Varentropy Decreases Under the Polar Transform

Erdal Arıkan, *Fellow, IEEE*

***Abstract*— We consider the evolution of variance of entropy (varentropy) in the course of a polar transform operation on binary data elements (BDEs). A BDE is a pair (X, Y) consisting of a binary random variable X and an arbitrary side information random variable Y. The varentropy of (X, Y) is defined as the variance of the random variable $-\log p_{X|Y}(X|Y)$. A polar transform of order two is a certain mapping that takes two independent BDEs and produces two new BDEs that are correlated with each other. It is shown that the sum of the varentropies at the output of the polar transform is less than or equal to the sum of the varentropies at the input, with equality if and only if at least one of the inputs has zero varentropy. This result is extended to polar transforms of higher orders and it is shown that the varentropy asymptotically decreases to zero when the BDEs at the input are independent and identically distributed.**

***Index Terms*— Polar coding, varentropy, dispersion.**

I. Introduction

WE USE the term "varentropy" as an abbreviation for "variance of the conditional entropy random variable" following the usage in [1]. In his pioneering work, Strassen [2] showed that the varentropy is a key parameter for estimating the performance of optimal block-coding schemes at finite (non-asymptotic) block-lengths. More recently, the comprehensive work by Polyanskiy *et al.* [3] further elucidated the significance of varentropy (under the name "dispersion") and rekindled interest in the subject. In this paper, we study varentropy in the context of polar coding. Specifically, we track the evolution of average varentropy in the course of polar transformation of independent identically distributed (i.i.d.) BDEs and show that it decreases to zero asymptotically as the transform size increases. As a side result, we obtain an alternative derivation of the polarization results of [4] and [5].

A. Notation and Basic Definitions

Our setting will be that of binary-input memoryless channels and binary memoryless sources. We treat source and channel coding problems in a common framework by using the neutral term "binary data element" (BDE) to cover both. Formally, a BDE is any pair of random variables (X, Y) where X takes values over $\mathcal{X} \stackrel{\Delta}{=} \{0, 1\}$ (not necessarily from the uniform distribution) and Y takes values over some alphabet $\mathcal{Y}$ which may be discrete or continuous. A BDE (X, Y) may represent, in a source-coding setting, a binary data source X that we wish to compress in the presence of some side information Y; or, it may represent, in a channel-coding setting, a channel with input X and output Y.

Given a BDE (X, Y), the information measures of interest in the sequel will be the *conditional entropy random variable*

$$h(X|Y) \stackrel{\Delta}{=} -\log p_{X|Y}(X|Y),$$

the *conditional entropy*

$$H(X|Y) \stackrel{\Delta}{=} \mathbb{E}\, h(X|Y),$$

and, the *varentropy*

$$\mathrm{V}(X|Y) \stackrel{\Delta}{=} \mathrm{Var}(h(X|Y)).$$

Throughout the paper, we use base-two logarithms.

The term *polar transform* is used in this paper to to refer to an operation that takes two *independent* BDEs (X_1, Y_1) and (X_2, Y_2) as input, and produces two new BDEs $(U_1, \mathbf{Y})$ and $(U_2; U_1, \mathbf{Y})$ as output, where $U_1 \stackrel{\Delta}{=} X_1 \oplus X_2$, $U_2 \stackrel{\Delta}{=} X_2$, and $\mathbf{Y} \stackrel{\Delta}{=} (Y_1, Y_2)$. The notation "$\oplus$" denotes modulo-2 addition.

B. Polar Transform and Varentropy

The main result of the paper is the following.

Theorem 1: The varentropy is nonincreasing under the polar transform in the sense that, if (X_1, Y_2), (X_2, Y_2) are any two independent BDEs at the input of the transform and $(U_1, \mathbf{Y})$, $(U_2; U_1, \mathbf{Y})$ are the BDEs at its output, then

$$\mathrm{V}(U_1|\mathbf{Y}) + \mathrm{V}(U_2|U_1, \mathbf{Y}) \leq \mathrm{V}(X_1|Y_1) + \mathrm{V}(X_2|Y_2), \quad (1)$$

with equality if and only if (iff) either $V(X_1|Y_1) = 0$ or $V(X_2|Y_2) = 0$.

For an alternative formulation of the main result, let us introduce the following notation:

$$h_{\mathrm{in},1} \stackrel{\Delta}{=} h(X_1|Y_1), \qquad h_{\mathrm{in},2} \stackrel{\Delta}{=} h(X_2|Y_2), \quad (2)$$

$$h_{\mathrm{out},1} \stackrel{\Delta}{=} h(U_1|\mathbf{Y}), \qquad h_{\mathrm{out},2} \stackrel{\Delta}{=} h(U_2|U_1, \mathbf{Y}). \quad (3)$$

Theorem 1 can be reformulated as follows.

Theorem 1′: The polar transform of conditional entropy random variables, $(h_{in,1}, h_{in,2}) \to (h_{out,1}, h_{out,2})$, produces positively correlated output entropy terms in the sense that

$$\mathrm{Cov}(h_{out,1}, h_{out,2}) \geq 0, \quad (4)$$

with equality iff either $\mathrm{Var}(h_{in,1}) = 0$ or $\mathrm{Var}(h_{in,2}) = 0$.

Manuscript received August 28, 2014; revised October 30, 2015; accepted March 24, 2016. Date of publication April 21, 2016; date of current version May 18, 2016. This work was supported in part by the Simons Institute for Theory of Computing, UC Berkeley, and in part by the Directorate-General for Research and Innovation within the European Commission Seventh Framework Programme Network of Excellence in Wireless Communications under Grant 318306.

The author is with the Department of Electrical-Electronics Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: arikan@ee.bilkent.edu.tr).

Communicated by H. Pfister, Associate Editor for Coding Theory.

Digital Object Identifier 10.1109/TIT.2016.2555841

0018-9448 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This second form makes it clear that any reduction in varentropy can be attributed entirely to the creation of a positive correlation between the entropy random variables $h_{\text{out},1}$ and $h_{\text{out},2}$ at the output of the polar transform.

Showing the equivalence of the two claims (1) and (4) is a simple exercise. We have, by the chain rule of entropy,

$$h_{\text{out},1} + h_{\text{out},2} = h_{\text{in},1} + h_{\text{in},2}; \tag{5}$$

hence, $\text{Var}(h_{\text{out},1} + h_{\text{out},2}) = \text{Var}(h_{\text{in},1} + h_{\text{in},2})$. Since $h_{\text{in},1}$ and $h_{\text{in},2}$ are independent, $\text{Var}(h_{\text{in},1} + h_{\text{in},2}) = \text{Var}(h_{\text{in},1}) + \text{Var}(h_{\text{in},2})$; while $\text{Var}(h_{\text{out},1} + h_{\text{out},2}) = \text{Var}(h_{\text{out},1}) + \text{Var}(h_{\text{out},2}) + 2\,\text{Cov}(h_{\text{out},1}, h_{\text{out},2})$. Thus, the claim (1), which can be written in the equivalent form

$$\text{Var}(h_{\text{out},1}) + \text{Var}(h_{\text{out},2}) \leq \text{Var}(h_{\text{in},1}) + \text{Var}(h_{\text{in},2}),$$

is true iff (4) holds.

A technical question that arises in the sequel is whether the varentropy is uniformly bounded across the class of all BDEs. This is indeed the case.

Lemma 1: For any BDE (X, Y), $V(X|Y) \leq 2.2434$.

Proof: It suffices to show that the second moment of $h(X|Y)$ satisfies the given bound.

$$\begin{aligned} E[h(X|Y)^2] &\leq \max_{0\leq x\leq 1} [x \log^2(x) + (1-x)\log^2(1-x)] \\ &\leq 2 \max_{0\leq x\leq 1} [x \log^2(x)] = 8e^{-2}\log^2(e) \approx 2.2434. \end{aligned}$$

(A numerical study shows that a more accurate bound on $V(X|Y)$ is 1.1716, but the present bound will be sufficient for our purposes.) □

This bound guarantees that all varentropy terms in this paper exist and are bounded; it also guarantees the existence of the covariance terms since by the Cauchy-Schwarz inequality we have $|\,\text{Cov}(h_{\text{out},1}, h_{\text{out},2})| \leq \sqrt{\text{Var}(h_{\text{out},1})\,\text{Var}(h_{\text{out},2})}$.

We will end this part by giving two examples in order to illustrate the behavior of varentropy under the polar transform. The terminology in both examples reflects a channel coding viewpoint; although, each model may also arise in a source coding context.

Example 1: In this example, (X, Y) *models a binary symmetric channel (BSC) with equiprobable inputs and a crossover probability* $0 \leq \epsilon \leq 1/2$*; in other words,* X *and* Y *take values in the set* $\{0, 1\}$ *with*

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{2}(1-\epsilon), & \text{if } x = y; \\ \frac{1}{2}\epsilon, & \text{if } x \neq y. \end{cases}$$

Fig. 1 gives a sketch of the varentropy and covariance terms defined above, with $\text{Var}(h_{in})$ *denoting the common value of* $\text{Var}(h_{in,1})$ *and* $\text{Var}(h_{in,2})$*). (Formulas for computing the varentropy terms will be given later in the paper.) The non-negativity of the covariance is an indication that the varentropy is reduced by the polar transform.*

Example 2: Here, (X, Y) *represents a binary erasure channel (BEC) with equiprobable inputs and an erasure probability* ϵ*. In other words,* X *takes values in* $\{0, 1\}$*,* Y *takes values in* $\{0, 1, 2\}$*, and*

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{2}(1-\epsilon), & \text{if } x = y; \\ \frac{1}{2}\epsilon, & \text{if } y = 2. \end{cases}$$

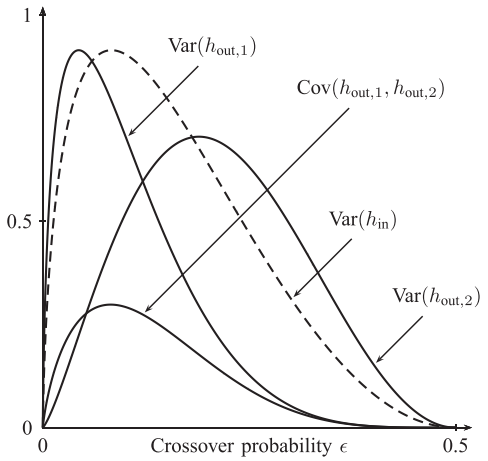

Fig. 1. Variance and covariance of entropy for BSC under polar transform.

