $d(X_1) = 1-(1-\delta)^2 = 2\delta - \delta^2$.

Now let us consider the conditional distribution of X_2 given X_1 denoted by $p_{X_2|X_1}(x)$. From standard results in probability theory [30], this conditional distribution is a well-defined mixture distribution for almost all realizations of X_1. Hence, the conditional RID of X_2 given X_1 denoted by $d(X_2|X_1 = x_1) \in [0,1]$ is well-defined almost surely and is a function of X_1. Define the conditional RID of X_2 given X_1 as $d(X_2|X_1) = \mathbb{E}_{X_1}[d(X_2|X_1 = x_1)]$, provided that $d(X_1|X_2 = x_1)$ is a random variable (i.e., a measurable function of x_1) with a well-defined expected value. In Section IV, we develop techniques to compute $d(X_2|X_1)$ for a large class of mixture distributions in a closed form, where in particular we prove that such a conditional RID is well-defined. We also extend those techniques to calculate the joint (e.g., $d(X_1, X_2)$), conditional (e.g., $d(X_2|X_1)$), and the mutual RID of mixture vector-valued random variables. As a result, we obtain that $d(X_2|X_1) = \delta^2$ and that

$$d(X_1) + d(X_2|X_1) = (2\delta - \delta^2) + \delta^2 = 2\delta = 2d(Z_1), \tag{9}$$

which is analogous to the chain rule for the mutual information [31]. In fact, we prove that such a chain rule holds for the RID and satisfies most of the properties of the traditional chain rule for the mutual information [31]. In particular,

$$d(X_1) + d(X_2|X_1) = d(X_1, X_2) \overset{(i)}{=} d(Z_1, Z_2) \overset{(ii)}{=} 2d(Z_1), \tag{10}$$

where (i) follows from the fact that $\mathbf{H}_2$ is an invertible matrix, and where (ii) is due to the fact that Z_1 and Z_2 are i.i.d.

It is, thus, seen that multiplying two i.i.d. random variables Z_1, Z_2, with a mixture distribution, by $\mathbf{H}_2$ modifies their conditional RIDs according to $(\delta, \delta) \mapsto (2\delta - \delta^2, \delta^2)$. This resembles the polarization of a BEC channel with a capacity $\delta \in (0,1)$ as in [15]. In Section V, we prove that such a polarization indeed occurs for the RID. To be more precise, let $\{Z_i : i \in [N]\}$ be a sequence of i.i.d. nonsingular random variables with an RID δ and let $X_1^N = \mathbf{H}_N Z_1^N$, where $N = 2^n$ is a power of two and where $\mathbf{H}_N = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes n}$ denotes the Hadamard matrix of order N. We prove that the sequence of conditional RIDs

$$\{d(X_i|X_1^{i-1}) : i \in [N]\},$$

for increasing values of $N = 2^n$, polarizes according to the polarization pattern of a BEC with a channel capacity $\delta \in (0,1)$. We further investigate the applications of the established polarization result in Section VII.

IV. Generalization of the RID

A. Space of Random Variables and Generalized RID

Our objective is to extend the definition of RID to vector-valued random variables, which are not necessarily continuous. Let $\mathcal{I}$ be a collection of independent and nonsingular random variables (with $\alpha_s = 0$ as in (5)). We define the space $\mathcal{L}$ of random variables generated by $\mathcal{I}$ as $\mathcal{L} = \cup_{n=1}^{\infty} \mathcal{L}_n$, where

$$\begin{aligned} \mathcal{L}_n = \{X_1^n : \exists\, k, \mathsf{A} \in \mathbb{R}^{n\times k}, Z_i \in \mathcal{I} \text{ for } i \in [k], \\ \text{such that } X_1^n = \mathsf{A} Z_1^k\}. \end{aligned} \tag{11}$$

It is seen that $\mathcal{L}_n$ consists of all n-dim random vectors generated by a linear mixture of *finitely many* elements of $\mathcal{I}$[1]. Note that $\mathcal{L}$ is stable under vector addition and concatenation, i.e., for arbitrary $W_1^n, X_1^n \in \mathcal{L}_n$ and $Y_1^m \in \mathcal{L}_m$, we have that $W_1^n + X_1^n \in \mathcal{L}_n$, and $[X_1^n; Y_1^m] \in \mathcal{L}_{n+m}$. Moreover, $\mathcal{L}$ is stable under an arbitrary linear transformation, i.e., $\psi(\mathcal{L}_n) \subseteq \mathcal{L}_m$ for any linear map $\psi : \mathbb{R}^n \to \mathbb{R}^m$.

We define the joint and the conditional RID, and the *Mutual Rényi Information* for the random variables in $\mathcal{L}$ by

$$d(X_1^n) = \lim_{q\to\infty} \frac{H(\langle X_1^n \rangle_q)}{\log_2(q)} \tag{12}$$

$$d(X_1^n|Y_1^m) = \lim_{q\to\infty} \frac{H(\langle X_1^n \rangle_q | Y_1^m)}{\log_2(q)} \tag{13}$$

$$\mathsf{RI}(X_1^n; Y_1^m) = d(X_1^n) - d(X_1^n|Y_1^m). \tag{14}$$

We will prove that all the limits above are well-defined. In general, computing the RID for a given multi-variate distribution is quite challenging since the distribution might contain a probability mass over complicated subsets or sub-manifolds of lower dimensions. Moreover, the limit might not even exist in some cases. Fortunately, using the linear structure in $\mathcal{L}$, we are able to obtain a simple formula for computing the RID via the rank characterization. A similar rank characterization was used in the context of finite fields for coding over the BEC and the BSC (*Binary Symmetric Channel*) in [32]. We first need some notation and definitions.

Definition 1. *Let* A *and* B *be two arbitrary matrices of dimension* $m_a \times n$ *and* $m_b \times n$*, and let* $\mathcal{C} \subseteq [n]$*. The residual of matrix* A *given* B *over the column set* $\mathcal{C}$ *is defined by*

$$\mathsf{Res}[\mathsf{A}|\mathsf{B}; \mathcal{C}] = \operatorname{rank}([\mathsf{A}; \mathsf{B}]_{\mathcal{C}}) - \operatorname{rank}(\mathsf{B}_{\mathcal{C}}). \tag{15}$$

It is seen that $\mathsf{Res}[\mathsf{A}|\mathsf{B}; \mathcal{C}]$ is the amount of increase in the rank of $\mathsf{B}_{\mathcal{C}}$ by adding the rows of $\mathsf{A}_{\mathcal{C}}$. In particular, if the rows of $\mathsf{A}_{\mathcal{C}}$ are in the row-span of $\mathsf{B}_{\mathcal{C}}$, then $\mathsf{Res}[\mathsf{A}|\mathsf{B}; \mathcal{C}]$ is 0.

Example 1. *Let* $\mathsf{A} = [1, 1]$ *and* $\mathsf{B} = [1, 0]$*. Then,*

$$\mathsf{Res}[\mathsf{A}|\mathsf{B}; \varnothing] = 0, \ \mathsf{Res}[\mathsf{A}|\mathsf{B}; \{1\}] = 0, \tag{16}$$

[1] In this paper, we mainly deal with linear transforms of i.i.d. variables, and our main motivation for defining this space is that it remains stable under linear operations. Moreover, using the underlying linear structure, we are able to extend the RID in a natural way to all the variables in this space.

$$\mathsf{Res}[\mathsf{A}|\mathsf{B};\{2\}] = 1,\ \mathsf{Res}[\mathsf{A}|\mathsf{B};\{1,2\}] = 1. \tag{17}$$

Many properties of Res simply follow from the algebraic properties of the rank. In this paper, we need additionally the following properties of Res summarized in Proposition 1.

Proposition 1. *Let* A, $\check{\mathsf{A}}$ *be* $m_a \times n$ *matrices. The operator* Res *satisfies the following properties:*

- *(chain rule) Let* $\{\mathcal{R}_i\}_{i=1}^p$ *with* $\mathcal{R}_i \subseteq [m_a]$ *be an arbitrary partition of the rows of* A*, and let* $\mathcal{C} \subseteq [n]$*. Then,*

$$\sum_{i=1}^{p} \mathsf{Res}\Big[\mathsf{A}[\mathcal{R}_i]|\mathsf{A}[\cup_{\ell=1}^{i-1}\mathcal{R}_\ell];\mathcal{C}\Big] = \mathrm{rank}\,(\mathsf{A}_\mathcal{C}). \tag{18}$$

- *(rank-*1 *innovation) Let* $\mathsf{a}, \check{\mathsf{a}} \in \mathbb{R}^{1\times n}$*. Suppose* $\mathcal{C}, \check{\mathcal{C}} \subseteq [n]$ *are arbitrary subsets of the columns of* A *and* $\check{\mathsf{A}}$*. Then,*

$$\begin{aligned}\mathsf{Res}\Big[[\mathsf{a},\check{\mathsf{a}}]|[\mathsf{A},\check{\mathsf{A}}];\mathcal{C}\sqcup\check{\mathcal{C}}\Big] \geqslant \mathsf{Res}\big[\mathsf{a}|\mathsf{A};\mathcal{C}\big] + \mathsf{Res}\big[\check{\mathsf{a}}|\check{\mathsf{A}};\check{\mathcal{C}}\big] \\ - \mathsf{Res}\big[\mathsf{a}|\mathsf{A};\mathcal{C}\big]\mathsf{Res}\big[\check{\mathsf{a}}|\check{\mathsf{A}};\,\check{\mathcal{C}}\big],\end{aligned} \tag{19}$$

where $\sqcup$ *denotes the disjoint union. The equality holds in* (19) *if* A *and* $\check{\mathsf{A}}$ *have non-overlapping set of nonzero rows, i.e.,* A *has zero rows in the row-set corresponding to the nonzero rows of* $\check{\mathsf{A}}$ *and vice versa.* □

Proof: Proof in Appendix A. ■

B. Properties of the RID over $\mathcal{L}$

We first need some notation to simplify the statement of the results in this section. Let $X_1^n \in \mathcal{L}_n$ and $Y_1^m \in \mathcal{L}_m$ be random vectors in $\mathcal{L}$. From the definition in (11), there are matrices A and B of dimension[2] $n \times k$ and $m \times k$ for some finite k and independent nonsingular random variables $Z_1^k \in \mathcal{I}$, such that $X_1^n = \mathsf{A}Z_1^k$ and $Y_1^m = \mathsf{B}Z_1^k$. Since each Z_i has a mixture distribution, it can be represented as $Z_i = \Theta_i C_i + (1-\Theta_i)D_i$, where $C_i \sim p_{c_i}$ and $D_i \sim p_{d_i}$ denote the continuous and the discrete part of Z_i and their corresponding distributions over $\mathbb{R}$, and where $\Theta_i \in \{0,1\}$ is a binary random variable independent of C_i and D_i with $\mathbb{P}[\Theta_i = 1] = d(Z_i)$. We define the support set of the random vector Z_1^k by

$$\mathcal{C} = \{i \in [k] : \Theta_i = 1\}. \tag{20}$$

It is seen that $\mathcal{C}$ is a random subset of $[k]$. Moreover, $\mathbb{P}[i \in \mathcal{C}] = \mathbb{P}[\Theta_i = 1] = d(Z_i)$, thus, $\mathcal{C}$ has the average cardinality $\mathbb{E}[|\mathcal{C}|] = \sum_{i=1}^k d(Z_i)$. We have the following result.

