On the Sub-Optimality of Single-Letter Coding Over Networks

Farhad Shirani and S. Sandeep Pradhan

***Abstract*— In this paper, we establish a new bound tying together the effective length and the maximum correlation between the outputs of an arbitrary pair of Boolean functions which operate on two sequences of correlated random variables. We derive a new upper bound on the correlation between the outputs of these functions. The upper bound may find applications in problems in many areas which deal with common information. We build upon Witsenhausen's result [1] on maximum correlation. The present upper bound takes into account the effective length of the Boolean functions in characterizing the correlation. We use the new bound to characterize the communication-cooperation tradeoff in multi-terminal communications. We investigate binary block-codes (BBC). A BBC is defined as a vector of Boolean functions. We consider an ensemble of BBCs which is randomly generated using single-letter distributions. We characterize the vector of dependency spectrums of these BBCs. We use this vector to bound the correlation between the outputs of two distributed BBCs. Finally, the upper bound is used to show that the large blocklength single-letter coding schemes studied in the literature are sub-optimal in various multi-terminal communication settings.**

***Index Terms*— Random coding, source coding, channel coding, maximum correlation.**

I. Introduction

MOST of the coding strategies developed in information theory are based on random code ensembles which are constructed using independent identically distributed (IID) random variables [2]. The codes associated with different nodes in the network are mutually independent. Moreover, the blocklength associated with these codes are asymptotically large. One can use the law of large numbers to characterize their performance in terms of information quantities that are the functionals of the underlying distribution used to construct the codes. They are called single-letter characterizations [3]. Although the original problem is to optimize performance of codes with asymptotically large blocklengths, the solution is characterized by a functional (such as mutual information) of just *one* realization of the source or the channel under consideration. At a high level, this is very similar to the characterizations of the probability of large deviations studied in probability theory [4], the simplest example being the Chernoff Bound. In network source coding problems, one can do better covering in larger dimensions so that source redundancy can be exploited more efficiently, and the sources can be represented and reconstructed with less distortion. In network channel coding problems, better packing can be done in larger dimensions so that the channel noise can be tackled in a better fashion. In summary, the efficiency of fundamental tasks of communication such as covering and packing increases as we increase the dimension more or less all the way to infinity. Recall that in point-to-point communication the key objective is to perform these tasks efficiently. Although the individual codewords are constructed using IID random variables, since the encoding and decoding processes are accomplished in large dimensions using the so-called typical sets, there is memory of arbitrary lengths among the source reconstruction vectors in source coding and channel input vectors in channel coding.

In network communication, one needs to (a) remove redundancy among correlated information sources [5] in a distributed manner in the source coding problems, and (b) induce redundancy among distributed terminals to facilitate [6] cooperation among them. For example, in the network source coding problems such as distributed data compression, the objective is to exploit the statistical correlation of the distributed information sources. Similarly, in the network channel coding problems, such as the interference networks and broadcast networks, correlation of information among different terminals are induced for better cooperation among them [7]. At a high level, efficient information coding strategies in networks exploit statistical correlation among distributed information sources or induce statistical correlation among information accessed by terminals in the network. Of course, the basic tasks such as packing and covering at every terminal need to be accomplished as well. Statistical correlation among information shared by the terminals in the network can be viewed as a resource that needs to be efficiently managed. Distributed statistical correlation can facilitate cooperation among the terminals in networks.

It was first observed by Gács, Körner and Witsenhausen [1], [8] that coding over blocks decreases distributed correlation. Consider a pair of distributed sources X and Y with a joint probability distribution P_{XY}. Let us assume that the joint distribution does not have any zeros. There are two distributed agents who observe the sources as shown in Figure 1. The observations include n memoryless copies of the sources. The objective is to encode the observations into one bit. Let e and f denote the encoding functions associated with the two encoders. Loosely speaking,

Manuscript received April 18, 2017; revised February 27, 2019; accepted April 20, 2019. Date of publication May 17, 2019; date of current version September 13, 2019. This work was supported by NSF under Grant CCF-1422284. This paper was presented in part at the 2017 IEEE International Symposium on Information Theory.

F. Shirani is with the Department of Electrical and Computer Engineering, New York University, New York, NY 10003 USA (e-mail: fshirani@umich.edu).

S. S. Pradhan is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA.

Communicated by V. M. Prabhakaran, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2019.2917434

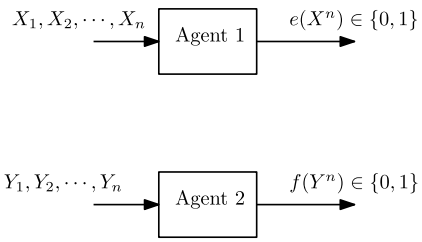

Fig. 1. Correlated Boolean decision functions.