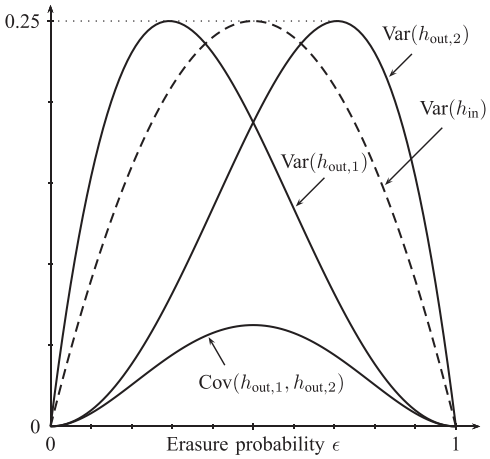

Fig. 2. Variance and covariance of entropy for BEC under polar transform.

In this case, there exist simple formulas for the varentropies. $\text{Var}(h_{in,1}) = \text{Var}(h_{in,2}) = \text{Var}(h_{in}) = \epsilon(1-\epsilon)$,$\text{Var}(h_{out,2}) = (2\epsilon - \epsilon^2)(1-\epsilon)^2$,$\text{Var}(h_{out,1}) = \epsilon^2(1-\epsilon^2)$. *The covariance is given by* $\text{Cov}(h_{out,1}, h_{out,2}) = \epsilon^2(1-\epsilon)^2$. *The corresponding curves are plotted in Fig. 2.*

C. Organization

The rest of the paper is organized as follows. In Section II, we define two canonical representations for a BDE (X, Y) that eliminate irrelevant details from problem description and simplify the analysis. In Section III, we review some basic facts about the covariance function that are needed in the remainder of the paper. Section IV contains the proof of Theorem $1'$. Section V considers the behavior of varentropy under higher-order polar transforms and contains a self-contained proof of the main polarization result of [4].

Throughout, we will often write $\overline{p}$ to denote $1 - p$ for a real number $0 \leq p \leq 1$. For $0 \leq p, q \leq 1$, we will write $p * q$ to denote the convolution $pq + \overline{p}\,\overline{q}$.

II. Canonical Representations

The information measures of interest relating to a given BDE (X, Y) are determined solely by the joint probability

distribution of (X, Y); the specific forms of the alphabets $\mathcal{X}$ and $\mathcal{Y}$ play no role. We have already fixed $\mathcal{X}$ as $\{0, 1\}$ so as to have a standard representation for X. It is possible and desirable to re-parametrize the problem, if necessary, so that $\mathcal{Y}$ also has a canonical form. Such canonical representations have been given for Binary Memoryless Symmetric (BMS) channels in [6]. The class of BDEs (X, Y) under consideration here is more general than the class of BMS channels, but similar ideas apply. We will give two canonical representations for BDEs, which we will call the α-representation and the β-representation. The α-representation replaces $\mathcal{Y}$ with a canonical alphabet $\mathcal{A} \subset [0, 1]$, and has the property of being "lossless". The β-representation replaces $\mathcal{Y}$ with $\mathcal{B} \subset [0, 1/2]$; it is "lossy", but happens to be more convenient than the α-representation for purposes of proving Theorem $1'$.

A. The α-Representation

Given a BDE (X, Y), we associate to each $y \in \mathcal{Y}$ the parameter

$$\alpha(y) = \alpha_{X|Y}(y) \stackrel{\Delta}{=} p_{X|Y}(0|y)$$

and define $A \stackrel{\Delta}{=} \alpha(Y)$. The random variable A takes values in the set $\mathcal{A} \stackrel{\Delta}{=} \{\alpha(y) : y \in \mathcal{Y}\}$, which is always a subset of $[0, 1]$. We refer to A as the α-representation of (X, Y). The α-representation provides economy by using a canonical alphabet $\mathcal{A}$ in which any two symbols $y, y' \in \mathcal{Y}$ are merged into a common symbol a whenever $\alpha(y) = \alpha(y') = a$.

We give some examples to illustrate the α-representation. For the BSC of Example 1, we have $\alpha(0) = 1-\epsilon$, $\alpha(1) = \epsilon$, $\mathcal{A} = \{\epsilon, 1-\epsilon\}$. In the case of the BEC of Example 2, we have $\alpha(0) = 1$, $\alpha(1) = 0$, $\alpha(2) = 1/2$, $\mathcal{A} = \{0, 1/2, 1\}$. As a third example, consider the channel $y = (-1)^x c + z$ where $c > 0$ is a constant and $z \sim N(0, 1)$ is a zero-mean unit-variance additive Gaussian noise, independent of x. In this case, we have

$$\alpha(y) = \frac{e^{-(y-c)^2/2}}{e^{-(y-c)^2/2} + e^{-(y+c)^2/2}} = \frac{1}{1 + e^{-2cy}},$$

giving $\mathcal{A} = (0, 1)$.

The α-representation provides "sufficient statistics" for computing the information measures of interest to us. To illustrate this, let (X, Y) be an arbitrary BDE and let $A = \alpha(Y)$ be its α-representation. Let F_A denote the cumulative distribution function (CDF) of A.

The conditional entropy random variable is given by

$$h(X|Y) = h(X|A) = \begin{cases} -\log A, & X = 0; \\ -\log \overline{A}, & X = 1. \end{cases} \tag{6}$$

Hence, the conditional entropy can be calculated as

$$\begin{aligned} H(X|Y) &= \mathbb{E}\, h(X|Y) = \mathbb{E}\, h(X|A) = \mathbb{E}_A \mathbb{E}_{X|A}\, h(X|A) \\ &= \mathbb{E}_A \mathcal{H}(A) = \mathbb{E}\, \mathcal{H}(A) = \int_0^1 \mathcal{H}(a)\, \mathrm{d}F_A(a), \end{aligned} \tag{7}$$

where $\mathcal{H}(a) \stackrel{\Delta}{=} -a \log a - \overline{a} \log \overline{a}$, $a \in [0, 1]$, is the binary entropy function. Likewise, the varentropy is given by

$$V(X|Y) = V(X|A) = \mathbb{E}\, \mathcal{H}_2(A) - \big[\mathbb{E}\, \mathcal{H}(A)\big]^2, \tag{8}$$

where $\mathcal{H}_2(a) \stackrel{\Delta}{=} -a \log^2 a - \overline{a} \log^2 \overline{a}$ and

$$\mathbb{E}\, \mathcal{H}_2(A) = \int_0^1 \mathcal{H}_2(a)\, \mathrm{d}F_A(a).$$

Finally, we note that $H(X) = \mathcal{H}(p_X(0)) = \mathcal{H}(\mathbb{E}\, A)$. Thus, all information measures of interest in this paper can be computed given knowledge of the distribution of A.

B. The β-Representation

Although the α-representation eliminates much of the irrelevant detail from (X, Y), there is need for an even more compact representation for the type of problems considered in the sequel. This more compact representation is obtained by associating to each $y \in \mathcal{Y}$ the parameter

$$\beta(y) = \beta_{X|Y}(y) \stackrel{\Delta}{=} \min\{p_{X|Y}(0|y), p_{X|Y}(1|y)\}.$$

We define the β-representation of (X, Y) as the random variable $B \stackrel{\Delta}{=} \beta(Y)$. We denote the range of B by $\mathcal{B} \stackrel{\Delta}{=} \{\beta(y) : y \in \mathcal{Y}\}$ and note that $\mathcal{B} \subset [0, 1/2]$.

The β-representation can be obtained from the α-representation by

$$\beta(y) = \min\{\alpha(y), 1 - \alpha(y)\}, \quad B = \min\{A, \overline{A}\};$$

but, in general, the α-representation cannot be recovered from the β-representation.

For the BSC of Example 1, we have $\beta(0) = \beta(1) = \epsilon$, giving $\mathcal{B} = \{\epsilon\}$. For the BEC of Example 2, we have $\beta(0) = \beta(1) = 0$, $\beta(2) = 1/2$, and $\mathcal{B} = \{0, 1/2\}$. For the binary-input additive Gaussian noise channel, we have

$$\beta(y) = \frac{1}{1 + e^{2c|y|}},$$

with $\mathcal{B} = (0, 1/2]$.

As it is evident from (6), the conditional entropy random variable $h(X|Y)$ cannot be expressed as a function of (X, B). However, if the CDF F_B of B is known, we can compute $H(X|Y)$ and $\mathrm{V}(X|Y)$ by the following formulas that are analogous to (7) and (8):

$$H(X|Y) = \mathbb{E}\, \mathcal{H}(B), \quad V(X|Y) = \mathbb{E}\, \mathcal{H}_2(B) - \big[\mathbb{E}\, \mathcal{H}(B)\big]^2.$$

To see that B is less than a "sufficient statistic" for information measures, one may note that $H(X)$ is not determined by knowledge of F_B alone. For example, for a BDE (X, Y) with $\Pr(Y = X) = 1$, we have $\Pr(B = 0) = 1$, independently of $p_X(0)$.

Despite its shortcomings, the β-representation will be useful for our purposes due to the fact that the binary entropy function $\mathcal{H}(p)$ is monotone over $p \in [0, 1/2]$ but not over $p \in [0, 1]$. Thus, the random variable $\mathcal{H}(B)$ is a monotone function of B over the range of B, but $\mathcal{H}(A)$ is not necessary so over the range of A. This monotonicity will be important in proving certain correlation inequalities later in the paper.

TABLE I
CLASSIFICATION OF BDES

Type	Property
pure	$P(B=b)=1$ for some $b \in [0,1/2]$
extreme	$P(B=0)=1$ or $P(B=1/2)=1$
perfect	$P(B=0)=1$
purely random (p.r.)	$P(B=1/2)=1$
erasing	$P(B=0)+P(B=1/2)=1$

C. Classification of Binary Data elements

Table I gives a classification of a BDE (X, Y) in terms of the properties of $B = \beta(Y)$. The classification allows an erasing BDE to be extreme as a special case.