Theorem 2. *Let* (X_1^n, Y_1^m) *and* $\mathcal{C}$ *be as before. Then,*

- $d(X_1^n) = \mathbb{E}[\mathrm{rank}\,(\mathsf{A}_\mathcal{C})]$,
- $d(X_1^n|Y_1^m) = \mathbb{E}\big[\mathsf{Res}[\mathsf{A}|\mathsf{B};\mathcal{C}]\big]$,

with the expectation taken over the random support set $\mathcal{C}$. □

Proof: Proof in Appendix B. ■

Remark 1. *Note that if one of the variables in* Z_1^k*, say* Z_1*, is discrete, then* $\mathbb{P}[\Theta_1 = 1] = 0$*, which implies that* $1 \notin \mathcal{C}$*. Hence, the first column of the matrices* A *and* B *will never be selected. From Theorem 2, this implies that we can drop the fully discrete constituents of* X_1^n *and* Y_1^m *(e.g.,* Z_1 *here) without changing their individual or joint RIDs.* ◊

[2]By adding zero columns whenever needed, without loss of generality, we can always assume that A and B have the same number of columns.

Using Theorem 2 and the properties of the Res operator in Proposition 1, we obtain the following properties of the RID.

Theorem 3. *Let* (X_1^n, Y_1^m) *be a random vector in* $\mathcal{L}$ *as in Theorem 2. Then, we have the following properties:*

- *(positivity)* $d(X_1^n) \geqslant 0$*, with the equality if and only if every* X_i*,* $i \in [n]$*, is discrete.*
- *(invariance)* $d(X_1^n) = d(\mathsf{L}X_1^n)$ *for any invertible* $n \times n$ *matrix* L.
- *(chain rule)* $d(X_1^n, Y_1^m) = d(X_1^n) + d(Y_1^m|X_1^n)$.
- *(symmetry)* $\mathsf{RI}(X_1^n; Y_1^m) = \mathsf{RI}(Y_1^m; X_1^n)$.
- *(positivity)* $\mathsf{RI}(X_1^n; Y_1^m) \geqslant 0$. □

Proof: Proof in Appendix C. ■

Example 2. *Let* Z_1^3 *be i.i.d. with* $d(Z_i) = 0.6$ *for* $i = 1, 2, 3$*. Let* $X = Z_1 + Z_2$ *and* $Y = Z_2 + Z_3$*. This can be written in the following form*

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \mathsf{A}Z_1^3. \tag{21}$$

For computing $d(X)$*, let* $\mathsf{a} = [1, 1, 0]$ *denote the first row of* A*. It is seen that for any* $\mathcal{C} \subseteq [3]$*, the rank of* $\mathsf{a}_\mathcal{C}$ *is equal to* 1 *except when* $\mathcal{C} = \varnothing$ *or* $\mathcal{C} = \{3\}$*. Thus,*

$$\begin{aligned} d(X) &= 1 - \mathbb{P}[\mathcal{C} = \varnothing] - \mathbb{P}[\mathcal{C} = \{3\}] &(22)\\ &= 1 - 0.4^3 - (0.4)^2(0.6) = 0.84. &(23)\end{aligned}$$

From symmetry, we also have $d(Y) = 0.84$*. To compute* $d(X, Y)$*, we can see that for any* $\mathcal{C} \subseteq [3]$*, the rank of* $\mathsf{A}_\mathcal{C}$ *is equal to* $|\mathcal{C}|$ *except when* $\mathcal{C} = [3]$*, where all the columns are selected. Hence,* $\mathrm{rank}\,(\mathsf{A}_\mathcal{C}) = |\mathcal{C}| - \mathbb{I}_{\{\mathcal{C}=[3]\}}$*, which gives*

$$\begin{aligned} d(X,Y) &= \mathbb{E}[|\mathcal{C}|] - \mathbb{P}[\mathcal{C} = [3]] &(24)\\ &= 3 \times 0.6 - 0.6^3 = 1.584. &(25)\end{aligned}$$

Using the chain rule for the RID, we obtain $d(X|Y) = d(Y|X) = 1.584 - 0.84 = 0.744$*, where it is seen that* $d(X|Y) < d(X)$*. We can also directly compute* $d(X|Y)$ *using*

$$d(X|Y) = \mathbb{E}\{\mathsf{Res}[\mathsf{a}|\mathsf{b};\mathcal{C}]\}, \tag{26}$$

where b *denotes the second row of* A*. We can simply check that* $\mathsf{Res}[\mathsf{a}|\mathsf{b};\mathcal{C}] = 1$ *except when* $\mathcal{C} \in \Big\{\varnothing, \{2\}, \{3\}\Big\}$*. Hence,*

$$\begin{aligned} d(X|Y) &= 1 - \mathbb{P}\Big[\mathcal{C} \in \Big\{\varnothing, \{2\}, \{3\}\Big\}\Big] &(27)\\ &= 1 - (0.4^3 + 2 \times 0.6(0.4)^2) &(28)\\ &= 1 - 0.256 = 0.744. &(29)\end{aligned}$$

The Mutual Rényi Information between X *and* Y *is given by*

$$\mathsf{RI}(X;Y) = d(X) - d(X|Y) = 0.84 - 0.744 = 0.096. \tag{30}$$

V. Polarization of the RID

A. Basic Definitions and Results

Before stating the polarization result for the RID, we first define the *erasure process*.

Definition 2. *Let* $\alpha \in [0, 1]$*. An "erasure process" with an initial value* α *is defined as follows:*

1) $e^{\varnothing} = \alpha$. $e^{+} = 2\alpha - \alpha^2$ *and* $e^{-} = \alpha^2$.
2) *Let* $e_n := e^{b_1 b_2 \ldots b_n}$ *for some* $\{+,-\}$*-valued sequence* b_1^n. *Define*

$$e_n^+ := e^{b_1 b_2 \ldots b_n +} = 2e_n - e_n^2,$$
$$e_n^- := e^{b_1 b_2 \ldots b_n -} = e_n^2.$$

Using the $\{+,-\}$ labeling, we can construct a binary tree where each leaf of the tree is labeled with a specific $\{+,-\}$-valued sequence and is assigned the erasure value corresponding to the same $\{+,-\}$-valued sequence.

Let $\{B_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. uniformly distributed $\{+,-\}$-valued random variables. By replacing B_1^n for $\{+,-\}$-labeling b_1^n in the definition of the erasure process, we obtain a stochastic process $E_n = e^{B_1 B_2 \ldots B_n}$. Let $\mathcal{F}_n$ be the σ-field generated by B_1^n. The BEC polarization can be summarized as follows [15, 33]:

1) $(E_n, \mathcal{F}_n, \mathbb{P})$ is a positive martingale bounded in $[0,1]$.
2) E_n converges to $E_\infty \in \{0,1\}$ with $\mathbb{P}(E_\infty = 1) = \alpha$.
3) For any $\beta \in (0, \frac{1}{2})$,

$$\liminf_{n\to\infty} \mathbb{P}(E_n \geqslant 1 - 2^{-2^{n\beta}}) = \alpha, \tag{31}$$
$$\liminf_{n\to\infty} \mathbb{P}(E_n \leqslant 2^{-2^{n\beta}}) = 1 - \alpha. \tag{32}$$

B. RID Polarization

Let $N = 2^n$ be a power of 2 and let Z_1^N be a sequence of i.i.d. nonsingular random variables with an RID $d(Z_i) = \delta \in (0,1)$. Let $\mathbb{H}_N$ be a Hadamard matrix of order N with the following recursive relation between $\mathbb{H}_N$ and $\mathbb{H}_{2N}$

$$\mathbb{H}_N = \begin{bmatrix} \mathsf{h}_1 \\ \vdots \\ \mathsf{h}_N \end{bmatrix} \quad \rightarrow \quad \mathbb{H}_{2N} = \begin{bmatrix} \mathsf{h}_1 & , & \mathsf{h}_1 \\ \mathsf{h}_1 & , & -\mathsf{h}_1 \\ \vdots & , & \vdots \\ \mathsf{h}_i & , & \mathsf{h}_i \\ \mathsf{h}_i & , & -\mathsf{h}_i \\ \vdots & , & \vdots \end{bmatrix}, \tag{33}$$

where h_i, $i \in [N]$, denotes the i-th row of $\mathbb{H}_N$. This corresponds to a standard Hadamard matrix with shuffled rows. This construction simplifies the proofs, but all the result are still valid for the standard Hadamard matrix $\mathbf{H}_N$ without any shuffling.

Let $\mathbb{H}_N$ be as in (33) and let $X_1^N = \mathbb{H}_N Z_1^N$ be the vector of variables obtained by the Hadamard transform of Z_1^N. Let us define

$$I_n : [N] \to [0,1],\ I_n(i) = d(X_i | X_1^{i-1}), i \in [N]. \tag{34}$$

Assume that b_1^n is the binary expansion of $i-1$. By replacing 0 by $+$ and 1 by $-$, we can equivalently label $I_n(i)$ by a sequence of $\{+,-\}$ of length n, i.e., $I_n(i) = I^{b_1 b_2 \ldots b_n}$. Similar to the erasure process, we can convert I_n to a stochastic process $I_n = I^{B_1 B_2 \ldots B_n}$ by using i.i.d. uniform $\{+,-\}$-valued random variables B_1^n. We can now prove the following theorem.