we wish to maximize correlation between the outputs such that $H(e(X^n)) > 0$ and $H(f(Y^n)) > 0$, where $H(\cdot)$ is the entropy function. It was shown that maximum correlation is achieved when the output depends only on one of the input samples at both encoders. In fact any block mapping strictly reduces the correlation between the output bits. In summary, uncoded mappings (mappings with blocklength 1) are optimal in terms of correlation preservation. A second observation that is made in these works is that if the sources have a common component, then and only then the output bits can be made perfectly correlated. In other words, if the objective is to generate one bit at both encoders that match with probability one, then this is possible if and only if the sources have a matching component to begin with. This matching component, if it exists, is called the common information of the two sources.[1] This observation suggests that common information is very fragile. Even a small perturbation of the source distribution can produce a large change in the correlation of the output bits [11]. The study of this setup has had impact on a variety of disciplines, for instance, by taking the agents to be two encoders in the distributed source coding problem [12], or two transmitters in the interference channel problem, or Alice and Bob in a secret key-generation problem [13], [14], or two agents in a distributed control problem [15], [16].

We have taken the fundamental observation made by Gács-Körner-Witsenhausen, and developed a framework for quantitatively characterizing the correlation preserving property of any pair of encoding functions with arbitrary blocklengths. It is harder to preserve distributed correlation in larger dimensions than in smaller dimensions. In other words, short blocklength codes are able to preserve and induce correlation in a distributed fashion in a better way than larger blocklength codes. At this point, it is worth noting that this strange behavior leads to a tension: to perform covering and packing we need large blocklength codes, whereas to preserve and induce correlation we need short blocklength codes. The overall network performance may be optimized by codes whose blocklength is some sweet finite value. Toward a characterization of this trade-off, consider a source encoder at a terminal in a network that maps n samples of an information source into k bits (for some n and k). Then the blocklength of the encoder is n. Suppose that each of the output bits depends only on (αn) samples of the input vector for some $\alpha < 1$. We can then define the *effective length* of the encoder as (αn). It is the conventional wisdom that performance of many coding strategies (characterized by a sequence of encoders and decoders with increasing blocklength) in network communication is super-additive in blocklength (e.g. see [17]). Our results state that the performance is not super-additive in effective length. In fact in network communication problems, to achieve optimality, certain components of the transmission system must have a finite effective length structure. In summary we have new a trade-off between covering and packing efficiency, and the correlation preserving/inducing ability of codes. That is, a trade-off between communication and cooperation in networks. Optimal codes have to straddle this trade-off.

In this work we make the following three contributions.

- We start with Section III. Consider a discrete memoryless source $X_1, X_2, \ldots$, with finite alphabet $\mathcal{X}$, and a generic distribution P_X. Consider a block encoder (a Boolean function) $e : \mathcal{X}^n \to \{0,1\}$, where n denotes the blocklength (see [8]). Given a pair consisting of a source and an encoder, we define its dependency spectrum (Definition 6) as a (unnormalized) 2^n-dimensional vector that captures the probabilistic as well as the functional memory structure of the pair. For example, for $n = 3$, the dependency spectrum is the following vector $[P_{000}, P_{001}, \ldots, P_{111}]$ charactering the contribution of the constant (P_{000}), single-letter (P_{001}, P_{010}, P_{100}), two-letter (P_{011}, P_{110}, P_{101}), and three-letter (P_{111}) component functions toward constructing e. Note that there are three one-letter and two-letter functions. As another example, for $n = 2$, and logical AND function $e(X^n) = X_1 \wedge X_2$, with binary uniform source, we get $P_{01} = P_{10} = P_{11} = \frac{1}{16}$, and $P_{00} = 0$. Logical AND function is two-thirds a single-letter function and one-third a two-letter function. This is a generalization of the effective length from a number to a vector.
  We use dependency spectrum to study distributed encoding of correlated sources in the following way and provide the first main result (Theorem 1 and 2) of the paper in Section IV. Consider a pair of discrete memoryless correlated sources $(X_1, Y_1), (X_2, Y_2), \ldots$ with finite alphabets $\mathcal{X} \times \mathcal{Y}$ and a generic distribution P_{XY}. Consider a pair of distributed block encoders $e : \mathcal{X}^n \to \{0,1\}$ and $f : \mathcal{Y}^n \to \{0,1\}$. We provide a characterization of the correlation between the outputs of these block encoders. In particular, we give a lower bound on the probability of disagreement between the outputs $P(e(X^n) \neq f(Y^n))$ in terms of the dependency spectra of (P_X, e) and (P_Y, f), and the correlation of the sources given by $P(X \neq Y)$. Roughly speaking, this is a quantitative characterization of the trade-off between the effective length of the encoders and the output correlation.
- We use this characterization to analyze a large class of sequence of code ensembles studied in the information theory literature in Section V. We call this class Single-letter coding ensembles (SLCE) (Definition 8). For example, this class subsumes Shannon-style IID unstructured code ensembles as well as structured linear code ensembles. Most, if not all, of the coding theorems of information theory that characterize asymptotic performance limits are based on this class. In short, a code ensemble is an infinite sequence

[1]This is also characterized via Rényi maximum correlation [9], [10]. If the sources have a common information then the maximum correlation of the sources is 1.

of collections of block encoders (indexed by blocklength $n = 1, 2, \ldots$) along with a probability distribution on the collection. We provide the following second main result of the paper (Theorem 3 and 4). For a discrete memoryless source $X_1, X_2, \ldots$, with finite alphabet $\mathcal{X}$ and a generic distribution P_X, the output of the SLCE has the following structure: for any fixed $1 < m < \infty$, the probability that the contribution of m-letter functions toward constructing e is large approaches zero as n tends to infinity. In other words, with high probability SLCE produces encoders which has either a single-letter component or an infinite-letter component. We call this the 1-∞ law. This is a structural deficiency of SLCE. Moreover, when applied on a pair of correlated sources, (X, Y), we provide a high probability upper bound on the correlation of the outputs.