For a pure (X, Y), we obtain from (7) and (8) that

$$H(X|Y) = \mathcal{H}(b), \quad V(X|Y) = b(1-b)\log^2\left(\frac{b}{1-b}\right),$$

where b is the value that $B = \beta(Y)$ takes with probability 1. A simple corollary to this is the following characterization of an extreme BDE.

Proposition 1: Let (X, Y) be a BDE and $B = \beta(Y)$. The following three statements are equivalent: (i) (X, Y) is extreme, (ii) $H(X|Y) = 0$ or $H(X|Y) = 1$, (iii) $\mathrm{V}(X|Y) = 0$.

We omit the proof since it is immediate from the above formulas for $H(X|Y)$ and $V(X|Y)$ for a pure BDE.

For an erasing (X, Y), it is easily seen that

$$H(X|Y) = p, \quad V(X|Y) = p(1-p)$$

where $p = P[\beta(Y) = 1/2]$ is the *erasure* probability.

Parenthetically, we note that while the entropy function satisfies $H(X|Y) \leq H(X)$, there is no such general relationship between $V(X|Y)$ and $V(X)$. For an erasing (X, Y) with $p_X(1) = 1 - p_X(0) = q$ and erasure probability p, we have $V(X) = q(1-q)\log^2[q/(1-q)]$ while $V(X|Y) = p(1-p)$. Either $V(X) < V(X|Y)$ or $V(X) > V(X|Y)$ is possible depending on q and p.

D. Canonical Representations Under Polar Transform

In this part, we explore how the α- and β-representations evolve as they undergo a polar transform. Let us return to the setting of Sect. I-B. Let $(U_1, \mathbf{Y})$ and $(U_2; U_1, \mathbf{Y})$ denote the two BDEs obtained from a pair of independent BDEs (X_1, Y_1) and (X_2, Y_2) by the polar transform. Let $h_{\mathrm{in},1}$, $h_{\mathrm{in},2}$, $h_{\mathrm{out},1}$, and $h_{\mathrm{out},2}$ denote the entropy random variables at the input and output of the polar transform. For $i = 1, 2$, let $A_{\mathrm{in},i}$ and $B_{\mathrm{in},i}$ be the α- and β-representations for the ith BDE at the input side; and let $A_{\mathrm{out},i}$ and $B_{\mathrm{out},i}$ be those for the ith BDE at the output side. Let the sample values of these variables be denoted by small-case letters, such as $a_{\mathrm{in},i}$ for $A_{\mathrm{in},i}$, $b_{\mathrm{in},i}$ for $B_{\mathrm{in},i}$, etc.

Proposition 2: The α-parameters at the input and output of a polar transform are related by

$$A_{out,1} = A_{in,1} * A_{in,2}, \tag{9}$$

$$A_{out,2} = \begin{cases} A_{in,1}A_{in,2}/(A_{in,1} * A_{in,2}), & U_1 = 0; \\ \overline{A}_{in,1}A_{in,2}/(\overline{A}_{in,1} * A_{in,2}), & U_1 = 1. \end{cases} \tag{10}$$

Remark 1: In (10), the event $\{A_{in,1} * A_{in,2} = 0\}$ leads to an indeterminate form $A_{out,2} = 0/0$, but the conditional probability of $\{A_{in,1} * A_{in,2} = 0\}$ given $\{U_1 = 0\}$ is zero: $A_{in,1} * A_{in,2} = 0$ implies $(A_{in,1}, A_{in,2}) \in \{(0,1),(1,0)\}$, which in turn implies $(X_1, X_2) \in \{(1,0),(0,1)\}$, giving $U_1 = 1$. Similarly, the event $\{\overline{A}_{in,1} * A_{in,2} = 0\}$ is incompatible with $\{U_1 = 1\}$.

Proof: For a fixed $\mathbf{Y} = (y_1, y_2)$, the sample values of $A_{\mathrm{out},1}$ are given by

$$\begin{aligned} a_{\mathrm{out},1}(y_1, y_2) &\stackrel{\Delta}{=} p_{U_1|Y_1,Y_2}(0|y_1, y_2) \\ &= \sum_{u_2} p_{U_1,U_2|Y_1,Y_2}(0, u_2|y_1, y_2) \\ &= \sum_{u_2} p_{X_1|Y_1}(u_2|y_1)\, p_{X_2|Y_2}(u_2|y_2) \\ &= a_{\mathrm{in},1}(y_1) * a_{\mathrm{in},2}(y_2). \end{aligned}$$

From this, the first statement (9) follows. The second statement (10) can be obtained by similar reasoning. □

The above result leads to the following "density evolution" formula. Let $F_{\mathrm{in},1}$, $F_{\mathrm{in},2}$, $F_{\mathrm{out},1}$, and $F_{\mathrm{out},2}$ be the CDFs of $A_{\mathrm{in},1}$, $A_{\mathrm{in},2}$, $A_{\mathrm{out},1}$, and $A_{\mathrm{out},2}$, respectively.

Proposition 3: The CDFs of the α-parameters at the output of a polar transform are related to the CDFs of the α-parameters at the input by

$$F_{out,1}(a) = \iint\limits_{a_1 * a_2 \leq a} \mathrm{d}F_{in,1}(a_1)\, \mathrm{d}F_{in,2}(a_2)$$

$$\begin{aligned} F_{out,2}(a) = &\iint\limits_{(a_1 a_2 / a_1 * a_2) \leq a} (a_1 * a_2)\, \mathrm{d}F_{in,1}(a_1)\, \mathrm{d}F_{in,2}(a_2) \\ &+ \iint\limits_{(\overline{a}_1 a_2 / \overline{a}_1 * a_2) \leq a} (\overline{a}_1 * a_2)\, \mathrm{d}F_{in,1}(a_1)\, \mathrm{d}F_{in,2}(a_2) \end{aligned}$$

These density evolution equations follow from (9) and (10). In the expression for $F_{\mathrm{out},2}(a)$, the integrands $(a_1 * a_2)$ and $(\overline{a}_1 * a_2)$ correspond to the conditional probability of U_1 being 0 and 1, respectively, given that $A_{\mathrm{in},1} = a_1$ and $A_{\mathrm{in},2} = a_2$. We omit the proof for brevity.

For the β-parameters, the analogous result to Proposition 2 is as follows.

$$\begin{aligned} B_{\mathrm{out},1} &= \gamma\,(B_{\mathrm{in},1} * B_{\mathrm{in},2}), \\ B_{\mathrm{out},2} &= \begin{cases} \gamma\,(B_{\mathrm{in},1}B_{\mathrm{in},2}/(B_{\mathrm{in},1} * B_{\mathrm{in},2})), & \Gamma > 0; \\ \gamma\,(\overline{B}_{\mathrm{in},1}B_{\mathrm{in},2}/(\overline{B}_{\mathrm{in},1} * B_{\mathrm{in},2})), & \Gamma \leq 0, \end{cases} \end{aligned}$$

where $\gamma(x) \stackrel{\Delta}{=} \min\{x, 1-x\}$ for any $x \in [0,1]$ and $\Gamma \stackrel{\Delta}{=} (1/2 - U_1)(1/2 - A_{\mathrm{in},1})(1/2 - A_{\mathrm{in},2})$. We omit the derivation of these evolution formulas for the β-parameters since they will not be used in the sequel. The main point to note here is that the knowledge of $(B_{\mathrm{in},1}, B_{\mathrm{in},2}, U_1)$ is not sufficient to determine Γ, hence not sufficient to determine $B_{\mathrm{out},2}$. So, there is no counterpart of Proposition 3 for the β-parameters.

Although there is no general formula for tracking the evolution of the β-parameters through the polar transform, there is an important exceptional case in which we can track that evolution, namely, the case where at least one of the BDEs

TABLE II
POLAR TRANSFORM OF EXTREME BDEs

$B_{\text{in},1}$	$B_{\text{in},2}$	$B_{\text{out},1}$	$B_{\text{out},2}$
perfect	any	$B_{\text{in},2}$	perfect
p.r.	any	p.r.	$B_{\text{in},2}$
any	perfect	$B_{\text{in},1}$	perfect
any	p.r.	p.r.	$B_{\text{in},1}$

at the transform input is extreme. This special case will be important in the sequel, hence we consider it in some detail.

Table II summarizes the evolution of the β-parameters for all possible situations in which at least one of the input BDEs is extreme. (In the table "p.r." stands for "purely random".)

The following proposition states more precisely the way the β-parameters evolve when one of the input BDEs is extreme.

Proposition 4: If $B_{in,1}$ is extreme, then the β-parameters at the output are given by

$$B_{out,1} = \begin{cases} B_{in,2}, & \text{if } B_{in,1} \text{ is perfect} \\ \frac{1}{2}, & \text{if } B_{in,1} \text{ is p.r.}; \end{cases} \tag{11}$$

$$B_{out,2} = \begin{cases} 0, & \text{if } B_{in,1} \text{ is perfect} \\ B_{in,2}, & \text{if } B_{in,1} \text{ is p.r.}. \end{cases} \tag{12}$$

If $B_{in,2}$ is extreme, then (11) *and* (12) *hold after interchanging $B_{in,1}$ and $B_{in,2}$.*

Proof: Suppose $B_{\text{in},1} \equiv 0$ (perfect), then $A_{\text{in},1}$ can only take the values 0 and 1, and we obtain from (9) that

$$A_{\text{out},1} = A_{\text{in},1} * A_{\text{in},2} = \begin{cases} A_{\text{in},2}, & A_{\text{in},1} = 0; \\ \overline{A}_{\text{in},2}, & A_{\text{in},1} = 1. \end{cases}$$

Thus, $B_{\text{out},1} = \min(A_{\text{out},1}, \overline{A}_{\text{out},1}) = \min(A_{\text{in},2}, \overline{A}_{\text{in},2}) = B_{\text{in},2}$, completing the proof of the first case in (11). We skip the proof of the remaining three cases since they follow by similar reasoning. □

III. COVARIANCE REVIEW

In this part, we collect some basic facts about the covariance function, which we will need in the following sections. The first result is the following formula for splitting a covariance into two parts.