Theorem 4 (RID Polarization)**.** $(I_n, \mathcal{F}_n, \mathbb{P})$ *is an erasure stochastic process with initial value* δ *polarizing to* $\{0,1\}$. □

Proof: For $n = 0$, we have a Hadamard matrix of order $N = 2^0 = 1$ which is simply a number, thus, $X_1 = Z_1$ and we have $I_0(1) = d(Z_1) = \delta$. Consider an arbitrary n, let $N = 2^n$ and let I_n be defined as in (34). We need to prove that I_n satisfies the following recursion for $i \in [2^n]$

$$I_n(i)^+ = I_{n+1}(2i-1) = 2I_n(i) - I_n(i)^2 \tag{35}$$
$$I_n(i)^- = I_{n+1}(2i) = I_n(i)^2. \tag{36}$$

As Z_1^N are i.i.d. nonsingular random variables, it results that $X_1^N = \mathbb{H}_N Z_1^N$ belongs to the space $\mathcal{L}$ generated by Z_1^N. Hence, using the rank characterization for the RID over $\mathcal{L}$ in Theorem 2, we have

$$I_n(i) = d(X_i | X_1^{i-1}) = \mathbb{E}_{\mathcal{C}}\big[\mathsf{Res}[\mathsf{h}_i | \mathbb{H}_N[1\!:\!i-1]; \mathcal{C}]\big], \tag{37}$$

where $\mathbb{H}_N[1\!:\!i-1]$ denotes the $(i-1)\times N$ matrix consisting of the first $i-1$ rows of $\mathbb{H}_N$ and where $\mathcal{C}$ denotes the random support of continuous parts of Z_1^N as defined in (20). Recall that $i \in \mathcal{C}$ if and only if the random variable Z_i is sampled according to the continuous part of its distribution.

At stage $n+1$, we have the term $I_n(i)^+$ which corresponds to the row $2i-1$ of $\mathbb{H}_{2N}$ as follows

$$\mathbb{H}_{2N}[1\!:\!2i-1] = \begin{bmatrix} \mathsf{h}_1 & , & \mathsf{h}_1 \\ \mathsf{h}_1 & , & -\mathsf{h}_1 \\ \vdots & , & \vdots \\ \mathsf{h}_{i-1} & , & \mathsf{h}_{i-1} \\ \mathsf{h}_{i-1} & , & -\mathsf{h}_{i-1} \\ \mathsf{h}_i & , & \mathsf{h}_i \end{bmatrix}, \tag{38}$$

where $I_n(i)^+$ is defined similar to (37) by

$$I_n(i)^+ = \mathbb{E}_{\mathcal{C},\check{\mathcal{C}}}\Big[\mathsf{Res}\Big[[\mathsf{h}_i, \mathsf{h}_i] \big| \mathbb{H}_{2N}[1\!:\!2(i-1)]; \mathcal{C} \sqcup \check{\mathcal{C}}\Big]\Big], \tag{39}$$

where $\mathcal{C}$ and $\check{\mathcal{C}}$ denote the support set of Z_1^N and $\check{Z}_1^N := Z_{N+1}^{2N}$. As $\check{Z}_1^N$ is an independent copy of Z_1^N, the support sets $\mathcal{C}$ and $\check{\mathcal{C}}$ are independent and identically distributed. Applying a simple row operation to $\mathbb{H}_{2N}[1\!:\!2(i-1)]$, which preserves the rank, we have that

$$\mathsf{Res}\Big[[\mathsf{h}_i, \mathsf{h}_i] \big| \mathbb{H}_{2N}[1\!:\!2(i-1)]; \mathcal{C} \sqcup \check{\mathcal{C}}\Big] = \mathsf{Res}\Big[[\mathsf{h}_i, \mathsf{h}_i] \big| [\mathsf{L}, \mathsf{R}]; \mathcal{C} \sqcup \check{\mathcal{C}}\Big],$$

where L and R are given by

$$\mathsf{L} = \begin{bmatrix} \mathsf{h}_1 \\ \mathbf{0} \\ \vdots \\ \mathsf{h}_{i-1} \\ \mathbf{0} \end{bmatrix}, \mathsf{R} = \begin{bmatrix} \mathbf{0} \\ -\mathsf{h}_1 \\ \vdots \\ \mathbf{0} \\ -\mathsf{h}_{i-1} \end{bmatrix}. \tag{40}$$

Since L and R have non-overlapping rows, using the rank-1 innovation property in Proposition 1, we obtain that

$$\begin{aligned}\mathsf{Res}\Big[[\mathsf{h}_i, \mathsf{h}_i] \big| [\mathsf{L}, \mathsf{R}]; \mathcal{C} \sqcup \check{\mathcal{C}}\Big] &= \mathsf{Res}\Big[\mathsf{h}_i \big| \mathsf{L}; \mathcal{C}\Big] + \mathsf{Res}\Big[\mathsf{h}_i \big| \mathsf{R}; \check{\mathcal{C}}\Big] \\ &\quad - \mathsf{Res}\Big[\mathsf{h}_i \big| \mathsf{L}; \mathcal{C}\Big] \mathsf{Res}\Big[\mathsf{h}_i \big| \mathsf{R}; \check{\mathcal{C}}\Big].\end{aligned} \tag{41}$$

From (40), it is also seen that

$$\mathsf{Res}[h_i | \mathsf{L}; \mathcal{C}] = \mathsf{Res}\Big[h_i \big| \mathbb{H}_N[1\!:\!i-1]; \mathcal{C}\Big], \tag{42}$$
$$\mathsf{Res}[h_i | \mathsf{R}; \check{\mathcal{C}}] = \mathsf{Res}\Big[h_i \big| \mathbb{H}_N[1\!:\!i-1]; \check{\mathcal{C}}\Big]. \tag{43}$$

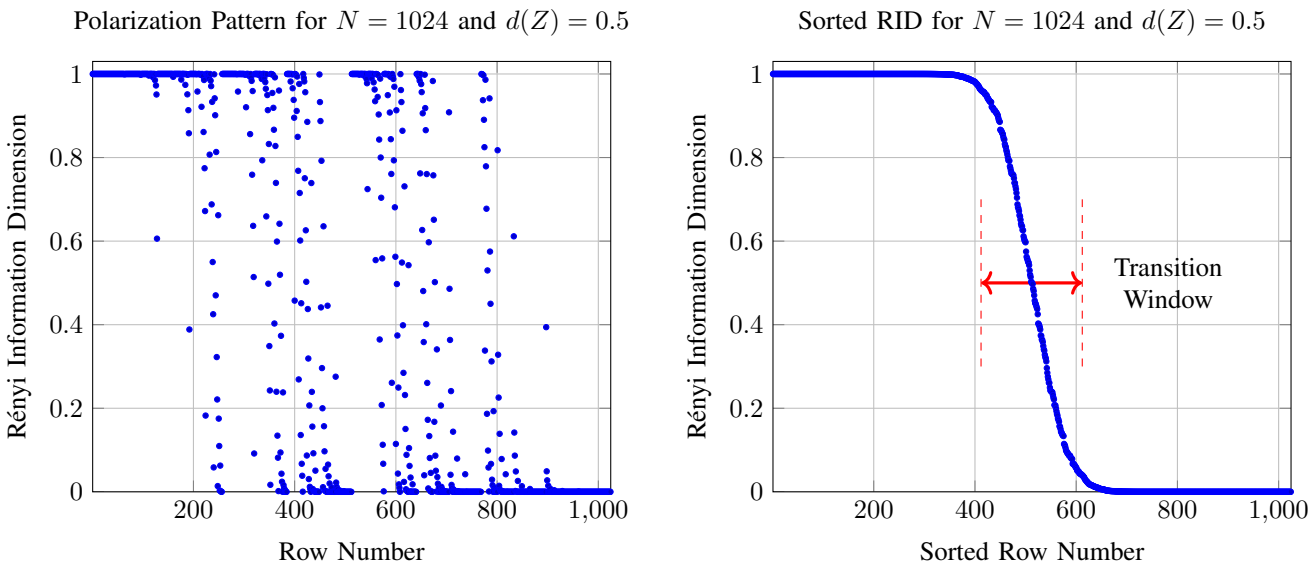

Fig. 1: Polarization pattern of an i.i.d. source $\{Z_i\}_{i=1}^{\infty}$ with $d(Z_1) = 0.5$ after being transformed by the Hadamard matrix of order $N = 1024$. It is seen that most of the rows are polarized to their corresponding RIDs, and the fraction of those rows polarized to 1 converges to $d(Z_1) = 0.5$.

As $\mathcal{C}$ and $\check{\mathcal{C}}$ are i.i.d., taking the expectation from (41), and using (42), (43), and also (37), we have

$$\begin{aligned} I_n(i)^+ &= I_n(i) + I_n(i) - I_n(i) \cdot I_n(i) \\ &= 2I_n(i) - I_n(i)^2, \end{aligned} \tag{44}$$

which proves the first identity in (35). To prove the second identity in (35), let $\check{Z}_1^N$ and Z_1^N be independent copies as defined before and let $W_1^N = \mathbb{H}_N Z_1^N$ and $\check{W}_1^N = \mathbb{H}_N \check{Z}_1^N$. We obtain that

$$X_{2i-1} = W_i + \check{W}_i, \;\; X_{2i} = W_i - \check{W}_i. \tag{45}$$

Moreover, by definition, we have

$$\begin{aligned} I_n(i)^+ &= d(X_{2i-1}|X_1^{2i-2}) = d(W_i + \check{W}_i|W_1^{i-1}, \check{W}_1^{i-1}) \\ I_n(i)^- &= d(X_{2i}|X_1^{2i-1}) \\ &= d(W_i - \check{W}_i|W_1^{i-1}, \check{W}_1^{i-1}, W_i + \check{W}_i). \end{aligned}$$

Applying the chain rule for RID from Theorem 3 and using the independence of W_1^N and $\check{W}_1^N$, we obtain

$$\begin{aligned} I_n(i)^+ + I_n(i)^- &= d(W_i + \check{W}_i, W_i - \check{W}_i|W_1^{i-1}, \check{W}_1^{i-1}) \\ &= d(W_i, \check{W}_i|W_1^{i-1}, \check{W}_1^{i-1}) \\ &= 2d(W_i|W_1^{i-1}) = 2I_n(i). \end{aligned} \tag{46}$$

From (44), this implies $I_n(i)^- = I_n(i)^2$, which proves the second identity in (35). From Definition 2, this indicates that I_n is an erasure process with the initial value $d(Z_1) = \delta$. ■

Fig. 1 illustrates the polarization pattern for an i.i.d. source $\{Z_i\}_{i=1}^{\infty}$ with $d(Z_1) = 0.5$ after being transformed by the Hadamard matrix of order $N = 2^{10} = 1024$. It is seen that even for $N = 1024$ more than 80% of the rows are polarized to their corresponding RIDs.