- We study two multi-terminal communication problems in Section VI: transmission of correlated sources over the interference channels and the multiple-access channels. These two problems have been studied in the literature extensively. Inner bounds to the asymptotic performance limits (achievable performance) based on SLCE have been developed in the literature. These bounds have been the de facto performance limits since the 1980s. We provide a novel coding technique based on finite-length codes along with two examples. Using the structural results from the previous section, we show analytically that this coding technique outperforms the inner bounds derived using arbitrarily long codes constructed using SLCE. This is the third main result of the paper (Proposition 5, 7 and 8). In other words, specifically designed finite-length codes can perform better than SLCE.

We discuss some related prior works. In the literature, it has been shown that the loss in correlation caused by the application of large effective-length codes causes a discontinuity in the performance of schemes using such codes in some multi-terminal problems. This was first observed in the Berger-Tung achievable rate-distortion region for the problem of distributed source coding [18] [11]. It was noted that when the common information is available to the two encoders in the distributed source coding problem, the performance is discontinuously better than when the common information is replaced with highly correlated components. In [12], we argued that the discontinuity in performance is due to the fact that the encoding functions in the Berger-Tung scheme preserve common information, but are unable to preserve correlation between highly correlated components. We proposed a new coding scheme, and derived an improved achievable rate-distortion region for the two user distributed source coding problem [19]. The new strategy uses a concatenated coding scheme which consists of one layer of codes with finite effective-length, and one layer of codes with asymptotically large effective-lengths.

II. Notation

In this section, we introduce the notation used in this paper. We represent random variables by capital letters such as X, U. Sets are denoted by calligraphic letters such as $\mathcal{X}, \mathcal{U}$. Particularly, the set of natural numbers and real numbers are shown by $\mathbb{N}$, and $\mathbb{R}$, respectively. For a random variable X, the corresponding probability space is $(\mathcal{X}, \mathbf{F}_X, P_X)$, where $\mathbf{F}$ is the underlying σ-field. The set of all subsets of $\mathcal{X}$ is written as $2^{\mathcal{X}}$. There are three different notations used for different classes of vectors. For random variables, the n-length vector $(X_1, X_2, \cdots, X_n)$, $X_i \in \mathcal{X}$ is denoted by $X^n \in \mathcal{X}^n$. For the vector of functions $(e_1(X), e_2(X), \cdots, e_n(X))$ we use the notation $\underline{e}(X)$. The binary string $(i_1, i_2, \cdots, i_n)$, $i_j \in \{0, 1\}$ is written as $\mathbf{i}$. As an example, the set of functions $\{\underline{e_{\mathbf{i}}}(X^n) | \mathbf{i} \in \{0, 1\}^n\}$ is the set of n-length vectors of functions $(e_{1,\mathbf{i}}, e_{2,\mathbf{i}}, \cdots, e_{n,\mathbf{i}})$ operating on the vector $(X_1, X_2, \cdots, X_n)$ each indexed by an n-length binary string $(i_1, i_2, \cdots, i_n)$. The vector of binary strings $(\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_n)$ denotes the standard basis for the n-dimensional space (e.g. $\mathbf{i}_1 = (0, 0, \cdots, 0, 1)$). The vector of random variables $(X_{j_1}, X_{j_2}, \cdots, X_{j_k})$, $j_i \in [1, n]$, $j_i \neq j_k$, is denoted by $X_{\mathbf{i}}$, where $i_{j_l} = 1, \forall l \in [1, k]$. For example, take $n = 3$, the vector (X_1, X_3) is denoted by X_{101}, and the vector (X_1, X_2) by X_{110}. Particularly, $X_{\mathbf{i}_j} = X_j$, $j \in [1, n]$. Also, for $\mathbf{t} = \underline{1}$, the all-ones vector, $X_{\mathbf{t}} = X^n$. For two binary strings $\mathbf{i}, \mathbf{j}$, we write $\mathbf{i} \leq \mathbf{j}$ if and only if $i_k \leq j_k, \forall k \in [1, n]$. Also, we write $\mathbf{i} < \mathbf{j}$ if $\mathbf{i} \leq \mathbf{j}$ and $\mathbf{i} \neq \mathbf{j}$. For a binary string $\mathbf{i}$ we define $N_{\mathbf{i}} \triangleq \omega_H(\mathbf{i})$, where ω_H denotes the Hamming weight. Lastly, the vector $\sim \mathbf{i}$ is the element-wise complement of $\mathbf{i}$. We use $\oplus_k$ to denote addition modulo k, where $k \in \mathbb{N}$.