Lemma 2: Let $\mathbf{S}$, $\mathbf{T}$ be jointly distributed random vectors over $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Let $f, g : \mathbb{R}^{m+n} \to \mathbb{R}$ be functions such that $\text{Cov}[f(\mathbf{S},\mathbf{T}), g(\mathbf{S},\mathbf{T})]$ exists, i.e., $\mathbb{E} f(\mathbf{S},\mathbf{T}) g(\mathbf{S},\mathbf{T})$, $\mathbb{E} f(\mathbf{S},\mathbf{T})$, and $\mathbb{E} g(\mathbf{S},\mathbf{T})$ all exist. Then,

$$\begin{aligned} \text{Cov}[f(\mathbf{S},\mathbf{T}), g(\mathbf{S},\mathbf{T})] &= \mathbb{E}_{\mathbf{T}} \text{Cov}_{\mathbf{S}|\mathbf{T}}[f(\mathbf{S},\mathbf{T}), g(\mathbf{S},\mathbf{T})] \\ &\quad + \text{Cov}_{\mathbf{T}}[\mathbb{E}_{\mathbf{S}|\mathbf{T}} f(\mathbf{S},\mathbf{T}), \mathbb{E}_{\mathbf{S}|\mathbf{T}} g(\mathbf{S},\mathbf{T})]. \end{aligned} \tag{13}$$

Although this is an elementary result, we give a proof here mainly for illustrating the notation. Our proof follows [7].

Proof: We will omit the arguments of the functions for brevity.

$$\begin{aligned} \text{Cov}(f,g) &= \mathbb{E}_{\mathbf{S},\mathbf{T}} fg - \mathbb{E}_{\mathbf{S},\mathbf{T}} f \cdot \mathbb{E}_{\mathbf{S},\mathbf{T}} g \\ &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_{\mathbf{S}|\mathbf{T}} fg - \mathbb{E}_{\mathbf{T}}\big[\mathbb{E}_{\mathbf{S}|\mathbf{T}} f \cdot \mathbb{E}_{\mathbf{S}|\mathbf{T}} g\big] \\ &\quad + \mathbb{E}_{\mathbf{T}}\big[\mathbb{E}_{\mathbf{S}|\mathbf{T}} f \cdot \mathbb{E}_{\mathbf{S}|\mathbf{T}} g\big] - \mathbb{E}_{\mathbf{T}} \mathbb{E}_{\mathbf{S}|\mathbf{T}} f \cdot \mathbb{E}_{\mathbf{T}} \mathbb{E}_{\mathbf{S}|\mathbf{T}} g \\ &= \mathbb{E}_{\mathbf{T}} \text{Cov}_{\mathbf{S}|\mathbf{T}}(f,g) + \text{Cov}_{\mathbf{T}}(\mathbb{E}_{\mathbf{S}|\mathbf{T}} f, \mathbb{E}_{\mathbf{S}|\mathbf{T}} g). \end{aligned}$$

□

The second result we recall is the following inequality.

Lemma 3 (Chebyshev's Covariance Inequality): Let X be a random variable taking values over $\mathbb{R}$ and let $f, g : \mathbb{R} \to \mathbb{R}$ be any two nondecreasing functions. Suppose that $\text{Cov}(f(X), g(X))$ exists, i.e., $\mathbb{E} f(X) g(X)$, $\mathbb{E} f(X)$, and $\mathbb{E} g(X)$ all exist. Then,

$$\text{Cov}(f(X), g(X)) \geq 0. \tag{14}$$

Proof: Let X' be an independent copy of X. Let $\mathbb{E}$ and $\mathbb{E}'$ denote expectation with respect to X and X', respectively. The proof follows readily from the following identity whose proof can be found in [8, p. 43].

$$\begin{aligned} \text{Cov}(f(X), g(X)) &= \mathbb{E} f(X) g(X) - \mathbb{E} f(X) \mathbb{E} g(X) \\ &= \frac{1}{2} \mathbb{E}' \mathbb{E}[(f(X) - f(X'))(g(X) - g(X'))]. \end{aligned}$$

Now note that for any $x, x' \in \mathbb{R}$, $f(x) - f(x')$ and $g(x) - g(x')$ have the same sign since both f and g are nondecreasing. Thus, $(f(x) - f(x'))(g(x) - g(x')) \geq 0$, and non-negativity of the covariance follows. □

IV. PROOF OF THEOREM 1′

Let us recall the setting of Theorem 1′. We have two independent BDEs (X_1, Y_1) and (X_2, Y_2) as inputs of a polar transform, and two BDEs $(U_1, \mathbf{Y})$ and $(U_2; U_1, \mathbf{Y})$ at the output, with $U_1 = X_1 \oplus X_2$, $U_2 = X_2$, and $\mathbf{Y} = (Y_1, Y_2)$. Associated with these BDEs are the conditional entropy random variables $h_{\text{in},1}$, $h_{\text{in},2}$, $h_{\text{out},1}$, and $h_{\text{out},2}$, as defined by (2) and (3). We will carry out the proof mostly in terms of the canonical parameters $A_i \triangleq \alpha_{X_i|Y_i}(Y_i)$ and $B_i \triangleq \beta_{X_i|Y_i}(Y_i)$, $i = 1, 2$. For shorthand, we will often write $\mathbf{X} = (X_1, X_2)$, $\mathbf{U} = (U_1, U_2)$, $\mathbf{A} = (A_1, A_2)$, and $\mathbf{B} = (B_1, B_2)$.

We will carry out our calculations in the probability space defined by the joint ensemble $(\mathbf{X}, \mathbf{Y})$. Probabilities over this ensemble will be denoted by $P(\cdot)$ and expectations by $\mathbb{E}[\cdot]$. Partial and conditional expectations and covariances will be denoted by $\mathbb{E}_{\mathbf{Y}}$, $\mathbb{E}_{\mathbf{X}|\mathbf{Y}}$, $\text{Cov}_{\mathbf{Y}}$, $\text{Cov}_{\mathbf{X}|\mathbf{Y}}$, etc. Due to the 1-1nature of the correspondence between $\mathbf{U}$ and $\mathbf{X}$, expectation and covariance operators such as $\mathbb{E}_{\mathbf{U}|\mathbf{Y}}$ and $\text{Cov}_{\mathbf{U}|\mathbf{Y}}$ will be equivalent to $\mathbb{E}_{\mathbf{X}|\mathbf{Y}}$ and $\text{Cov}_{\mathbf{X}|\mathbf{Y}}$, respectively. We will prefer to use expectation operators in terms of the primary variables $\mathbf{X}$ and $\mathbf{Y}$ rather than the secondary (derived) variables such as $\mathbf{U}$, $\mathbf{A}$, $\mathbf{B}$, to emphasize that the underlying space is $(\mathbf{X}, \mathbf{Y})$. We note that, due to the independence of Y_1 and Y_2, A_1 and A_2 are independent; likewise, B_1 and B_2 are independent.

A. Covariance Decomposition Step

As the first step of the proof of Theorem 1′, we use the covariance decomposition formula (13) to write

$$\begin{aligned} \text{Cov}(h_{\text{out},1}, h_{\text{out},2}) &= \mathbb{E}_{\mathbf{Y}} \text{Cov}_{\mathbf{X}|\mathbf{Y}}(h_{\text{out},1}, h_{\text{out},2}) \\ &\quad + \text{Cov}_{\mathbf{Y}}(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},1}, \mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},2}). \end{aligned} \tag{15}$$

For brevity, we will use the notation

$$\text{Cov}_1 \triangleq \mathbb{E}_{\mathbf{Y}} \text{Cov}_{\mathbf{X}|\mathbf{Y}}(h_{\text{out},1}, h_{\text{out},2})$$

$$\text{Cov}_2 \triangleq \text{Cov}_{\mathbf{Y}}(\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},1}, \mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},2})$$

to denote the two terms on the right hand side of (15). Our proof of Theorem 1′ will consist in proving the following two statements.

Proposition 5: We have $\mathrm{Cov}_1 \geq 0$*, with equality iff either* (X_1, Y_1) *or* (X_2, Y_2) *is an erasing BDE.*

Proposition 6: We have $\mathrm{Cov}_2 \geq 0$*.*

Remark 2: We note that $\mathrm{Cov}_2 = 0$ *iff, of the two BDEs* (X_1, Y_1) *and* (X_2, Y_2)*, either one is extreme or both are pure. We note this only for completeness but do not use it in the paper.*

The rest of the section is devoted to the proof of the above propositions.

B. Proof of Proposition 5

For $p, q \in [0, 1]$, define

$$f(p,q) \stackrel{\Delta}{=} (p*q)(p*\overline{q}) \log\left(\frac{p*q}{p*\overline{q}}\right) \times \left[\mathcal{H}\left(\frac{p\,\overline{q}}{p*\overline{q}}\right) - \mathcal{H}\left(\frac{p\,q}{p*q}\right)\right]. \tag{16}$$

We will soon give a formula for Cov_1 in terms of this function. First, a number of properties of $f(p, q)$ will be listed. The following symmetry properties are immediate:

$$f(p,q) = f(\overline{p},q) = f(\overline{p},q) = f(\overline{p},\overline{q}), \tag{17}$$

$$f(p,q) = f(q,p). \tag{18}$$

Lemma 4: We have $f(p, q) \geq 0$ *for all* $p, q \in [0, 1]$ *with equality iff* $p \in \{0, 1/2, 1\}$ *or* $q \in \{0, 1/2, 1\}$.

Proof: We use (17) to write

$$f(p,q) = f(r,s) \tag{19}$$

where $r \stackrel{\Delta}{=} \min\{p, \overline{p}\}$ and $s \stackrel{\Delta}{=} \min\{q, \overline{q}\}$. Thus, instead of proving $f(p, q) \geq 0$, it suffices to prove $f(r, s) \geq 0$ for $0 \leq r, s \leq 1/2$. In fact, using (18), it suffices to prove $f(r, s) \geq 0$ for $0 \leq r \leq s \leq 1/2$. Assuming $0 \leq r \leq s \leq 1/2$, it is straightforward to show that

$$r*s \geq r*\overline{s} \quad \text{and} \quad \frac{rs}{r*s} \leq \frac{r\,\overline{s}}{r*\overline{s}} \leq \frac{1}{2}. \tag{20}$$

Thus, if we write out the expression for $f(r, s)$, as in (16) with (r, s) in place of (p, q), we can see easily that each of the four factors on the right hand side of that expression are non-negative. More specifically, the logarithmic term is non-negative due to the first inequality in (20) and the bracketed term is non-negative due to the second inequality in (20). This completes the proof that $f(p, q) \geq 0$ for all $p, q \in [0, 1]$.