C. RID-Preserving Matrices

Let $\{Z_i\}_{i=1}^{\infty}$ be an i.i.d. source with $d(Z_1) \in (0, 1)$. Let $\{\mathsf{A}^{(k)}\}_{k=1}^{\infty}$ be a sequence of matrices of order $m(k) \times k$. We say that $\{\mathsf{A}^{(k)}\}_{k=1}^{\infty}$ is an RID-preserving family for $\{Z_i\}_{i=1}^{\infty}$ if and only if

$$d(\mathsf{A}^{(k)} Z_1^k) \geqslant d(Z_1^k) - o(k), \tag{47}$$

where $o(k)$ denotes a vanishing term compared with k as k tends to infinity. We define the asymptotic measurement rate of the family $\{\mathsf{A}^{(k)}\}_{k=1}^{\infty}$ by $\rho := \limsup_{k\to\infty} \frac{m(k)}{k}$. From (47), it is seen that taking measurements with this family of matrices asymptotically preserves the whole RID of the source.

Proposition 5. *Let $\{Z_i\}_{i=1}^{\infty}$ be an i.i.d. source with a given $d(Z_1) \in (0, 1)$ and let $\{\mathsf{A}^{(k)}\}_{k=1}^{\infty}$ be an RID-preserving family for the source. Then, $\rho \geqslant d(Z_1)$.* □

Proof: First note that for every k, the random vector $\mathsf{A}^{(k)} Z_1^k$ belongs the space $\mathcal{L}$ generated by the i.i.d. nonsingular variables $\{Z_i\}_{i=1}^{\infty}$, thus, it has a well-defined RID. In particular, if $\mathcal{C}$ is the support set of the i.i.d. nonsingular variables Z_1^k, as defined in (20), from the rank property of the RID proved in Theorem 2, we have

$$m(k) \geqslant \mathbb{E}[\mathrm{rank}\,(\mathsf{A}_{\mathcal{C}}^{(k)})] = d(\mathsf{A}^{(k)} Z_1^k) \tag{48}$$

$$\geqslant d(Z_1^k) - o(k) = kd(Z_1) - o(k). \tag{49}$$

Dividing both sides by k, and taking the limit as k tends to infinity, we obtain the desired result $\rho \geqslant d(Z_1)$. ■

Proposition (5) implies that to be RID-preserving, any family of matrices needs to have a measurement rate at least as large as the RID of the source. To show that this measurement rate is indeed sufficient, we build an RID-preserving family $\{\mathsf{A}^{(k)}\}$, labeled with $k \in \{2^n \colon n \in \mathbb{N}\}$, where $\mathsf{A}^{(k)}$ is a submatrix of the Hadamard matrix $\mathbb{H}_k$, obtained by selecting

specific rows of $\mathbb{H}_k$. We then prove that the measurement rate of the constructed family is $d(Z_1)$.

Let $N = 2^n$ be a power of two, let $\mathbb{H}_N$ be the shuffled Hadamard matrix of order N defined in (33), and let $X_1^N = \mathbb{H}_N Z_1^N$. Set $\beta \in (0, \frac{1}{2})$, and $\epsilon^{(N)} = 2^{-N^\beta}$, and let $\mathsf{H}^{(N)}$ be a submatrix of $\mathbb{H}_N$ obtained by selecting those rows of $\mathbb{H}_N$ belonging to the index set

$$\mathcal{S}^{(N)} = \{i \in [N] : d(X_i|X_1^{i-1}) \geqslant \epsilon^{(N)}\}. \tag{50}$$

Let $m(N) = |\mathcal{S}^{(N)}|$ be the number of rows of $\mathsf{H}^{(N)}$ and let $\{\mathsf{H}^{(N)}\}$ be the resulting family of matrices indexed with N, where N is a power of 2.

Proposition 6. *The sequence of matrices $\{\mathsf{H}^{(N)}\}$ is RID-preserving for the i.i.d. source $\{Z_i\}_{i=1}^\infty$ and has a measurement rate $\rho = d(Z_1)$.* □

Proof: Let $X_1^N = \mathbb{H}_N Z_1^N$. Note that from Theorem 4, the sequence of conditional RIDs $\{d(X_i|X_1^{i-1})\}_{i=1}^N$ is an erasure process with an initial value $d(Z_1)$. Thus, applying the polarization rate result in (31), we obtain that for a $\beta \in (0, \frac{1}{2})$ and $\epsilon^{(N)} = 2^{-N^\beta}$, the fraction of those conditional RIDs with a value larger than $\epsilon^{(N)}$, i.e., those belonging to $\mathcal{S}^{(N)}$, must converge to $d(Z_1)$. This implies that $\rho = \limsup_{N\to\infty} \frac{m(N)}{N} = d(Z_1)$. To prove the RID-preserving property, let $\mathsf{r}(\ell) \in \mathcal{S}^{(N)}$ be the index of the ℓ-th row of $\mathsf{H}^{(N)}$ among the rows of $\mathbb{H}_N$. Then, we have

$$\begin{aligned}
d(\mathsf{H}^{(N)} Z_1^N) &= \sum_{\ell=1}^{m(N)} d(X_{\mathsf{r}(\ell)}|X_{\mathsf{r}(1)}, X_{\mathsf{r}(2)}, \ldots, X_{\mathsf{r}(\ell-1)}) \\
&\overset{(a)}{\geqslant} \sum_{\ell=1}^{m(N)} d(X_{\mathsf{r}(\ell)}|X_1^{\mathsf{r}(\ell)-1}) \\
&\overset{(b)}{=} d(X_1^N) - \sum_{s\notin\mathcal{S}^{(N)}} d(X_s|X_1^{s-1}) \\
&\geqslant d(Z_1^N) - (N - |\mathcal{S}^{(N)}|)\epsilon^{(N)} \\
&\geqslant d(Z_1^N) - o(N),
\end{aligned} \tag{51}$$

where in (a) we used the positivity of the Mutual Rényi Information proved in Theorem 3 and the fact that conditioning reduces the RID, and where in (b) we used the chain rule for the RID given by $d(X_1^N) = \sum_{s=1}^N d(X_s|X_1^{s-1})$. The Eq. (51) confirms the RID-preserving property of $\{\mathsf{H}^{(N)}\}$. This completes the proof. ■

VI. Multi-terminal (Distributed) Polarization

A. Multi-terminal RID Polarization

The RID polarization proved in Theorem 4 can be extended to the multi-terminal signals. Let $\{(U_i, V_i)\}_{i=1}^\infty$ be a sequence of i.i.d. 2-dim vectors in $\mathcal{L}$. Since $[U_1, V_1] \in \mathcal{L}$, there are $\mathsf{a}, \mathsf{b} \in \mathbb{R}^{1\times k}$ and i.i.d. nonsingular variables Z_1^k such that $[U_1; V_1] = [\mathsf{a}; \mathsf{b}]Z_1^k$. Let $N = 2^n$ be power of 2 and let $X_1^N = \mathbb{H}_N U_1^N$ and $Y_1^N = \mathbb{H}_N V_1^N$, where $\mathbb{H}_N$ is as in (33). In the *single-terminal* case in Section V-B, we used the chain rule for the variables $X_1^N = \mathbb{H}_N Z_1^N$ to expand $d(X_1^N)$ in terms of the conditional RIDs $\{d(X_i|X_1^{i-1})\}_{i=1}^N$, thus, obtaining an erasure process with initial value $d(Z_1)$ polarizing to $\{0, 1\}$. In the multi-terminal case, however, we obtain different erasure processes by applying the chain rule to $d(X_1^N, Y_1^N)$ with different expansion orders. For example, if we expand first in terms of X_1^N and then in terms of Y_1^N, we obtain the following two sequences for $i \in [N]$:

$$I_n(i) = d(X_i|X_1^{i-1}), J_n(i) = d(Y_i|Y_1^{i-1}, X_1^N). \tag{52}$$

To show that I_n and J_n are indeed erasure processes, similar to Section V-B, we label different components of I_n and J_n with $\{+, -\}$-valued sequences. We remove the details for brevity. We obtain the following result.