III. The *Effective-Length* of an Encoder

In this section, we define a set of parameters which measure the effective-length of an encoding function. We consider general Boolean functions, and find a decomposition of these functions into components which operate over specific subsets of the input sequence. The proposed decomposition builds upon the analysis in [1]. The first subsection summarizes the well-known results. The second subsection contains some new results (Proposition 1 and 3).

A. Mathematical Preliminaries

It turns out that when the input sequence is a vector of independent binary symmetric variables, the decomposition that we provide is equivalent to the Fourier transform of Boolean functions [20]. The Fourier transform does not take into account the underlying probability distribution of the sources. The connections between Fourier transforms and the correlation between outputs of pairs of functions was previously studied in [21], where the decidability of the non-interactive simulation problem was considered. We propose the decomposition for general finite input alphabets with arbitrary input distributions. We only consider encoders with binary outputs.[2] The encoder can be viewed as a vector of Boolean functions. Based on the decomposition, we define a generalization of the effective-length called the '*dependency spectrum*' of a Boolean function.

We proceed by formally defining the problem. We assume that two correlated DMS's are being fed to two arbitrary encoders, and analyze the correlation between the outputs of

[2]The analysis provided in this paper can be generalized to arbitrary finite output alphabets. The interested reader can refer to Section 7 in [1].

these encoders. The following gives the formal definition for DMS's.

Definition 1. *(X, Y) is called a pair of DMS's if we have $P_{X^n,Y^n}(x^n, y^n) = \prod_{i\in[1,n]} P_{X_i,Y_i}(x_i, y_i), \forall n \in \mathbb{N}, x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$, where $P_{X_i,Y_i} = P_{X,Y}, \forall i \in [1, n]$, for some joint distribution $P_{X,Y}$.*

Akin to the results presented in [1] and [8], we restrict our attention to the binary block encoders (BBE), which are defined below.

Definition 2. *A Binary-Block-Encoder is characterized by the triple $(\underline{e}, \mathcal{X}, n)$, where $\underline{e}$ is a mapping $\underline{e} : \mathcal{X}^n \to \{0, 1\}^n$, $\mathcal{X}$ is a finite set, and n is an integer.*

We refer to a BBE by its corresponding mapping $\underline{e}$. The mapping $\underline{e}$ can be viewed as a vector of functions $(e_i)_{i\in[1,n]}$, where $e_i : \mathcal{X}^n \to \{0, 1\}$. We convert the problem of analyzing a BBE into one where the encoder is a binary real-valued function. Converting the discrete-valued encoding function into a real-valued one is crucial since it allows us to use the rich set of tools available in functional analysis. We present a summary of the functional analysis apparatus used in this work.

Definition 3. *Fix a discrete memoryless source X, and a BBE $\underline{e} : \mathcal{X}^n \to \{0, 1\}^n$. Let $P\,(e_i(X^n) = 1) = q_i$. For each Boolean function $e_i, i \in [1, n]$, the real-valued function corresponding to e_i is defined as follows*:

$$\tilde{e}_i(X^n) = \begin{cases} 1 - q_i, & \text{if } e_i(X^n) = 1, \\ -q_i. & \text{otherwise.} \end{cases} \tag{1}$$

Remark 1. *Note that $\tilde{e}_i, i \in [1, n]$ has zero mean and variance $q_i(1 - q_i)$.*

The random variable $\tilde{e}_i(X^n)$ has finite variance on the probability space $(\mathcal{X}^n, 2^{\mathcal{X}^n}, P_{X^n})$. The set of all such functions is denoted by $\mathcal{H}_{X,n}$. More precisely, we define $\mathcal{H}_{X,n} \triangleq L_2(\mathcal{X}^n, 2^{\mathcal{X}^n}, P_{X^n})$ as the separable Hilbert space of all functions $\tilde{h} : \mathcal{X}^n \to \mathbb{R}$ with inner product given by $\tilde{h} \cdot \tilde{g} = \sum_{x^n} \tilde{h}(x^n)\tilde{g}(x^n)P_{X^n}(x^n)$. Since X is a DMS, the isomorphism relation

$$\mathcal{H}_{X,n} = \mathcal{H}_{X,1} \otimes \mathcal{H}_{X,1} \cdots \otimes \mathcal{H}_{X,1} \tag{2}$$

holds [22], where $\otimes$ indicates the tensor product.

Example 1. Let $n = 1$. Let $\mathcal{X} = \{0, 1\}$. The Hilbert space $\mathcal{H}_{X,1}$ is the space of all functions $\tilde{h} : \mathcal{X} \to \mathbb{R}$. The space is spanned by the two linearly independent functions $\tilde{h}_1(X) = \mathbb{1}_{\{X=1\}}$ and $\tilde{h}_2(X) = \mathbb{1}_{\{X=0\}}$. We conclude that the space is two-dimensional.