Next, we identify the necessary and sufficient conditions for $f(p, q)$ to be zero over $0 \leq p, q \leq 1$. Clearly, $f(p, q) = 0$ iff one of the four factors on the right hand side of (16) equals zero. By straightforward algebra, one can verify the following statements. The first factor $p * q$ equals zero iff $(p, q) \in \{(0, 1), (1, 0)\}$. The second factor $p*\overline{q}$ equals zero iff $(p, q) \in \{(0, 0), (1, 1)\}$. The log term equals zero iff $p = 1/2$ or $q = 1/2$. Finally the difference of the entropy terms equals zero iff $pq/p*q = p\overline{q}/p*\overline{q}$ or $pq/p*q = 1 - p\overline{q}/p*\overline{q}$ which in turn is true iff $p \in \{0, 1/2, 1\}$ or $q \in \{0, 1/2, 1\}$. Taking the logical combination of these conditions we conclude that $f(p, q) = 0$ iff $p \in \{0, 1/2, 1\}$ or $q \in \{0, 1/2, 1\}$. □

Lemma 5: We have

$$\mathrm{Cov}_1 = \mathbb{E} f(\mathbf{A}) = \mathbb{E} f(\mathbf{B}). \tag{21}$$

Proof: Fix a sample $\mathbf{y} = (y_1, y_2)$. Note that

$$\begin{aligned}\mathrm{Cov}_{\mathbf{X}|\mathbf{y}}(h_{\mathrm{out},1}, h_{\mathrm{out},2}) &= \mathrm{Cov}_{\mathbf{X}|\mathbf{y}}(h(U_1|\mathbf{y}), h(U_2|U_1,\mathbf{y})) \\ &= \mathbb{E}_{\mathbf{X}|\mathbf{y}}\big\{\big[h(U_1|\mathbf{y}) - H(U_1|\mathbf{y})\big] \\ &\qquad \times h(U_2|U_1,\mathbf{y})\big\} \\ &= \sum_{u_1} p_{U_1|\mathbf{Y}}(u_1|\mathbf{y}) \\ &\qquad \big[h(u_1|\mathbf{y}) - H(U_1|\mathbf{y})\big] H(U_2|u_1,\mathbf{y}).\end{aligned}$$

After some algebra, the term $\big[h(u_1|\mathbf{y}) - H(U_1|\mathbf{y})\big]$ simplifies to

$$(1 - p_{U_1|\mathbf{Y}}(u_1|\mathbf{y})) \log \frac{1 - p_{U_1|\mathbf{Y}}(u_1|\mathbf{y})}{p_{U_1|\mathbf{Y}}(u_1|\mathbf{y})}.$$

Substituting this in the preceding equation and writing out the sum over U_1 explicitly, we obtain

$$\begin{aligned}\mathrm{Cov}_{\mathbf{X}|\mathbf{y}}(h_{\mathrm{out},1}, h_{\mathrm{out},2}) &= p_{U_1|\mathbf{Y}}(0|\mathbf{y}) p_{U_1|\mathbf{Y}}(1|\mathbf{y}) \\ &\cdot \log \frac{p_{U_1|\mathbf{Y}}(0|\mathbf{y})}{p_{U_1|\mathbf{Y}}(1|\mathbf{y})} \big[H(U_2|U_1 = 1,\mathbf{y}) - H(U_2|U_1 = 0,\mathbf{y})\big].\end{aligned}$$

Expressing each factor on the right side of the above equation in terms of $a_i = \alpha(y_i)$, $i = 1, 2$, we see that it equals $f(a_1, a_2)$. Taking expectations, we obtain $\mathrm{Cov}_1 = \mathbb{E} f(\mathbf{A})$. The alternative formula $\mathrm{Cov}_1 = \mathbb{E} f(\mathbf{B})$ follows from the fact that $f(\mathbf{B}) = f(\mathbf{A})$ due to the symmetries (17). □

Proposition 5 now follows readily. We have $\mathrm{Cov}_1 \geq 0$ since $f(a_1, a_2) \geq 0$ for all $a_1, a_2 \in [0, 1]$ by Lemma 4. By the same lemma, strict positivity, $\mathbb{E} f(\mathbf{A}) > 0$, is possible iff the events $A_1 \notin \{0, 1/2, 1\}$ and $A_2 \notin \{0, 1/2, 1\}$ can occur simultaneously with non-zero probability, *i.e.*, iff

$$P\left(A_1 \notin \{0, \frac{1}{2}, 1\}\right) P\left(A_2 \notin \{0, \frac{1}{2}, 1\}\right) > 0, \tag{22}$$

since A_1 and A_2 are independent. Condition (22) is true iff

$$P\left(B_1 \notin \{0, \frac{1}{2}\}\right) P\left(B_2 \notin \{0, \frac{1}{2}\}\right) > 0, \tag{23}$$

which in turn is true iff neither B_1 nor B_2 is erasing. This completes the proof of Proposition 5.

C. Proof of Proposition 6

Let $g_1(p, q) \stackrel{\Delta}{=} \mathcal{H}(p * q)$ and $g_2(p, q) \stackrel{\Delta}{=} \mathcal{H}(p) + \mathcal{H}(q) - \mathcal{H}(p*q)$ for $p, q \in [0, 1]$. These functions will be used to give an explicit expression for Cov_2. First, we note some symmetry properties of the two functions. For $i = 1, 2$, we have

$$g_i(p,q) = g_i(\overline{p},q) = g_i(\overline{p},q) = g_i(\overline{p},\overline{q}), \tag{24}$$

$$g_i(p,q) = g_i(q,p). \tag{25}$$

We omit the proofs since they are immediate.

Lemma 6: We have, for $i = 1, 2$*,*

$$\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{out,i} = g_i(\mathbf{A}) = g_i(\mathbf{B}). \tag{26}$$

Proof: These results follow from (6), (9), and (10). We compute $\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},1}$ as follows.

$$\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},1} = \mathbb{E}_{\mathbf{U}|\mathbf{A}} h_{\text{out},1} = \mathcal{H}(A_1 * A_2) = g_1(\mathbf{A}).$$

For the second term, we use the entropy conservation (5).

$$\begin{aligned}\mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},2} &= \mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{in},1} + \mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{in},2} - \mathbb{E}_{\mathbf{X}|\mathbf{Y}} h_{\text{out},1} \\ &= \mathcal{H}(A_1) + \mathcal{H}(A_2) - \mathcal{H}(A_1 * A_2) = g_2(\mathbf{A}).\end{aligned}$$

The second form of the formulas in terms of $\mathbf{B}$ follow from the symmetry properties (24). □

As a corollary to Lemma 6, we now have

$$\text{Cov}_2 = \text{Cov}[g_1(\mathbf{B}), g_2(\mathbf{B})]. \tag{27}$$

In order to prove that $\text{Cov}_2 \geq 0$, we will apply Lemma 3 to (27). First, we need to establish some monotonicity properties of the functions g_1 and g_2. We insert here a general definition.

Definition 1: A function $g : \mathbb{R}^n \to \mathbb{R}$ is called nondecreasing if, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $g(\mathbf{x}) \leq g(\mathbf{y})$ whenever $x_i \leq y_i$ for all $i = 1, \ldots, n$.

Lemma 7: $g_1 : [0, 1/2]^2 \to \mathbb{R}^+$ is nondecreasing.

Proof: Since $g_1(b_1, b_2) = g_1(b_2, b_1)$, it suffices to show that $g_1(b_1, b_2)$ is nondecreasing in $b_1 \in [0, 1/2]$ for fixed $b_2 \in [0, 1/2]$. So, fix $b_2 \in [0, 1/2]$ and consider $g_1(b_1, b_2)$ as a function of $b_1 \in [0, 1/2]$. Recall the well-known facts that the function $\mathcal{H}(p)$ over $p \in [0, 1]$ is a strictly concave non-negative function, symmetric around $p = 1/2$, attaining its minimum value of 0 at $p \in \{0, 1\}$, and its maximum value of 1 at $p = 1/2$. It is readily verified that, for any fixed $b_2 \in [0, 1/2]$, as b_1 ranges from 0 to $1/2$, $b_1 * b_2$ decreases from $\overline{b}_2$ to $1/2$, hence $g_1(b_1, b_2) = \mathcal{H}(b_1 * b_2)$ increases from $\mathcal{H}(\overline{b}_2)$ to $\mathcal{H}(1/2) = 1$, with strict monotonicity if $b_2 \neq 1/2$. This completes the proof. □

Lemma 8: $g_2 : [0, 1/2]^2 \to \mathbb{R}^+$ is nondecreasing.