Theorem 7. *$(I_n, \mathcal{F}_n, \mathbb{P})$ and $(J_n, \mathcal{F}_n, \mathbb{P})$ are erasure processes with initial value $d(U_1)$ and $d(V_1|U_1)$ respectively, polarizing to $\{0, 1\}$.* □

Proof: Proof in Appendix D. ■

By changing the order of expansion of $d(X_1^N, Y_1^N)$, i.e., first expanding with respect to Y_1^N and then with respect to X_1^N, we obtain another 2-dim erasure process (I_n, J_n) with the initial value $(d(U_1|V_1), d(V_1))$, rather than $(d(U_1), d(V_1|U_1))$. In fact, by applying the monotone chain rule expansion introduced in [34], we can expand $d(X_1^N, Y_1^N)$ jointly (and simultaneously) in terms of Xs and Ys, thus, we can construct different 2-dim polarizing erasure processes (I_n, J_n) that converge almost surely to $(I_\infty, J_\infty) \in \{0, 1\}^2$. Also, the closure of the region of all possible $(\bar{I}, \bar{J}) := \mathbb{E}[(I_\infty, J_\infty)]$ for polarizing processes (I_n, J_n) contains the dominant face of the 2-dim region given by

$$\begin{aligned}
\mathcal{R}_2 = \{\boldsymbol{\rho} \in \mathbb{R}_+^2 : \rho_1 \geqslant d(U_1|V_1), \rho_2 \geqslant d(V_1|U_1), \\
\rho_1 + \rho_2 \geqslant d(U_1, V_1)\},
\end{aligned} \tag{53}$$

which is a line connecting two points $(d(U_1), d(V_1|U_1))$ and $(d(U_1|V_1), d(V_1))$ in $\mathbb{R}_+^2$. This resembles the Slepian-Wolf region for the distributed source coding [35]. The results can be extended to an i.i.d. sequence of d-dim signal $\mathbf{U} = (U_1, U_2, \ldots, U_d)$, for some $d \geqslant 3$. By applying the chain rule in different orders and using the results in [34], it is possible to build a d-dim erasure process $(I_n^{(1)}, I_n^{(2)}, \ldots, I_n^{(d)})$ that converges almost surely to $(I_\infty^{(1)}, I_\infty^{(2)}, \ldots, I_\infty^{(d)}) \in \{0, 1\}^d$. Moreover, the closure of the region of all d-dim averages $\mathbb{E}[(I_\infty^{(1)}, I_\infty^{(2)}, \ldots, I_\infty^{(d)})]$ corresponds to the dominant face of the region

$$\mathcal{R}_d = \{\boldsymbol{\rho} \in \mathbb{R}_+^d : \sum_{i\in\mathcal{T}} \rho_i \geqslant d(\mathbf{U}_\mathcal{T}|\mathbf{U}_{\mathcal{T}^c}),\ \forall \mathcal{T} \subseteq [d]\}, \tag{54}$$

where $\mathbf{U}_\mathcal{T}$ denotes the subvector of $\mathbf{U}$ obtained by selecting the components in $\mathcal{T} \subseteq [d]$.

B. RID-Preserving Matrices for Multi-terminal Signals

The RID-preserving property in Section V-C can also be extended in a natural way to multi-terminal signals. For simplicity, we focus on the 2-dim case. The results can be extended to dimensions larger than 2.

Let $\{(U_i, V_i)\}_{i=1}^\infty$ be a sequence of i.i.d. 2-dim signals belonging to $\mathcal{L}$ and let $\{\mathsf{A}^{(k)}, \mathsf{B}^{(k)}\}_{k=1}^\infty$ be a sequence of matrices of order $m_a(k) \times k$ and $m_b(k) \times k$. We call

$\{\mathsf{A}^{(k)},\mathsf{B}^{(k)}\}_{k=1}^{\infty}$ an RID-preserving sequence for $\{(U_i,V_i)\}_{i=1}^{\infty}$ if and only if

$$d(\mathsf{A}^{(k)}U_1^k,\mathsf{B}^{(k)}V_1^k)\geqslant d(U_1^k,V_1^k)-o(k), \tag{55}$$

where $o(k)$ denotes a term vanishing in the dimension k. We define $\rho_a := \limsup_{k\to\infty}\frac{m_a(k)}{k}$, and $\rho_b := \limsup_{k\to\infty}\frac{m_b(k)}{k}$ the asymptotic measurement rate of the family. We obtain the following result.

Theorem 8. *Let $\{(U_i,V_i)\}_{i=1}^{\infty}$ be an i.i.d. source with the joint RID $d(U_1,V_1)$ and conditional RIDs $d(U_1|V_1)$ and $d(V_1|U_1)$. Let $\{\mathsf{A}^{(k)},\mathsf{B}^{(k)}\}_{k=1}^{\infty}$ be a family of RID-preserving matrices for the source. Then,*

$$\rho_a\geqslant d(U_1|V_1),\rho_b\geqslant d(V_1|U_1),\rho_a+\rho_b\geqslant d(U_1,V_1). \tag{56}$$

Proof: First note that from the rank property for the RID proved in Theorem 2 and the RID-preserving property in (55), it results that

$$m_a(k)+m_b(k)\geqslant d(\mathsf{A}^{(k)}U_1^k,\mathsf{B}^{(k)}V_1^k)\geqslant d(U_1^k,V_1^k)-o(k),$$

simply because the rank of $[\mathsf{A}^{(k)};\mathsf{B}^{(k)}]$ is less than its number of rows $m_a(k)+m_b(k)$. This implies that $\rho_a+\rho_b\geqslant d(U_1,V_1)$. To prove the other two inequalities, note that the RID-preserving property in (55) can be written as

$$d(U_1^k,V_1^k|\mathsf{A}^{(k)}U_1^k,\mathsf{B}^{(k)}V_1^k)\leqslant o(k). \tag{57}$$

Applying the chain rule, we have

$$d(U_1^k|\mathsf{A}^{(k)}U_1^k,\mathsf{B}^{(k)}V_1^k)+d(V_1^k|\mathsf{B}^{(k)}V_1^k,U_1^k)\leqslant o(k). \tag{58}$$

Using the positivity of the RID, this immediately implies that $d(V_1^k|\mathsf{B}^{(k)}V_1^k,U_1^k)\leqslant o(k)$, which using the rank property for the RID gives

$$m_b(k)\geqslant d(\mathsf{B}^{(k)}V_1^k|U_1^k)\geqslant d(V_1^k|U_1^k)-o(k), \tag{59}$$

which implies the desired result $\rho_b\geqslant d(V_1|U_1)$. The other inequality $\rho_a\geqslant d(U_1|V_1)$ follows similarly. ∎

Theorem 8 can be extended in a natural way to the d-dim sources, where it can be shown that for a family of matrices to be RID-preserving for the d-dim source, it is necessary that their measurement rate belong to the d-dim region in (54).

VII. Applications in Compressed Sensing

In Compressed Sensing, the aim is to recover a structured signal $\mathbf{x}=x_1^N$ by taking only a few number of linear measurements $\mathbf{y}=\mathsf{A}\mathbf{x}$, where $\mathbf{y}=y_1^m$ denotes the vector of m linear measurements taken via the $m\times N$ matrix A. If the signal $\mathbf{x}$ has a sparse representation in a basis with at most $k\ll N$ nonzero elements (k-sparse) and if A is suitably designed with respect to this basis, $\mathbf{x}$ can be recovered by taking $m\ll N$ measurements [16–19]. Fix a $\delta\in(0,1)$ and consider an N-dimensional signal $X_1^N\in\mathbb{R}^N$ whose components are sampled i.i.d. from the distribution

$$p_X(x)=(1-\delta)\mathbb{I}_0(x)+\delta p_c(x), \tag{60}$$

where $\mathbb{I}_0(x)$ denotes a delta measure at point 0 and where p_c is a continuous probability distribution. For a large block-length N, almost all the realizations $\mathbf{x}=x_1^n$ of the signal X_1^N have approximately $k=N\delta$ nonzero components, thus, a sparse signal with a sparsity ratio $\delta=\frac{k}{N}=d(X)$.

Let $N=2^n$ be a power of 2 and let $\mathcal{S}^{(N)}$ be as in (50). Let $\mathsf{H}^{(N)}$ be the submatrix of $\mathbb{H}_N$ consisting of the rows in $\mathcal{S}^{(N)}$ and let $Y_1^{m(N)}=\mathsf{H}^{(N)}X_1^N$ be the measurements. From the RID-preserving property of $\mathsf{H}^{(N)}$ proved in Proposition 6, we have that $d(X_1^N|Y_1^{m(N)})=o(N)$. From the definition of the RID, this implies that for a sufficiently large q, the measurements $Y_1^{m(N)}$ capture a significant fraction of the information of the quantized signal $\langle X_1^N\rangle_q$.

In this paper, we mainly focused on the polarization of the RID as an information measure. It is interesting to know whether the RID polarization for the infinite-alphabet signals proved in this paper, can be exploited as in the case of discrete polarization for finite-alphabet sources [20] to build a decoder that recovers the initial signal X_1^N from the collection of linear measurements $Y_1^{m(N)}$ up to a negligible distortion (e.g., error probability or l_2-distortion). In this section, we propose an approach to establish such an *operational* aspect of the problem although we do not prove it.

Let $\epsilon_N=2^{-N^\beta}$ for some $\beta\in(0,\frac{1}{2})$ be the threshold value used for constructing H_N in (50). Then, for a sufficiently large block-length N, it results that

$$\lim_{q\to\infty}\frac{H(\langle X_1^N\rangle_q|Y_1^{m(N)})}{\log_2(q)}=d(X_1^N|Y_1^{m(N)})\leqslant N\epsilon_N. \tag{61}$$

Let $\alpha_N\in\mathbb{R}_+$ and $q_N\in\mathbb{N}$ be two sequences of $N\in\{2^n:n\in\mathbb{Z}_+\}$ such that $\lim_{N\to\infty}q_N=\infty$ and

$$H(\langle X_1^N\rangle_{q_N}|Y_1^{m(N)})\leqslant\alpha_N N\epsilon_N\log_2(q_N)=:p_N. \tag{62}$$

Note that α_N is a scaling factor used to ensure that a sequence of q_N satisfying (61) exists. We prove that under the stated conditions, if p_N tends to zero as N tends to infinity, then we can decode the quantized signal $\langle X_1^N\rangle_{q_N}$ with a vanishing error probability p_N, using the *MAP* (Maximum a posteriori Probability) decoder. We use the following simple lemma.

Lemma 9. *Let D be a discrete random variable taking values in the countable alphabet $\mathcal{D}$ and let Y be an arbitrary random variable, jointly distributed with D, such that the conditional distribution (probability mass function) $p(d|y)$ is well-defined almost surely. Let $\mathcal{E}=\{(d,y):\widehat{D}(y)\neq d\}$ be the error event of the MAP decoder defined by $\widehat{D}(y)=\arg\max_{d'\in\mathcal{D}}p(d'|y)$. Then, the average error probability satisfies $\mathbb{P}[\mathcal{E}]\leqslant H(D|Y)\log_2(e)$, where $H(D|Y)$ denotes the conditional entropy of D given Y in bits.* □

Proof: Proof in Appendix E. ∎

Using Lemma 9, we can see from (62) that the quantized components $\langle X_1^N\rangle_{q_N}$ can be recovered up to an average error probability $p_N\log_2(e)$. This implies that, with a very high probability, the desired signal X_1^N can be recovered up to a vanishing distortion $\frac{1}{q_N}$ provided that p_N tends to 0. We state this as the following conjecture.