As a reminder, the following defines the tensor product of vector spaces.

Definition 4 ([22])**.** *Let $\mathcal{H}_i, i \in [1, n]$ be vector spaces over a field F. Also, let $\mathcal{B}_i = \{v_{i,j} | j \in [1, d_i]\}$ be the basis for $\mathcal{H}_i$ where d_i is the dimension of $\mathcal{H}_i$. Then, the tensor product space $\otimes_{i\in[1,n]}\mathcal{H}_i$ is defined as the set of elements $v = \sum_{j_1\in[1,d_1]} \sum_{j_2\in[1,d_2]} \cdots \sum_{j_n\in[1,d_n]} c_{j_1,j_2,\cdots,j_n} v_{j_1} \otimes v_{j_2} \cdots \otimes v_{j_n}$.*

Remark 2. *The tensor product operation in $\mathcal{H}_{X,n}$ is real multiplication (i.e. $f_1, f_2 \in \mathcal{H}_{X,1} : f_1(X_1) \otimes f_2(X_2) \triangleq f_1(X_1)f_2(X_2)$). So, if $\{f_i(X) | i \in [1, d]\}$ is a basis for $\mathcal{H}_{X,1}$ when $|\mathcal{X}| = d$, a basis for $\mathcal{H}_{X,n}$ would be the set of all the real multiplications of these basis elements: $\{\Pi_{j\in[1,n]} f_{i_j}(X_j), i_j \in [1, d]\}$.*

Example 1 gives a decomposition of the space $\mathcal{H}_{X,1}$ for binary input alphabets. Next, we introduce a decomposition of $\mathcal{H}_{X,1}$ for general alphabets which turns out to be very useful. Particularly, we argue that every Boolean function $\tilde{e}(X) \in \mathcal{H}_{X,1}$ can be written as a summation of two functions, one function whose expected value is 0, and a constant function. More precisely, let $\mathcal{I}_{X,1}$ be the subset of all functions of X which have 0 mean, and let $\gamma_{X,1}$ be the set of constant real functions of X. $\mathcal{I}_{X,1}$ and $\gamma_{X,1}$ are linear subspaces of $\mathcal{H}_{X,1}$. $\mathcal{I}_{X,1}$ is the null space of the functional which takes an arbitrary function $\tilde{f} \in \mathcal{H}_{X,1}$ to its expected value $\mathbb{E}_X(\tilde{f})$. The null space of any non-zero linear functional is a hyper-space in $\mathcal{H}_{X,1}$. So, $\mathcal{I}_{X,1}$ is a $(|\mathcal{X}| - 1)$-dimensional subspace of $\mathcal{H}_{X,1}$. On the other hand, $\gamma_{X,1}$ is a one dimensional subspace which is not contained in $\mathcal{I}_{X,1}$. It is spanned by the function $\tilde{g}(X) \equiv 1$. Consider an arbitrary element $\tilde{f} \in \mathcal{H}_{X,1}$. One can write $\tilde{f} = \tilde{f}_1 + \tilde{f}_2$ where $\tilde{f}_1 = \tilde{f} - \mathbb{E}_X(\tilde{f}) \in \mathcal{I}_{X,1}$, and $\tilde{f}_2 = \mathbb{E}_X(\tilde{f}) \in \gamma_{X,1}$. Hence, $\mathcal{H}_{X,1} = \mathcal{I}_{X,1} \oplus \gamma_{X,1}$ gives a decomposition of $\mathcal{H}_{X,1}$. Replacing $\mathcal{H}_{X,1}$ with $\mathcal{I}_{X,1} \oplus \gamma_{X,1}$ in (2), we have:

$$\begin{aligned} \mathcal{H}_{X,n} &= \otimes_{i=1}^n \mathcal{H}_{X,1} = \otimes_{i=1}^n (\mathcal{I}_{X,1} \oplus \gamma_{X,1}) \\ &\overset{(a)}{=} \oplus_{\mathbf{i}\in\{0,1\}^n} (\mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}), \end{aligned} \tag{3}$$

where

$$\mathcal{G}_j = \begin{cases} \gamma_{X,1} & j = 0, \\ \mathcal{I}_{X,1} & j = 1, \end{cases}$$

and, in (a), we have used the distributive property of tensor products over direct sums. Using equation (3) we can define the following:

Definition 5. *For any $\tilde{e} \in \mathcal{H}_{X,n}, n \in \mathbb{N}$, define the decomposition $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$, where $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$. Then, $\tilde{e}_{\mathbf{i}}$ is the component of $\tilde{e}$ which is only a function of $\{X_{i_j} | i_j = 1\}$. The collection $\{\tilde{e}_{\mathbf{i}} | \sum_{j\in[1,n]} i_j = k\}$, is called the set of k-letter components of $\tilde{e}$. The vector $(\tilde{e}_{\mathbf{i}})_{\mathbf{i}\in\{0,1\}^n}$ is called the real decomposition vector corresponding to $\tilde{e}$.*

In order clarify the notation, we provide the following two examples.