Proof: Again, since $g_2(b_1, b_2) = g_2(b_2, b_1)$, it suffices to show that $g_2(b_1, b_2)$ is nondecreasing in $b_1 \in [0, 1/2]$ for fixed $b_2 \in [0, 1/2]$. Recall that $g_2(b_1, b_2) = \mathcal{H}(b_1) + \mathcal{H}(b_2) - \mathcal{H}(b_1 * b_2)$. Exclude the constant term $\mathcal{H}(b_2)$ and focus on the behavior of $I(b_1) \stackrel{\Delta}{=} \mathcal{H}(b_1 * b_2) - \mathcal{H}(b_1)$ over $b_1 \in [0, 1/2]$. Observe that $I(b_1)$ is the mutual information between the input and output terminals of a BSC with crossover probability b_1 and a Bernoulli-b_2 input. The mutual information between the input and output of a discrete memoryless channel is a convex function of the set of channel transition probabilities for any fixed input probability assignment [9, p. 90]. So, $I(b_1)$ is convex in $b_1 \in [0, 1/2]$. Since $I(0) = \mathcal{H}(b_2)$ and $I(1/2) = 0$, it follows from the convexity property that $I(b_1)$ is decreasing in $b_1 \in [0, 1/2]$, and strictly decreasing if $b_2 \neq 0$. This completes the proof. □

Proposition 6 can now be proved as follows. First, we apply Lemma 2 to (27) to decompose Cov_2 as

$$\begin{aligned}\text{Cov}(g_1(\mathbf{B}), g_2(\mathbf{B})) = \mathbb{E}_{B_1} \text{Cov}_{B_2}(g_1(\mathbf{B}), g_2(\mathbf{B})) \\ + \text{Cov}_{B_1}(\mathbb{E}_{B_2} g_1(\mathbf{B}), \mathbb{E}_{B_2} g_2(\mathbf{B})).\end{aligned}$$

Each covariance term on the right side is positive by Chebyshev's correlation inequality (Lemma 3) and the fact that g_1 and g_2 are nondecreasing in the sense of Def. 1. More specifically, Chebyshev's inequality implies that

$$\text{Cov}_{B_2}(g_1(b_1, B_2), g_2(b_1, B_2)) \geq 0$$

for any fixed $b_1 \in [0, 1/2]$ since $g_1(b_1, b_2)$ and $g_2(b_1, b_2)$ are nondecreasing functions of b_2 when b_1 is fixed. Likewise, Chebyshev's inequality implies that

$$\text{Cov}_{B_1}(\mathbb{E}_{B_2} g_1(\mathbf{B}), \mathbb{E}_{B_2} g_2(\mathbf{B})) \geq 0$$

since $\mathbb{E}_{B_2} g_1(b_1, B_2)$ and $\mathbb{E}_{B_2} g_2(b_1, B_2)$ are, as a simple consequence of Lemma 8, nondecreasing functions of b_1.

D. Proof of Theorem 1′

The covariance inequality (4) is an immediate consequence of (15) and Propositions 5 and 6. We only need to identify the necessary and sufficient conditions for the covariance to be zero. For brevity, let us define

$$T \stackrel{\Delta}{=} \text{“}B_1 \text{ or } B_2 \text{ is extreme”}.$$

The present goal is to prove that

$$\text{Cov}(h_{\text{out},1}, h_{\text{out},2}) = 0 \quad \text{iff} \quad T \text{ holds.} \tag{28}$$

The proof will make us of the decomposition

$$\begin{aligned}\text{Cov}(h_{\text{out},1}, h_{\text{out},2}) &= \text{Cov}_1 + \text{Cov}_2 \\ &= \mathbb{E} f(\mathbf{B}) + \text{Cov}(g_1(\mathbf{B}), g_2(\mathbf{B}))\end{aligned} \tag{29}$$

that we have already established. Let us define

$$R \stackrel{\Delta}{=} \text{“}B_1 \text{ or } B_2 \text{ is erasing”}$$

and note that R appears in Proposition 5 as the necessary and sufficient conditions for Cov_1 to be zero. Note also that T implies R since “extreme” is a special instance “erasing” according to definitions in Table I.

We begin the proof of (28) with the sufficiency part. in other words, by assuming that T holds. Since T implies R, T is sufficient for $\text{Cov}_1 = 0$. To show that T is sufficient for $\text{Cov}_2 = 0$, we recall Proposition 4, which states that, if T is true, then either $B_{\text{out},1}$ or $B_{\text{out},2}$ is extreme. To be more specific, if $B_{\text{in},1}$ or $B_{\text{in},2}$ is p.r., then $B_{\text{out},1} \equiv 1/2$ and $g_1(\mathbf{B}) \equiv 1$; if $B_{\text{in},1}$ or $B_{\text{in},2}$ is perfect, then $B_{\text{out},2} \equiv 0$ and $g_2(\mathbf{B}) \equiv 0$. (The notation “$\equiv$” should be read as “equals with probability one”.) In either case, $\text{Cov}_2 = \text{Cov}(g_1(\mathbf{B}), g_2(\mathbf{B})) = 0$. This completes the proof of the sufficiency part.

To prove necessity in (28), we write T as

$$T = R \wedge (R^c \vee T) \tag{30}$$

where R^c denotes the complement (negation) of R. The validity of (30) follows from $R \wedge T = T$. To prove necessity, we will use contraposition and show that T^c implies $\text{Cov}(h_{\text{out},1}, h_{\text{out},2}) > 0$. Note that $T^c = R^c \vee (R \wedge T^c)$. If T^c is true, then either R^c or $(R \wedge T^c)$ is true. If R^c is true, then $\text{Cov}_1 > 0$ by Proposition 5. We will complete the proof by showing that $R \wedge T^c$ implies $\text{Cov}(h_{\text{out},1}, h_{\text{out},2}) > 0$. For this, we note that when one of the BDEs is erasing, there is an explicit formula for Cov_2. We state this result as follows.

Lemma 9: Let B_1 be erasing with erasure probability $\epsilon \triangleq P(B_1 = 1/2)$ and let B_2 be arbitrary with $\delta \triangleq H(X_2|Y_2)$. Then,

$$\mathrm{Cov}_2 = \epsilon(1-\epsilon)\delta(1-\delta) \tag{31}$$

This formula remains valid if B_2 is erasing with erasure probability $\epsilon \triangleq P(B_2 = 1/2)$ and B_1 is arbitrary with $\delta \triangleq H(X_1|Y_1)$.

Proof: We first observe that

$$g_1(B_1, B_2) = \begin{cases} \mathcal{H}(B_2), & B_1 = 0; \\ 1, & B_1 = \frac{1}{2}; \end{cases}$$

$$g_2(B_1, B_2) = \begin{cases} 0, & B_1 = 0; \\ \mathcal{H}(B_2), & B_1 = \frac{1}{2}. \end{cases}$$

Now, the claim (31) is obtained by simply computing the covariance of these two random variables. The second claim follows by the symmetry property (25). □

Returning to the proof of Theorem 1′, the proof of the necessity part is now completed as follows. If $R \wedge T^c$ holds, then at least one of the BDEs is *strictly* erasing (has erasure probability $0 < \epsilon < 1$) and the other is non-extreme. By Proposition 1, the conditional entropy $H(X|Y)$ of a non-extreme BDE (X, Y) is strictly between 0 and 1. So, by Lemma 9, we have $\mathrm{Cov}_2 > 0$. This completes the proof.

V. Varentropy Under Higher-Order Transforms

In this part, we consider the behavior of varentropy under higher-order polar transforms. The section concludes with a proof of the polarization theorem using properties of varentropy.

A. Polar Transform of Higher Orders

For any $n \geq 1$, there is a polar transform of order $N = 2^n$. A polar transform of order $N = 2^n$ is a mapping ψ_N that takes N BDEs $\{(X_i, Y_i)\}_{i=1}^N$, as input, and produces a new set of N BDEs $\{(U_i; \mathbf{U}^{i-1}, \mathbf{Y})\}_{i=1}^N$, where $\mathbf{Y} = (Y_1, \ldots, Y_N)$ and $\mathbf{U}^{i-1} = (U_1, \ldots, U_{i-1})$ is a subvector of $\mathbf{U} = (U_1, \ldots, U_N)$, which in turn is obtained from $\mathbf{X} = (X_1, \ldots, X_N)$ by the transform

$$\mathbf{U} = \mathbf{X}\mathbf{G}_N, \qquad \mathbf{G}_N \triangleq \mathbf{F}^{\otimes n}, \qquad \mathbf{F} \triangleq \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{32}$$

The sign "$\otimes n$" in the exponent denotes the nth Kronecker power. We allow Y_i to take values in some arbitrary set $\mathcal{Y}_i$, $1 \leq i \leq N$, which is not necessarily discrete. We assume that (X_i, Y_i), $1 \leq i \leq N$, are independent but not necessarily identically-distributed.

(An alternate form of the polar transform matrix, as used in [4], is $\mathbf{G}_N = \mathbf{B}_N\mathbf{F}^{\otimes n}$, in which $\mathbf{B}_N$ is a permutation matrix known as *bit-reversal*. The form of $\mathbf{G}_N$ that we are using here is less complex and adequate for the purposes of this paper. However, if desired, the results given below can be proved under bit-reversal (or, any other permutation) after suitable re-indexing of variables.)

B. Polarization Results

The first result in this section is a generalization of Theorem 1 to higher order polar transforms.

Theorem 2: Let $N = 2^n$ for some $n \geq 1$. Let (X_i, Y_i), $1 \leq i \leq N$, be independent but not necessarily identically distributed BDEs. Consider the polar transform $\mathbf{U} = \mathbf{X}\mathbf{G}_N$ and let $(U_i; \mathbf{U}^{i-1}, \mathbf{Y})$, $1 \leq i \leq N$, be the BDEs at the output of the polar transform. The varentropy is nonincreasing under any such polar transform in the sense that

$$\sum_{i=1}^N V(U_i|\mathbf{U}^{i-1}, \mathbf{Y}) \leq \sum_{i=1}^N V(X_i|Y_i). \tag{33}$$

The next result considers the special case in which the BDEs at the input of the polar transform are i.i.d. and the transform size goes to infinity.

Theorem 3: Let (X_i, Y_i), $1 \leq i \leq N$, be i.i.d. copies of a given BDE (X, Y). Consider the polar transform $\mathbf{U} = \mathbf{X}\mathbf{G}_N$ and let $(U_i; \mathbf{U}^{i-1}, \mathbf{Y})$, $1 \leq i \leq N$, be the BDEs at the output of the polar transform. Then, the average varentropy at the output goes to zero asymptotically:

$$\frac{1}{N}\sum_{i=1}^N V(U_i|\mathbf{U}^{i-1}, \mathbf{Y}) \to 0, \quad \text{as } N \to \infty. \tag{34}$$

C. Proof of Theorem 2

We will first bring out the recursive nature of the polar transform by giving a more abstract formulation in terms of the α-parameters of the variables involved. Let us recall that a polar transform of order two is essentially a mapping of the form

$$(A_{\mathrm{in},1}, A_{\mathrm{in},2}) \to (A_{\mathrm{out},1}, A_{\mathrm{out},2}), \tag{35}$$

where $A_{\mathrm{in},1}$ and $A_{\mathrm{in},2}$ are the α-parameters of the input BDEs (X_1, Y_1) and (X_2, Y_2), and $A_{\mathrm{out},1}$ and $A_{\mathrm{out},2}$ are the α-parameters of the output BDEs $(U_1, \mathbf{Y})$ and $(U_2; U_1, \mathbf{Y})$.