Conjecture 1. *Let $N\in\{2^n:n\in\mathbb{Z}_+\}$ and let $\beta\in(0,\frac{1}{2})$ and $\epsilon_N=2^{-N^\beta}$ as before. There exists a scaling factor α_N and a quantization factor $q_N\overset{N\to\infty}{\longrightarrow}\infty$ with $p_N:=\alpha_N N\epsilon_N\log_2(q_N)\overset{N\to\infty}{\longrightarrow}0$.*

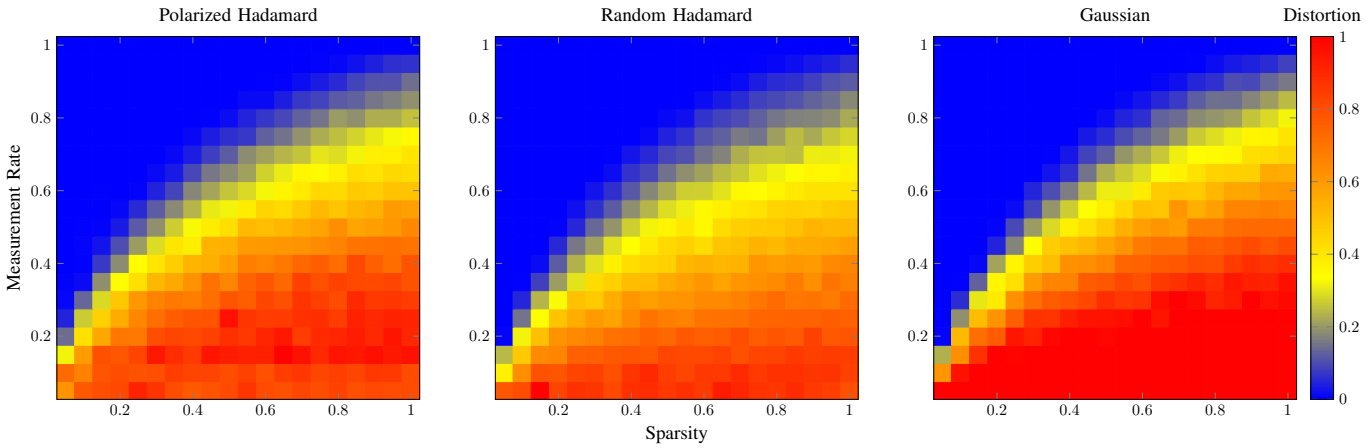

Fig. 2: The Rate-Distortion (RD) curve for partial Hadamard matrices and its comparison with that of random Hadamard and random Gaussian matrices. The horizontal axis indicates the sparsity level of the signal (the fraction of nonzero components in the signal), and the vertical axis shows the measurement rate (the number of measurements per dimension of the signal).

For example, if we set $\alpha_N = 1$ for all N, due to the logarithmic dependence of p_N on q_N, a sequence $q_N = o(2^{2^{N^\beta}})$ would be sufficient for the Conjecture 1 to be true. Considering the doubly-exponential growth rate of $2^{2^{N^\beta}}$ as a function of N, we believe that such a sequence q_N should exist. This would establish the operational performance of our constructed matrices for Compressed Sensing of i.i.d. sources. Although we do not directly prove Conjecture 1, we use numerical simulations in Section VII-A to illustrate that our constructed polarized Hadamard matrices along with the off-the-shelf low-complexity l_1-norm minimization algorithm in Compressed Sensing (instead of the more complicated MAP decoder) still have a promising operational performance.

Using partial Hadamard matrices has several practical advantages. Their components are ± 1 and can be robustly implemented as on-off pattern in many practical measurement devices and easily stored in a computer. Partial Hadamard matrices also yield computationally efficient recovery algorithms. In brief, a crucial step in all recovery algorithms in Compressed Sensing is computing $\mathsf{A}^{\mathsf{T}}\mathbf{y}$ (*matched-filtering*), in which the inner product of the columns of the measurement matrix A with the observations $\mathbf{y} = \mathsf{A}\,\mathbf{x}$ is calculated. Using the structure of the Hadamard matrices, this can be done with $O(N \log_2(N))$ rather than $O(N^2)$ operations needed for the traditional matrix-vector multiplication. Even for small dimensions such as $N = 1000$ this is around 100 times faster.

A. Simulation Results

In this section, we assess the operational performance of the partial Hadamard matrices constructed in Section V-C via numerical simulations.

1) Measurement Matrix and Recovery Algorithm

For simulations, we use a zero-mean and unit-variance sparse distribution as in (60)

$$p_X(x) = (1-\delta)\mathbb{I}_0(x) + \delta p_c(x), \tag{63}$$

where $\mathbb{I}_0(x)$ denotes the delta measure at zero, and where p_c is a fixed zero-mean continuous distribution with a variance $\frac{1}{\delta}$. We do the simulations for different sparsity levels $\delta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ of the underlying signal. Note that for a given δ in this list, the RID of the generated signal $X \sim p_X$ is given by $d(X) = \delta$. We use the mean square error (MSE) as the distortion measure $d(\mathbf{x}, \widehat{\mathbf{x}}) = \|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2$ between the target signal $\mathbf{x}$ and the estimate $\widehat{\mathbf{x}}$ obtained via the recovery algorithm. The simulations are done with the Hadamard matrix of order $N = 1024$. To build the measurement matrix A, we sort the rows of $\mathbb{H}_N$ according to their conditional RIDs and select those rows with highest RID. We use the l_1-norm minimization algorithm to recover the signal:

$$\widehat{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{w}\in\mathbb{R}^N}{\arg\min} \|\mathbf{w}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathsf{A}\mathbf{w}, \tag{64}$$

where the input to this algorithm is the vector of linear measurements $\mathbf{y} = \mathsf{A}\,\mathbf{x}$ for the given signal $\mathbf{x}$. We use the CVX package [36] to solve (64).

2) Comparison with other Measurement Matrices

We compare the performance of our constructed matrices with random Hadamard matrices and random Gaussian matrices extensively used in Compressed Sensing. Fig. 2 illustrates the resulting rate-distortion curve of the l_1-norm minimization (64) for these three families of matrices. It is seen that our constructed matrices have a performance very close to that of other two families, while being deterministically constructed.

VIII. Conclusion

In this paper, we generalized the definition of RID as an information measure from scalar random variables initially proposed in [3] to a larger family of vector-valued random variables. We proved that for such a family the joint and the conditional RIDs are well-defined and can be computed with a closed form formula. Using this, we proved that the RID of a sequence of i.i.d. nonsingular random variables

polarizes to the extreme values of 0 and 1 when transformed by Hadamard matrices. We also gave a closed-form expression for the polarization pattern using the BEC polarization in the discrete case. This gives a natural counterpart of the finite-alphabet source polarization in the infinite-alphabet case. We investigated some of the applications of the new polarization phenomenon in Compressed Sensing.

Appendix A
Proof of Proposition 1

For the first part, let $\mathscr{R}_1 = \mathcal{R}_1$, and for $i = 2, 3, \dots, p$, set $\mathscr{R}_i = \cup_{\ell=1}^{i} \mathcal{R}_\ell$. Note that since $\{\mathcal{R}_i\}_{i=1}^{p}$ is a partition of the rows of A, we have $\mathscr{R}_p = [m_a]$. Using the definition of Res in (15), we obtain

$$\sum_{i=1}^{p} \mathsf{Res}\Big[\mathsf{A}[\mathcal{R}_i] \big| \mathsf{A}[\cup_{\ell=1}^{i-1}\mathcal{R}_\ell]; \mathcal{C}\Big] = \operatorname{rank}\big[\mathsf{A}[\mathscr{R}_1]_{\mathcal{C}}\big] \tag{65}$$

$$+ \sum_{i=2}^{p} \Big(\operatorname{rank}\big[\mathsf{A}[\mathscr{R}_i]_{\mathcal{C}}\big] - \operatorname{rank}\big[\mathsf{A}[\mathscr{R}_{i-1}]_{\mathcal{C}}\big]\Big) \tag{66}$$

$$= \operatorname{rank}\big[\mathsf{A}[\mathscr{R}_p]_{\mathcal{C}}\big] = \operatorname{rank}\big[\mathsf{A}_{\mathcal{C}}\big]. \tag{67}$$

To prove the rank-1 innovation property in the second part, first note that from the definition of Res operator in (15), $\mathsf{Res}\big[\mathsf{a}, \check{\mathsf{a}} | [\mathsf{A}, \check{\mathsf{A}}]; \mathcal{C} \sqcup \check{\mathcal{C}}\big] \in \{0, 1\}$. In particular, it is zero if adding the individual row $\big[\mathsf{a}, \check{\mathsf{a}}\big]$ does not increase the rank of the matrix $[\mathsf{A}, \check{\mathsf{A}}]$ at column set $\mathcal{C} \sqcup \check{\mathcal{C}}$, where in that case, by simply restricting to $\mathcal{C}$ or $\check{\mathcal{C}}$, we must have $\mathsf{Res}[\mathsf{a}|\mathsf{A}; \mathcal{C}] = \mathsf{Res}[\check{\mathsf{a}}|\check{\mathsf{A}}; \check{\mathcal{C}}] = 0$. This immediately gives the desired inequality in (19). Moreover, if A and $\check{\mathsf{A}}$ have non-overlapping set of nonzero rows, $\mathsf{Res}\big[\mathsf{a}, \check{\mathsf{a}} | [\mathsf{A}, \check{\mathsf{A}}]; \mathcal{C} \sqcup \check{\mathcal{C}}\big]$ would be 1 if and only if either a increases the rank of A at column set $\mathcal{C}$, or $\check{\mathsf{a}}$ increases the rank of $\check{\mathsf{A}}$ at column set $\check{\mathcal{C}}$, or both, where in that case the reverse inequality in (19) also holds. This completes the proof.

Appendix B
Proof of Theorem 2

To simplify the proof, we first prove following two lemmas.