Example 2. Let (X_1, X_2) be two independent symmetric binary random variables. Assume $e(X_1, X_2) = X_1 \oplus X_2$ is the binary addition function. In this example $P(e = 1) = \frac{1}{2}$. The corresponding real function is given as follows:

$$\tilde{e}(X_1, X_2) = \begin{cases} -\frac{1}{2} & X_1 + X_2 \in \{0, 2\}, \\ \frac{1}{2} & X_1 + X_2 = 1, \end{cases}$$

Using Lagrange interpolation [23], we can write $\tilde{e}$ as follows:

$$\tilde{e} = -\frac{1}{2}(X_1 + X_2 - 2)(X_1 + X_2) - \frac{1}{4}(X_1 + X_2 - 1)(X_1 + X_2 - 2) - \frac{1}{4}(X_1+X_2)(X_1+X_2 - 1) = -X_1^2 - X_2^2 - 2X_1X_2 + 2X_1 + 2X_2 - \frac{1}{2}.$$

The decomposition of $\tilde{e}$ in the form given in (3) is

$$\tilde{e}_{1,1} = X_1 + X_2 - 2X_1X_2 - \frac{1}{2} = -\frac{1}{2}(1 - 2X_1)(1 - 2X_2),$$
$$\tilde{e}_{1,0} = -X_1^2 + X_1 = X_1(1 - X_1) \overset{(a)}{=} 0,$$
$$\tilde{e}_{0,1} = -X_2^2 + X_2 = X_2(1 - X_2) \overset{(a)}{=} 0,$$
$$\tilde{e}_{0,0} = 0.$$

where (a) holds since the input is chosen from $\{0, 1\}$. Note that $\tilde{e}$ has a single non-zero component in its decomposition. This component is the two-letter function $\tilde{e}_{1,1} \in \mathcal{I}_{X,1} \otimes \mathcal{I}_{X,1}$. This is to be expected since the binary addition of two symmetric variables is independent of each variable. So there are no single-letter components. In fact one can verify this directly as follows:

$$\mathbb{E}_{X_2|X_1}(\tilde{e}|X_1) = X_1 - X_1 = 0, \quad \mathbb{E}_{X_1|X_2}(\tilde{e}|X_2) = X_2 - X_2 = 0.$$

Remark 3. *In the previous example, we found that the binary summation of two independent binary symmetric variables is a two-letter function (i.e. it only has a two-letter component). However, this is not true when the source is not symmetric. When $P(X = 1) \neq P(X = 0)$, the output of the summation is not independent of each of the inputs. One can show that the single-letter components of the summation are non-zero in this case.*

Example 3. Let $e(X_1, X_2) = X_1 \wedge X_2$ be the binary logical AND function. The corresponding real function is:

$$\tilde{e}(X_1, X_2) = \begin{cases} -\frac{1}{4} & (X_1, X_2) \neq (1, 1), \\ \frac{3}{4} & (X_1, X_2) = (1, 1). \end{cases}$$

Lagrange interpolation gives $\tilde{e} = X_1X_2 - \frac{1}{4}$. The decomposition is given by:

$$\tilde{e}_{1,1} = (X_1 - \frac{1}{2})(X_2 - \frac{1}{2}), \qquad \tilde{e}_{1,0} = \frac{1}{2}(X_1 - \frac{1}{2}),$$
$$\tilde{e}_{0,1} = \frac{1}{2}(X_2 - \frac{1}{2}), \qquad \tilde{e}_{0,0} = 0.$$

The variances of these functions are given below:

$$Var(\tilde{e}) = \frac{3}{16}, \qquad Var(\tilde{e}_{0,1}) = Var(\tilde{e}_{1,0}) = Var(\tilde{e}_{1,1}) = \frac{1}{16}.$$

As we shall see in the next sections, these variances play a major role in determining the correlation preserving properties of $\tilde{e}$. In the perspective of the effective-length, the function $\tilde{e}$ has $\frac{2}{3}$ of its variance distributed between $\tilde{e}_{0,1}$, and $\tilde{e}_{1,0}$ which are single-letter functions, and $\frac{1}{3}$ of the variance is on $\tilde{e}_{1,1}$ which is a two-letter function.

Similar to the previous examples, for arbitrary $\tilde{e} \in \mathcal{H}_{X,n}, n \in \mathbb{N}$, we can find a decomposition $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$, where $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$. We can characterize $\tilde{e}_{\mathbf{i}}$ in terms of products of the basis elements of $\mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$ as follows.

Lemma 1. *For an arbitrary input alphabet $\mathcal{X}$, let $\tilde{h}_l(X), l \in \{1, 2, \cdots, |\mathcal{X}| - 1\}$ be an orthogonal basis for $\mathcal{I}_{X,1}$, such that $E(\tilde{h}_l^2(X)) = q(1 - q), \forall l \in \{1, 2, \cdots, |\mathcal{X}| - 1\}$, where $q = P(X \neq 0)$. Let $\tau = \{t : i_t = 1\}$, then*:

$$\tilde{e}_{\mathbf{i}}(X^n) = \sum_{\forall t \in \tau : l_t \in [1, |\mathcal{X}|-1]} c_{\mathbf{i},(l_t)_{t\in\tau}} \prod_{t \in \tau} \tilde{h}_{l_t}(X_t), \tag{4}$$

where $c_{\mathbf{i}}, (l_t)_{t\in\tau} \in \mathbb{R}$, and $(l_t)_{t\in\tau}$ is the sequence of l_t's for $t \in \tau$.