Alternatively, the polar transform may be viewed as an operation in the space of CDFs of α-parameters and represented in the form

$$(F_{\mathrm{out},1}, F_{\mathrm{out},2}) = \psi_2(F_{\mathrm{in},1}, F_{\mathrm{in},2}) \tag{36}$$

where $F_{\mathrm{in},i}$ and $F_{\mathrm{out},i}$ are the CDFs of $A_{\mathrm{in},i}$ and $A_{\mathrm{out},i}$, respectively.

Let $\mathcal{M}$ be the space of all CDFs belonging to random variables defined on the interval $[0, 1]$. The CDF of any α-parameter A belongs to $\mathcal{M}$, and conversely, each CDF $F \in \mathcal{M}$ defines a valid α-parameter A. Thus, we may regard the polar transform of order two (36) as an operator of the form

$$\psi_2 : \mathcal{M}^2 \to \mathcal{M}^2. \tag{37}$$

We will define higher order polar transforms following this viewpoint.

For each $i = 1, \ldots, N$, let $A_{\mathrm{in},i}$ denote the α-parameter of the ith BDE (X_i, Y_i) at the input, and let $F_{\mathrm{in},i}$ denote the CDF of $A_{\mathrm{in},i}$. Likewise, let $A_{\mathrm{out},i}$ denote the α-parameter of the ith BDE $(U_i; \mathbf{U}^{i-1}, \mathbf{Y})$ at the output, and let $F_{\mathrm{out},i}$ be

the CDF of $A_{\text{out},i}$. Let $\mathbf{F}_{\text{in}} = (F_{\text{in},1}, \ldots, F_{\text{in},N})$ and $\mathbf{F}_{\text{out}} = (F_{\text{out},1}, \ldots, F_{\text{out},N})$. We will represent a polar transform of order N abstractly as $\mathbf{F}_{\text{out}} = \psi_N(\mathbf{F}_{\text{in}})$.

There is a recursive formula that defines the polar transform of order N in terms of the polar transform of order $N/2$. Let us split the output $\mathbf{F}_{\text{out}}$ into two halves as $\mathbf{F}_{\text{out}} = (\mathbf{F}'_{\text{out}}, \mathbf{F}''_{\text{out}})$. Each half is obtained by a size-$N/2$ transform of the form

$$\mathbf{F}'_{\text{out}} = \psi_{N/2}(\mathbf{F}'_{\text{in}}), \quad \mathbf{F}''_{\text{out}} = \psi_{N/2}(\mathbf{F}''_{\text{in}}),$$

in which $\mathbf{F}'_{\text{in}} = (F'_{\text{in},1}, \ldots, F'_{\text{in},N/2})$, $\mathbf{F}''_{\text{in}} = (F''_{\text{in},1}, \ldots, F''_{\text{in},N/2})$ are obtained from $\mathbf{F}_{\text{in}}$ through a series of size-2 transforms

$$(F'_{\text{in},i}, F''_{\text{in},i}) = \psi_2(F_{\text{in},i}, F_{\text{in},i+N/2}), \quad 1 \le i \le N/2. \quad (38)$$

The derivation of the above recursion from the algebraic definition (32) is standard knowledge in polar coding, and will be omitted.

Let us write $V(F)$ to denote the varentropy $V(X|Y)$ of a BDE (X, Y) whose α-parameter has CDF F. Using (8), we can write $V(F)$ as

$$V(F) = \int_0^1 \mathcal{H}_2(a)\, \mathrm{d}F(a) - \left(\int_0^1 \mathcal{H}(a)\, \mathrm{d}F(a) \right)^2. \quad (39)$$

We are now ready to prove Theorem 2. The proof will be by induction. First note that the claim (33) is true for $N = 2$ by Theorem 1. Let $N \ge 4$ and suppose, as induction hypothesis, that the claim is true for transforms of orders $N/2$ and smaller. We will show that the claim is true for order N. By the induction hypothesis, we have

$$\sum_{i=1}^{N/2} V(F'_{\text{out},i}) \le \sum_{i=1}^{N/2} V(F'_{\text{in},i}) \quad (40)$$

and

$$\sum_{i=1}^{N/2} V(F''_{\text{out},i}) \le \sum_{i=1}^{N/2} V(F''_{\text{in},i}). \quad (41)$$

Summing (40) and (41) side by side,

$$\sum_{i=1}^{N} V(F_{\text{out},i}) \le \sum_{i=1}^{N/2} \left[V(F'_{\text{in},i}) + V(F''_{\text{in},i}) \right] \quad (42)$$

Using the induction hypothesis again, we obtain

$$V(F'_{\text{in},i}) + V(F''_{\text{in},i}) \le V(F_{\text{in},i}) + V(F_{\text{in},i+N/2}) \quad (43)$$

for all $i = 1, \ldots, N/2$. The proof if completed by using (43) to upper-bound the right side of (42) further.

D. Proof of Theorem 3

In this proof we will consider a sequence of polar transforms indexed by $n \ge 1$. For a given n, the size of the transform is $N = 2^n$; the inputs of the transform are (X_i, Y_i), $1 \le i \le N$, which are i.i.d. copies of a given BDE (X, Y); the outputs of the transform, which we will refer to as "the nth generation BDEs", are $(U_i; \mathbf{U}^{i-1}, \mathbf{Y})$, $1 \le i \le N$. Let F_0 denote the CDF of (X, Y). Let $F_{n,i}$ denote the CDF of $(U_i; \mathbf{U}^{i-1}, \mathbf{Y})$, the ith BDE in the nth generation, $n \ge 1$, $1 \le i \le 2^n$, and set $F_{0,1} = F_0$. In this notation, we can express the normalized varentropy compactly as

$$\overline{V}_n \stackrel{\Delta}{=} \frac{1}{2^n} \sum_{i=1}^{2^n} V(U_i|\mathbf{U}^{i-1}, \mathbf{Y}) = \frac{1}{2^n} \sum_{i=1}^{2^n} V(F_{n,i}), \quad n \ge 1,$$

and $\overline{V}_0 \stackrel{\Delta}{=} V(F_0)$. The sequence $\{\overline{V}_n\}$ is non-negative (since each $\overline{V}_n$ is a sum of varentropies), and nonincreasing by Theorem 2. Thus $\{\overline{V}_n\}$ converges to a limit $c \ge 0$. Our goal is to prove that $c = 0$.

The analysis in the proof of Theorem 2 covers the present case as a special instance. In the present notation, the recursive relation (38) takes the form

$$(F_{n,i}, F_{n,i+2^{n-1}}) = \psi_2(F_{n-1,i}, F_{n-1,i}), \quad 1 \le i \le 2^{n-1},$$

since here we have $F_{n-1,i} = F_{n-1,i+2^{n-1}}$ due to i.i.d. BDEs at the transform input. Using this relation, we obtain readily an explicit formula for the incremental change in normalized varentropy from generation n to $(n+1)$, namely,

$$D_{n+1} \stackrel{\Delta}{=} \overline{V}_{n+1} - \overline{V}_n = -\sum_{i=1}^{2^n} C(F_{n,i}), \quad n \ge 0, \quad (44)$$

where

$$C(F_{n,i}) \stackrel{\Delta}{=} V(F_{n,i}) - \left[V(F_{n+1,i}) + V(F_{n+1,i+2^n}) \right]/2. \quad (45)$$

If we denote the conditional entropy random variables in the polar transform as $\{h_{n,i}\}$, it can be seen that

$$C(F_{n,i}) = \mathrm{Cov}(h_{n+1,i}, h_{n+1,i+2^n}).$$

Thus, we have $C(F_{n,i}) \ge 0$ by Theorem 1′, implying that $D_n \le 0$ for all $n \ge 1$. It is useful to note here that

$$c \stackrel{\Delta}{=} \lim_{n\to\infty} \overline{V}_n = V(F_0) - \sum_{i=1}^{\infty} D_n, \quad (46)$$

showing explicitly that c is the limit of a monotone nonincreasing sequence of sums.

For $\delta \ge 0$, let

$$\mathcal{M}_\delta \stackrel{\Delta}{=} \{F \in \mathcal{M} : V(F) \ge \delta\}. \quad (47)$$

and

$$\Delta(\delta) \stackrel{\Delta}{=} \inf\{C(F) : F \in \mathcal{M}_\delta\}. \quad (48)$$

As we will see in a moment, the main technical problem that remains is to show that

$$\delta > 0 \implies \Delta(\delta) > 0. \quad (49)$$

While this proposition seems plausible in view of the fact that $C(F) = 0$ iff $V(F) = 0$ (by Theorem 1′), there is the technical question of whether the "inf" in (48) is achieved as a "min" by some $F \in \mathcal{M}_\delta$. We will first complete the proof of Theorem 3 by assuming that (49) holds. Then, we will give a proof of (49) in the Appendix.

Let $J_n(\delta) \stackrel{\Delta}{=} \{1 \le i \le 2^n : F_{n,i} \in \mathcal{M}_\delta\}$, and $P_n(\delta) \stackrel{\Delta}{=} |J_n(\delta)|/2^n$. For $\delta > 0$, we may think of $J_n(\delta)$ as the set of

"bad" BDEs in the nth generation and $P_n(\delta)$ as their fraction in the same population. From (44), we obtain the bound

$$D_n \leq -P_n(\delta)\Delta(\delta), \quad \delta \geq 0. \tag{50}$$

To apply this bound effectively, we need a lower bound on $P_n(\delta)$. To derive such a lower bound, we observe that, for any $\delta \geq 0$,

$$\overline{V}_n \leq [1 - P_n(\delta)]\delta + P_n(\delta)M \leq \delta + P_n(\delta)M \tag{51}$$

where $M \stackrel{\Delta}{=} 2.3434$ is the bound on varentropy provided by Lemma 1. Let n_0 be such that for all $n \geq n_0$, $\overline{V}_n \geq c/2$. Since $\{\overline{V}_n\}$ converges to $c \geq 0$, n_0 exists and is finite. This, combined with (51), implies the following bound on the fraction of bad indices.

$$P_n(\delta) \geq \frac{\overline{V}_n - \delta}{M} \geq \frac{c/2 - \delta}{M}, \qquad n \geq n_0. \tag{52}$$

Using (52) in (50) with $\delta = c/4$ gives

$$D_n \leq -(c/4M) \cdot \Delta(c/4), \quad n \geq n_0. \tag{53}$$

From (46), we see that having $c > 0$ is incompatible with (53). This completes the proof that $c = 0$ (subject to the assumption that (49) holds, which is proved in the Appendix).