Lemma 10. *Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors in $\mathcal{L}$. Suppose that there is a matrix L such that $\mathbf{Y} = \mathsf{L}\mathbf{X}$. Then*

$$\lim_{q\to\infty} \frac{H(\langle \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q)}{\log_2(q)} = 0. \tag{68}$$

Proof: For a vector $\mathbf{V}$, we define $\Delta[\mathbf{V}] = \mathbf{V} - \langle \mathbf{V} \rangle_q$ as the vector of quantization residual. Let $\boldsymbol{\Delta}_1 = \Delta[\mathbf{X}]$, and $\boldsymbol{\Delta}_2 = \Delta[\mathsf{L}\langle \mathbf{X} \rangle_q]$. Then, we have

$$\begin{aligned}
\langle \mathbf{Y} \rangle_q = \langle \mathsf{L}\mathbf{X} \rangle_q &= \langle \mathsf{L}\langle \mathbf{X} \rangle_q + \mathsf{L}\boldsymbol{\Delta}_1 \rangle_q \\
&= \big\langle \langle \mathsf{L}\langle \mathbf{X} \rangle_q \rangle_q + \boldsymbol{\Delta}_2 + \mathsf{L}\boldsymbol{\Delta}_1 \big\rangle_q \\
&\overset{(i)}{=} \langle \mathsf{L}\langle \mathbf{X} \rangle_q \rangle_q + \langle \boldsymbol{\Delta}_2 + \mathsf{L}\boldsymbol{\Delta}_1 \rangle_q,
\end{aligned}$$

where in (i) we used the fact that $\langle \mathsf{L}\langle \mathbf{X} \rangle_q \rangle_q$ is already in the quantized form, so it can go outside the quantization operator. Since $\langle \mathsf{L}\langle \mathbf{X} \rangle_q \rangle_q$ is a function of $\langle \mathbf{X} \rangle_q$, we obtain that

$$H(\langle \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q) = H(\langle \boldsymbol{\Delta}_2 + \mathsf{L}\boldsymbol{\Delta}_1 \rangle_q | \langle \mathbf{X} \rangle_q). \tag{69}$$

Let $\boldsymbol{\Xi} = \boldsymbol{\Delta}_2 + \mathsf{L}\boldsymbol{\Delta}_1$. Applying the triangle inequality we have

$$\|\boldsymbol{\Xi}\|_\infty \leqslant \|\boldsymbol{\Delta}_2\|_\infty + \|\mathsf{L}\boldsymbol{\Delta}_1\|_\infty \leqslant \frac{1}{q}(1 + \|\mathsf{L}\|_{\infty,\infty}), \tag{70}$$

where $\|\mathsf{L}\|_{\infty,\infty} = \sup_{\mathbf{x}:\mathbf{x}\neq 0} \frac{\|\mathsf{L}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$ is the operator norm of L, and where $\frac{1}{q}$ results from the fact that every component of the quantization residual is always bounded by $\frac{1}{q}$. This implies that $\langle \boldsymbol{\Xi} \rangle_q$ can take at most $\lceil 1 + \|\mathsf{L}\|_{\infty,\infty} \rceil^n$ different values, thus, from (69), we have that $H(\langle \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q)$ is upper bounded, independent of the value of q, by $n \log_2(\lceil 1 + \|\mathsf{L}\|_{\infty,\infty} \rceil)$, where n denotes the dimension of the vector $\mathbf{Y}$, and where for a real number r, we define $\lceil r \rceil = \min\{n \in \mathbb{Z} : n \geqslant r\}$. Hence, dividing (69) by $\log_2(q)$ and taking the limit as q tends to infinity, we obtain the desired result. ∎

Remark 2. *Note that Lemma 10 still holds if we replace the quantized values $\langle \mathbf{X} \rangle_q$ in (68) by the unquantized random variables $\mathbf{X}$, or keep any mixture thereof.* ◊

Lemma 11. *Suppose that all the conditions of Lemma 10 hold, and let $\mathbf{Z}$ be another vector in $\mathcal{L}$ with the same dimension as $\mathbf{Y}$. Then*

$$\limsup_{q\to\infty} \frac{H(\langle \mathbf{Z} + \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q)}{\log_2(q)} = \limsup_{q\to\infty} \frac{H(\langle \mathbf{Z} \rangle_q | \langle \mathbf{X} \rangle_q)}{\log_2(q)},$$

with a similar equality holding for $\liminf$ *instead of* $\limsup$.

Proof: The proof follows from Lemma 10. We have

$$\begin{aligned}
H(\langle \mathbf{Z} + \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q) &= H(\langle \mathbf{Z} + \mathbf{Y} \rangle_q | \langle \mathbf{X} \rangle_q, \langle \mathbf{Z} \rangle_q) \\
&\quad + H(\langle \mathbf{Z} \rangle_q | \langle \mathbf{X} \rangle_q) - H(\langle \mathbf{Z} \rangle_q | \langle \mathbf{Z} + \mathbf{Y} \rangle_q, \langle \mathbf{X} \rangle_q).
\end{aligned} \tag{71}$$

Note that $\mathbf{Y}$ is a linear function of $\mathbf{X}$, thus, there are matrices L_1 and L_2 such that $\mathbf{Z}+\mathbf{Y} = \mathsf{L}_1[\mathbf{X}; \mathbf{Z}]$ and $\mathbf{Z} = \mathsf{L}_2[\mathbf{Z}+\mathbf{Y}; \mathbf{X}]$. Dividing both sides of (71) by $\log_2(q)$ and using Lemma 10 yields that the first and the last term on the right hand side of (71) tend to 0 as q tends to infinity. Thus, applying $\limsup$ and $\liminf$, we obtain the desired result. ∎

Using Lemma 11 and Remark 2, we now prove Theorem 2. For the first part, from the definition of RID in (12), we have

$$\begin{aligned}
H(\langle X_1^n \rangle_q) &\doteq H(\langle X_1^n \rangle_q | \mathcal{C}) && (72)\\
&\doteq H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}} + \mathsf{A}_{\mathcal{C}^c}\mathbf{D}_{\mathcal{C}^c} \rangle_q | \mathcal{C}) && (73)\\
&\doteq H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}} + \mathsf{A}_{\mathcal{C}^c}\mathbf{D}_{\mathcal{C}^c} \rangle_q | \mathcal{C}, \mathbf{D}_{\mathcal{C}^c}) && (74)\\
&\doteq H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}} \rangle_q | \mathcal{C}, \mathbf{D}_{\mathcal{C}^c}) && (75)\\
&= \mathbb{E}_{\mathcal{C}}\Big\{ H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}} \rangle_q | \mathcal{C} = \mathcal{C}^*, \mathbf{D}_{\mathcal{C}^c}) \Big\} && (76)\\
&= \mathbb{E}_{\mathcal{C}}\Big\{ H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}} \rangle_q | \mathcal{C} = \mathcal{C}^*) \Big\} && (77)
\end{aligned}$$

where $\mathcal{C}$ is the support set defined by (20) and $\mathbf{C} = C_1^k$ and $\mathbf{D} = D_1^k$ are the vector of continuous and discrete parts of Z_1^k as defined in Section IV-B, and where in (72), (74), and (75), we use the fact that $\mathcal{C}$ and $\mathbf{D}_{\mathcal{C}^c}$ are discrete variables with a finite entropy independent of q and they can be arbitrarily added or removed from the conditioning part of the entropy. Recall that from our notation in Section I-A, $f_q \doteq h_q$ for two sequences f, h (parametrized with $q \in \mathbb{N}$), whenever $\limsup_{q\to\infty} \frac{f_q - h_q}{\log_2(q)} = 0$. For (75), we used Lemma 11 to remove the variable $\mathsf{A}_{\mathcal{C}^c}\mathbf{D}_{\mathcal{C}^c}$ in (74) since it is a linear function

of $\mathbf{D}_{\mathcal{C}^c}$ appearing in the conditioning part. For (77), we used the independence of $\mathbf{D}_{\mathcal{C}^c}$ from $\mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}}$ conditioned on $\mathcal{C}=\mathcal{C}^*$.

Now, consider a specific realization of the support set $\mathcal{C}=\mathcal{C}^*$ of size $k^*=|\mathcal{C}^*|$ and notice that since $\mathcal{C}$ is independent of the continuous component $\mathbf{C}$, conditioning on $\mathcal{C}=\mathcal{C}^*$ does not change the distribution of $\mathbf{C}_{\mathcal{C}^*}$, which is a k^*-dimensional continuous distribution. Let $m^*=\operatorname{rank}(\mathsf{A}_{\mathcal{C}^*})$ and let $\check{\mathsf{A}}$ be a maximal submatrix of $\mathsf{A}_{\mathcal{C}^*}$ consisting of linearly independent rows of $\mathsf{A}_{\mathcal{C}^*}$. Then, we have

$$H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q)=H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q,\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q) \tag{78}$$

$$=H(\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q)+H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q). \tag{79}$$

Note that the $\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}$ has a well-defined m^*-dimensional continuous distribution, thus, $\lim_{q\to\infty}\frac{H(\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q)}{\log_2(q)}=m^*$. Moreover, from Lemma 11, the second term in (79) vanishes in the limit when divided by $\log_2(q)$ because $\mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}$ is a linear function of $\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}$. Thus, from (77), we obtain

$$d(X_1^n)=\lim_{q\to\infty}\frac{H(\langle X_1^n\rangle_q)}{\log_2(q)} \tag{80}$$

$$=\lim_{q\to\infty}\frac{\mathbb{E}_{\mathcal{C}}\Big[H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}}\rangle_q|\mathcal{C}=\mathcal{C}^*)\Big]}{\log_2(q)} \tag{81}$$

$$=\mathbb{E}_{\mathcal{C}}\Big[\lim_{q\to\infty}\frac{H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}}\rangle_q|\mathcal{C}=\mathcal{C}^*)}{\log_2(q)}\Big] \tag{82}$$

$$=\mathbb{E}_{\mathcal{C}}\Big[\operatorname{rank}(\mathsf{A}_{\mathcal{C}})\Big], \tag{83}$$

where in (81), we used the fact that $\mathcal{C}$ takes only finitely many values and exchanged the expectation and the limit. This completes the proof of the first part of the theorem.