Proof. Follows from Definition 4. □

Example 4. Let $\mathcal{X} = \{0, 1\}$. Since $\mathcal{G}_{i_j}$'s, $j \in [1, n]$ take values from the set $\{\mathcal{I}_{X,1}, \gamma_{X,1}\}$, they are all one-dimensional. Let $\tilde{h}$ be defined as follows:

$$\tilde{h}(X) = \begin{cases} 1 - q, & \text{if } X = 1, \\ -q. & \text{if } X = 0, \end{cases} \tag{5}$$

where $q \triangleq P(X = 1)$. Then, the single element set $\{\tilde{h}(X)\}$ is a basis for $\mathcal{I}_{X,1}$. Hence, using the previous lemma:

$$\tilde{e}_{\mathbf{i}}(X^n) = c_{\mathbf{i}} \prod_{t : i_t = 1} \tilde{h}(X_t), \tag{6}$$

where $c_{\mathbf{i}} \in \mathbb{R}$.

B. Properties of the Real Decomposition

The dependency spectrum and effective length of an arbitrary Boolean function are defined below.

Definition 6. *For a function $e : \mathcal{X}^n \to \{0, 1\}$, with real decomposition vector $(\tilde{e}_{\mathbf{i}})_{\mathbf{i}\in\{0,1\}^n}$, the dependency spectrum is defined as the vector $(\mathbf{P_i})_{\mathbf{i}\in\{0,1\}^n}$ of the variances, where $\mathbf{P_i} = Var(\tilde{e}_{\mathbf{i}}), \mathbf{i} \in \{0, 1\}^n$. The effective length is defined as the expected value $\widetilde{\mathbf{L}} = \frac{1}{n}\sum_{\mathbf{i}\in\{0,1\}^n} w_H(\mathbf{i}) \cdot \mathbf{P_i}$, where $w_H(\cdot)$ is the Hamming weight.*

Remark 4. *The dependency spectrum $(\mathbf{P_i})_{\mathbf{i}\in\mathbb{N}}$ characterizes the 'effect' of each component $\tilde{e}_{\mathbf{i}}$ on the output of the function $\tilde{e}$. A relevant work can be found in [24] where the output correlation of two functions with i.i.d inputs is characterized using the singular values vector of the matrix $P_X^{-\frac{1}{2}\otimes_n} P_{X,Y}^{\otimes_n} P_Y^{-\frac{1}{2}\otimes_n}$.*

In the next proposition, we show that the $\tilde{e}_{\mathbf{i}}$'s are uncorrelated and we derive an expression for $\mathbf{P_i}$ using the notation in Lemma 1.

Proposition 1. *Let X^n be a sequence of independent and identically distributed (i.i.d.) random variables and let $(\tilde{e}_{\mathbf{i}})_{\mathbf{i}\in\{0,1\}^n}$ be the real decomposition vector corresponding to the Boolean function $e(X^n)$. Define $\mathbf{P_i}$ as the variance of $\tilde{e}_{\mathbf{i}}(X_{\mathbf{i}})$. The following hold:*
1) $\mathbb{E}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{j}}) = 0, \mathbf{i} \neq \mathbf{j}$, in other words $\tilde{e}_{\mathbf{i}}$'s are uncorrelated.
2) $\mathbf{P_i} = \mathbb{E}(\tilde{e}_{\mathbf{i}}^2) = \sum_{\forall t\in\tau : l_t \in [1,|\mathcal{X}|-1]} c^2_{\mathbf{i},(l_t)_{t\in\tau}} (q(1 - q))^{w_H(\mathbf{i})}$, where w_H is the Hamming weight function. Particularly, if $\mathcal{X} = \{0, 1\}$, then $\mathbf{P_i} = \mathbb{E}(\tilde{e}_{\mathbf{i}}^2) = c_{\mathbf{i}}^2 (q(1 - q))^{w_H(\mathbf{i})}$.

Proof. 1) follows by direct calculation. 2) holds from the independence of X_i's. □

In the next lemma we find a characterization of $\tilde{e}_{\mathbf{i}}, \mathbf{i} \in \{0,1\}^n$ for general $\tilde{e}$.

Proposition 2. *Let X^n be a sequence of i.i.d. random variables and let $(\tilde{e}_{\mathbf{i}})_{\mathbf{i}\in\{0,1\}^n}$ be the real decomposition vector corresponding to the Boolean function $e(X^n)$. $\tilde{e}_{\mathbf{i}} = \mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \tilde{e}_{\mathbf{j}}$ gives the orthogonal decomposition of $\tilde{e}$ into the Hilbert spaces $\mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \cdots \otimes \mathcal{G}_{i_n}, \mathbf{i} \in \{0,1\}^n$.*

Proof. We prove that the $\tilde{e}_{\mathbf{i}}$ given in the lemma are indeed the decomposition into the components of the direct sum. Equivalently, we show that 1) $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$, and 2) $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}, \forall \mathbf{i} \in \{0,1\}^n$.