VI. Concluding Remarks

One of the implications of the convergence of average varentropy to zero is that the entropy random variables "concentrate" around their means along almost all trajectories of the polar transform. This concentration phenomenon provides a theoretical basis for understanding why polar decoders are robust against quantization of likelihoood ratios [10].

Theorem 3 may be seen as an alternative version of the "polarization" results of [4]. In [4], the analysis was centered around the mutual information function and martingale methods were used to establish asymptotic results. The present study is centered around the varentropy and uses weak convegence of probability distributions. The use of weak convergence in such problems is not new; Richardson and Urbanke [6, pp. 187 and 188] used similar methods to deal with problems of convergence of functionals defined on the space of binary memoryless channels.

We should mention that Alsan and Telatar [11] have given an elementary proof of polarization that avoids martingale theory, and instead, uses Mrs. Gerber's lemma [12]. It appears possible to adopt the method of [11] to establish Theorem 3 without using weak convergence.

Appendix
Proof of (49)

Lemma 10: The space $\mathcal{M}$ of CDFs on $[0, 1]$ is a compact metric space.

Proof: This follows from a general result about probability measures on compact metric spaces. [14, p. 45, Th. 6.4] states that, for any compact metric space X, the space $\mathcal{M}(X)$ of all probability measures defined on the σ-algebra of Borel sets in X is compact. Our definition of $\mathcal{M}$ above coincides with the $\mathcal{M}(X)$ with $X = [0, 1]$. □

For $F \in \mathcal{M}$, let F^- and F^+ be defined by (see (37))

$$(F^-, F^+) = \psi_2(F, F).$$

Define $C : \mathcal{M} \to \mathbb{R}$ as the mapping

$$C(F) \stackrel{\Delta}{=} V(F) - \big[V(F^-) + V(F^+)\big]/2. \tag{54}$$

This definition is a repetition of (45) in a more convenient notation. We have already seen the interpretation of $C(F)$ as a covariance and mentioned that $C(F) \geq 0$. It is also clear that $C(F)$ is bounded: $C(F) \leq V(F) \leq M$, where $M = 2.3434$. Thus, we may restrict the range of C and write it as a mapping $C : \mathcal{M} \to [0, M]$.

Lemma 11: The mapping $C : \mathcal{M} \to [0, M]$ is continuous (w.r.t. the weak topology on $\mathcal{M}$ and the usual topology of Borel sets in $\mathbb{R}$).

Proof: We wish to show that if $F_n \Rightarrow F_0$ (in the sense of weak-convergence), then $|C(F_n) - C(F_0)| \to 0$. We observe from (39) that $V(F)$ is given in terms of expectations of two bounded uniformly continuous functions, $\mathcal{H} : [0, 1] \to [0, 1]$ and $\mathcal{H}_2 : [0, 1] \to [0, M]$. Thus, by definition of weak convergence ([14, p. 40]), we have $|V(F_n) - V(F_0)| \to 0$. In view of (54), the proof will be complete if we can show that $(F_n \Rightarrow F_0)$ implies $(F_n^- \Rightarrow F_0^-)$ and $(F_n^+ \Rightarrow F_0^+)$, where $F_n^- \stackrel{\Delta}{=} (F_n)^-$, etc. By the "portmanteau" theorem (see, *e.g.*, Theorem 6.1 in [14, p. 40]), it is sufficient to show that for every open set $G \subset [0, 1]$,

$$\liminf_n \int_G \mathrm{d}F_n^-(a) \geq \int_G \mathrm{d}F_0^-(a), \tag{55}$$

$$\liminf_n \int_G \mathrm{d}F_n^+(a) \geq \int_G \mathrm{d}F_0^+(a). \tag{56}$$

To prove (55), let $f_1 : [0, 1]^2 \to [0, 1]$ be such that $f_1(a_1, a_2) = a_1 * a_2$. Then, we can write

$$P_n^-(G) \stackrel{\Delta}{=} \int_G \mathrm{d}F_n^-(a) = \iint\limits_{f_1^{-1}(G)} \mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2),$$

which follows from the density evolution equation

$$F_n^-(a) = \iint\limits_{a_1 * a_2 \leq a} \mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2)$$

that was proved as part of Proposition 3. We note that (i) the pre-image $f_1(G) \subset [0, 1]^2$ is an open set since the function f is a continuous and (ii) the product measure $F_n \times F_n$ converges weakly to $F_0 \times F_0$ [15, p. 21, Th. 3.2]; so, again by the portmanteau theorem,

$$\liminf_n \iint\limits_{f_1^{-1}(G)} \mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2) \geq \iint\limits_{f_1^{-1}(G)} \mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2).$$

Since

$$\iint\limits_{f_1^{-1}(G)} \mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2) = \int_G \mathrm{d}F_0^-(a),$$

the proof is complete.

The second condition (56) can be proved in a similar manner. We will sketch the steps of the proof but leave out the details. The relevant form of the density evolution equation is now

$$F_n^+(a) = \iint\limits_{(a_1a_2/a_1*a_2)\le a} (a_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2) + \iint\limits_{(\overline{a}_1a_2/\overline{a}_1*a_2)\le a} (\overline{a}_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2).$$

We define $f_{21}(a_1, a_2) = a_1a_2/a_1 * a_2$ and $f_{22}(a_1, a_2) = \overline{a}_1a_2/\overline{a}_1 * a_2$, and write

$$P_n^+(G) \stackrel{\Delta}{=} \int_G \mathrm{d}F_n^+(a) = \iint\limits_{f_{21}^{-1}(G)} (a_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2) + \iint\limits_{f_{22}^{-1}(G)} (\overline{a}_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2).$$

Next, we note that, by a general result on the preservation of weak convergence [15, Th. 5.1],

$$(a_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2) \Rightarrow (a_1 * a_2)\,\mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2),$$
$$(\overline{a}_1 * a_2)\,\mathrm{d}F_n(a_1)\,\mathrm{d}F_n(a_2) \Rightarrow (\overline{a}_1 * a_2)\,\mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2).$$

(The important point here is that the functions $(a_1 * a_2)$ and $(\overline{a}_1 * a_2)$ are uniformly continuous and bounded over the domain $(a_1, a_2) \in [0, 1]^2$. The claimed convergences follow readily from the definition of weak convergence.) The proof is completed by writing

$$\begin{aligned}\liminf_n P_n^+(G) &\ge \iint\limits_{f_{21}^{-1}(G)} (a_1 * a_2)\,\mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2) + \iint\limits_{f_{22}^{-1}(G)} (\overline{a}_1 * a_2)\,\mathrm{d}F_0(a_1)\,\mathrm{d}F_0(a_2)\\ &= \int_G \mathrm{d}F_0^+(a).\end{aligned}$$

□

Lemma 12: For $\delta > 0$, $\Delta(\delta) > 0$.

Proof: Fix $\delta > 0$. The set $\mathcal{M}_\delta$ can be written as the pre-image of a closed set under a continuous function: $\mathcal{M}_\delta = C^{-1}([\delta, M])$. Hence, by a general result about continuity ([16, p. 86, Th. 4.8]), $\mathcal{M}_\delta$ is closed; and, being a subset of the compact set [0, 1], it is compact ([16, p. 37, Th. 2.35]). Since C is continuous and $\mathcal{M}_\delta$ is compact, the "inf" in (48) is achieved by some $F_0 \in \mathcal{M}_\delta$ ([16, p. 89, Th. 4.16]): $\Delta(\delta) = C(F_0)$. Since $V(F_0) \ge \delta > 0$, F_0 is not extreme, so by Theorem 1′, $C(F_0) > 0$. □

ACKNOWLEDGMENT

The author would like to thank the Associate Editor and anonymous referees for helpful comments and suggestions.

REFERENCES

[1] I. Kontoyiannis and S. Verdú, "Optimal lossless compression: Source varentropy and dispersion," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013, pp. 1739–1743.
[2] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Proc. Trans. 3rd Prague Conf. Inf. Theory*, Prague, Czech Republic, 1962, pp. 689–723.
[3] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[4] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.
[5] E. Arıkan, "Source polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 899–903.
[6] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[7] J. D. Esary, F. Proschan, and D. W. Walkup, "Association of random variables, with applications," *Ann. Math. Statist.*, vol. 38, no. 5, pp. 1466–1474, 1967.
[8] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[9] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.
[10] S. H. Hassani and R. Urbanke, "Polar codes: Robustness of the successive cancellation decoder with respect to quantization," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 1962–1966.
[11] M. Alsan and E. Telatar, "A simple proof of polarization and polarization for non-stationary channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 301–305.
[12] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications—I," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973.
[13] E. Şaşoğlu, "Polar coding theorems for discrete systems," Ph.D. dissertation, Lab. De Théorie De L'Inf., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2011.
[14] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. San Francisco, CA, USA: Academic, 1967.
[15] P. Billingsley, *Convergence of Probability Measures*. New York, NY, USA: Wiley, 1968.
[16] W. Rudin, *Principles of Mathematical Analysis*. New York, NY, USA: McGraw-Hill, 1976.

Erdal Arıkan (S'84–M'79–SM'94–F'11) was born in Ankara, Turkey, in 1958. He received the B.S. degree from the California Institute of Technology, Pasadena, CA, in 1981, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1982 and 1985, respectively, all in Electrical Engineering. Since 1987 he has been with the Electrical-Electronics Engineering Department of Bilkent University, Ankara, Turkey, where he works as a professor. He is the receipient of 2010 IEEE Information Theory Society Paper Award and the 2013 IEEE W.R.G. Baker Award, both for his work on polar coding.