To prove the second part, recall that $X_1^n=\mathsf{A}Z_1^k$ and $Y_1^m=\mathsf{B}Z_1^k$. We follow similar steps as in the first part, where from (77), we essentially need to compute the expression

$$H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}}\rangle_q|\mathsf{B}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}},\mathcal{C}=\mathcal{C}^*) \tag{84}$$

for different realizations of support set $\mathcal{C}^*$. Note that $\mathsf{B}_{\mathcal{C}^*}$ is not necessarily full-rank. Let $\check{\mathsf{B}}$ be a maximal submatrix of $\mathsf{B}_{\mathcal{C}^*}$ consisting of those rows that are linearly independent. Also, let $\check{\mathsf{A}}$ be the maximal submatrix of $\mathsf{A}_{\mathcal{C}^*}$ consisting of those linearly independent rows that are also linearly independent of the rows of $\check{\mathsf{B}}$. From the definition of Res operator in (15), it is not difficult to check that the number of rows of $\check{\mathsf{A}}$ is given by $\mathsf{Res}[\mathsf{A}|\mathsf{B};\mathcal{C}^*]$. Hence, we have

$$H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\mathsf{B}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*})=H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q,\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*})$$

$$=H(\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*}) \tag{85}$$

$$+H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q,\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*}). \tag{86}$$

Note that $[\check{\mathsf{A}};\check{\mathsf{B}}]$ is a full rank matrix, thus, $[\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*};\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*}]$ has a well-defined continuous distribution. In particular, for almost all realizations of $\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*}$, the random variable $\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}$ has a continuous distribution, thus, for the first term in (85) we have that $\lim_{q\to\infty}\frac{H(\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*})}{\log_2(q)}=\dim(\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*})$, which is equal to $\mathsf{Res}[\mathsf{A}|\mathsf{B};\mathcal{C}^*]$. For the second term in (86), we have that $\lim_{q\to\infty}\frac{H(\langle \mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}\rangle_q|\langle \check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}\rangle_q,\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*})}{\log_2(q)}=0$ since $\mathsf{A}_{\mathcal{C}^*}\mathbf{C}_{\mathcal{C}^*}$ is a linear function of $\check{\mathsf{A}}\mathbf{C}_{\mathcal{C}^*}$ and $\check{\mathsf{B}}\mathbf{C}_{\mathcal{C}^*}$ appearing in the conditioning part, and the result follows from Lemma 10 and Remark 2. Thus, we have

$$\lim_{q\to\infty}\frac{H(\langle \mathsf{A}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}}\rangle_q|\mathsf{B}_{\mathcal{C}}\mathbf{C}_{\mathcal{C}},\mathcal{C}=\mathcal{C}^*)}{\log_2(q)}=\mathsf{Res}[\mathsf{A}|\mathsf{B};\mathcal{C}^*], \tag{87}$$

and taking the average over $\mathcal{C}$, we obtain the desired result.

Appendix C
Proof of Theorem 3

We use the rank characterization of the RID proved in Theorem 2. The positivity simply follows from the positivity of the rank of a matrix. Also, if $d(X_1^n)=0$, from the property of the rank and the definition of X_1^n in (11), we have that $0\leqslant d(X_i)\leqslant d(X_1^n)=0$ for every $i\in[n]$, which implies that all the X_i, $i\in[n]$, are discrete variables. The invariance results from the fact that for any $n\times k$ matrix A, any invertible $n\times n$ matrix L, and any subset $\mathcal{C}\subseteq[k]$ of columns of A, we have $\operatorname{rank}(\mathsf{A}_{\mathcal{C}})=\operatorname{rank}(\mathsf{L}\mathsf{A}_{\mathcal{C}})$. The chain rule follows from the chain rule for the Res operator in Proposition 1:

$$d(X_1^n,Y_1^m)=\mathbb{E}\big[\mathsf{Res}\big[[\mathsf{A};\mathsf{B}];\mathcal{C}\big]\big] \tag{88}$$

$$=\mathbb{E}\big[\mathsf{Res}[\mathsf{A};\mathcal{C}]\big]+\mathbb{E}\big[\mathsf{Res}[\mathsf{B}|\mathsf{A};\mathcal{C}]\big] \tag{89}$$

$$=d(X_1^n)+d(Y_1^m|X_1^n). \tag{90}$$

To prove the symmetry, note that $\mathsf{RI}(X_1^n;Y_1^m)=d(X_1^n)-d(X_1^n|Y_1^m)$. Thus, using the chain rule property, we obtain that $\mathsf{RI}(X_1^n;Y_1^m)=d(X_1^n)+d(Y_1^m)-d(X_1^n,Y_1^m)$, and the symmetry follows from the symmetry of $d(X_1^n,Y_1^m)$.

Finally, for the last part note that from the definition of the Res operator, it results that $\operatorname{rank}(\mathsf{A}_{\mathcal{C}})=\mathsf{Res}[\mathsf{A}|\varnothing;\mathcal{C}]\geqslant\mathsf{Res}[\mathsf{A}|\mathsf{B};\mathcal{C}]$. Taking the average over $\mathcal{C}$, we have $d(X_1^n)\geqslant d(X_1^n|Y_1^m)$, which implies the desired result.

Appendix D
Proof of Theorem 7

Note that U_1^N are i.i.d. random variables obtained via a linear transform of the variables in $\mathcal{L}$, thus, they belong to $\mathcal{L}$. Hence, from Theorem 4, it immediately results that I_n is an erasure process with initial value $I_0(1)=d(U_1)$ polarizing to $\{0,1\}$.

To prove that J_n is also a polarizing erasure process, first note that the recursive structure in (33) remains intact if we transform h_i into $\mathsf{h}_i\otimes\mathsf{b}$ and $\mathbb{H}_N$ into $\mathbb{H}_N\otimes\mathsf{b}$, where $\otimes$ denotes the Kronecker product. Moreover, it is not difficult to see that there are indeed two main ingredients in the proof of Theorem 4: Applying the recursive structure of $\mathbb{H}_N$ as in (38), in order to obtain an expression for the plus-branch ($I_n(i)\to I_n(i)^+$), and using the chain rule in (46), in order to compute the minus-branch ($I_n(i)\to I_n(i)^-$). It is not difficult to check that both conditions remain valid after the mentioned transformation. This implies that J_n is also an erasure process, whose initial value, from chain rule, is given by $J_0(1)=d(V_1|U_1)$.

Appendix E
Proof of Lemma 9

Suppose that given $Y=y$, we use the MAP decoder defined by $\widehat{D}(y)=\arg\max_{d'\in\mathcal{D}}p(d'|y)$ to decode D. Then, for any

arbitrary $d \in \mathcal{D}$, we have

$$\mathbb{P}[\mathcal{E}|Y=y] = 1 - \max_{d' \in \mathcal{D}} p(d'|y) \tag{91}$$

$$\leqslant 1 - p(d|y) = 1 - e^{\log(p(d|y))}. \tag{92}$$

Taking the average over the joint distribution of (D, Y) and using the Jensen's inequality [37] for the convex function $u \mapsto e^u$, we have

$$\mathbb{P}[\mathcal{E}] \leqslant 1 - e^{\mathbb{E}_{D,Y}[\log(p(d|y))]} = 1 - e^{-H(D|Y)\log_2(e)} \tag{93}$$

$$\overset{(i)}{\leqslant} 1 - (1 - H(D|Y)\log_2(e)) \tag{94}$$

$$= H(D|Y)\log_2(e), \tag{95}$$

where in (i) we used the inequality $e^{-u} \geqslant 1 - u$ for $u \in \mathbb{R}$. This completes the proof.

REFERENCES

[1] S. Haghighatshoar and E. Abbe, "Polarization of the rényi information dimension for single and multi terminal analog compression," *arXiv preprint arXiv:1301.6388*, 2013.

[2] ——, "Polarization of the rényi information dimension for single and multi terminal analog compression," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 779–783.

[3] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 1-2, pp. 193–215, 1959.

[4] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.

[5] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1564–1572, 1994.

[6] K. Falconer, *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[7] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3748, 2010.

[8] G. Alberti, H. Bölcskei, C. De Lellis, G. Koliander, and E. Riegler, "Lossless linear analog compression," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 2789–2793.

[9] D. Stotz, E. Riegler, E. Agustsson, and H. Bolcskei, "Almost lossless analog signal separation and probabilistic uncertainty relations," *IEEE Transactions on Information Theory*, 2017.

[10] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7434–7464, 2013.

[11] L. Li, H. Mahdavifar, and I. Kang, "A structured construction of optimal measurement matrix for noiseless compressed sensing via analog polarization," *arXiv preprint arXiv:1212.5577*, 2012.

[12] S. Haghighatshoar, E. Abbe, and E. Telatar, "Adaptive sensing using deterministic partial hadamard matrices," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2012, pp. 1842–1846.

[13] Y. Wu, S. Shamai, and S. Verdú, "Degrees of freedom of the interference channel: A general formula." in *ISIT*, 2011, pp. 1362–1366.

[14] D. Stotz and H. Bolcskei, "Degrees of freedom in vector interference channels," in *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2012*. IEEE, 2012, pp. 1755–1760.

[15] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[16] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[17] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[18] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[19] ——, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[20] E. Arikan, "Source polarization," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2010, pp. 899–903.

[21] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1751–1768, 2010.

[22] E. Abbe and E. Telatar, "Polar codes for the-user multiple access channel," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5437–5448, 2012.

[23] S. B. Korada and R. Urbanke, "Polar codes for slepian-wolf, wyner-ziv, and gelfand-pinsker," in *Information Theory Workshop (ITW), 2010 IEEE*. IEEE, 2010, pp. 1–5.

[24] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6428–6443, 2011.

[25] E. Sasoglu and A. Vardy, "A new polar coding scheme for strong security on wiretap channels," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1117–1121.

[26] Y. Yan, C. Ling, and X. Wu, "Polar lattices: where arıkan meets forney," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1292–1296.

[27] E. Abbe and A. Barron, "Polar coding schemes for the awgn channel," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 194–198.

[28] S. Haghighatshoar, E. Abbe, and E. Telatar, "A new entropy power inequality for integer-valued random variables," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2013, pp. 589–593.

[29] P. R. Halmos, *Measure theory*. Springer, 2013, vol. 18.

[30] K. L. Chung, *A course in probability theory*. Academic press, 2001.

[31] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[32] E. Abbe and Y. Wigderson, "High-girth matrices and polarization," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 2461–2465.

[33] E. Arikan and I. Telatar, "On the rate of channel polarization," in *IEEE International Symposium on Information Theory (ISIT)*, 2009, pp. 1493–1495.

[34] E. A. Bilkent, "Polar coding for the slepian-wolf problem based on monotone chain rules," in *2012 IEEE International Symposium on Information Theory Proceedings*, 2012.

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

[36] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[37] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta mathematica*, vol. 30, no. 1, pp. 175–193, 1906.