First we check the equality $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$. Let $\mathbf{t}$ denote the n-length vector whose elements are all ones. We have:

$$\tilde{e}_{\mathbf{t}} = \mathbb{E}_{X^n|X_{\mathbf{t}}}(\tilde{e}|X_{\mathbf{t}}) - \sum_{\mathbf{i}<\mathbf{t}} \tilde{e}_{\mathbf{i}} \overset{(a)}{\Rightarrow} \tilde{e}_{\mathbf{t}} + \sum_{\mathbf{i}<\mathbf{t}} \tilde{e}_{\mathbf{i}} = \tilde{e} \overset{(b)}{\Rightarrow} \tilde{e} = \sum_{\mathbf{i}\in\{0,1\}^n} \tilde{e}_{\mathbf{i}},$$

where in (a) we have used 1) $X_{\mathbf{t}} = X^n$ and 2) for any function $\tilde{f}$ of X^n, $\mathbb{E}_{X^n|X^n}(\tilde{f}|X^n) = \tilde{f}$, and (b) holds since $\mathbf{i} < \mathbf{t} \Leftrightarrow \mathbf{i} \neq \mathbf{t}$. It remains to show that $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}, \forall \mathbf{i} \in \{0,1\}^n$. The next lemma provides a means to verify this property.

Lemma 2. *Fix $\mathbf{i} \in \{0,1\}^n$, define $\mathcal{A}_0 \triangleq \{s|i_s = 0\}$, and $\mathcal{A}_1 \triangleq \{s|i_s = 1\}$. Then, $\tilde{f}$ is an element of $\mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$ if and only if (1) it is constant in all X_s, $s \in \mathcal{A}_0$, and (2) $\mathbb{E}_{X^n|X_{\sim \mathbf{i}_s}}(\tilde{f}|X_{\sim \mathbf{i}_s}) = 0$ for all s, when $s \in \mathcal{A}_1$.*

Proof. Please refer to the appendix. □

Returning to the original problem, it is enough to show that $\tilde{e}_{\mathbf{i}}$'s satisfy the conditions in Lemma 2. We prove the stronger result presented in the next lemma.

Lemma 3. *Let X^n be a sequence of i.i.d. random variables and let $(\tilde{e}_{\mathbf{i}})_{\mathbf{i}\in\{0,1\}^n}$ be the real decomposition vector corresponding to the Boolean function $e(X^n)$. The following hold:*
1) $\forall \mathbf{i}, \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}) = 0$.
2) $\forall \mathbf{i} \leq \mathbf{k}$, we have $\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{i}}|X_{\mathbf{k}}) = \tilde{e}_{\mathbf{i}}$.
3) $\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{k}}) = 0$, for $\mathbf{i} \neq \mathbf{k}$.
4) $\forall \mathbf{k} \leq \mathbf{i} : \mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{i}}|X_{\mathbf{k}}) = 0$.

Proof. Please refer to the appendix. □

The second condition in Lemma 3 is equivalent to condition (2) in Lemma 2. The fourth condition in Lemma 3 is equivalent to condition (1) in Lemma 2. Using Lemma 2 and 3, we conclude that $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}, \forall \mathbf{i} \in \{0,1\}^n$. This completes the proof of Proposition 2. □

The following example clarifies the notation used in Proposition 2.

Example 5. Consider the case where $n = 2$. We have the following decomposition of $\mathcal{H}_{X,2}$:

$$\mathcal{H}_{X,2} = (\mathcal{I}_{X,1} \otimes \mathcal{I}_{X,1}) \oplus (\mathcal{I}_{X,1} \otimes \gamma_{X,1}) \oplus (\gamma_{X,1} \otimes \mathcal{I}_{X,1}) \oplus (\gamma_{X,1} \otimes \gamma_{X,1}). \tag{7}$$

Let $\tilde{e}(X_1, X_2)$ be an arbitrary function in $\mathcal{H}_{X,2}$. The decomposition of $\tilde{e}$ in the form given in (7) is as follows:

$$\begin{aligned}
\tilde{e} &= \tilde{e}_{1,1} + \tilde{e}_{1,0} + \tilde{e}_{0,1} + \tilde{e}_{0,0},\\
\tilde{e}_{1,1} &= \tilde{e} - \mathbb{E}_{X_2|X_1}(\tilde{e}|X_1) - \mathbb{E}_{X_1|X_2}(\tilde{e}|X_2)\\
&\quad + \mathbb{E}_{X_1,X_2}(\tilde{e}) \in \mathcal{I}_{X,1} \otimes \mathcal{I}_{X,1},\\
\tilde{e}_{1,0} &= \mathbb{E}_{X_2|X_1}(\tilde{e}|X_1) - \mathbb{E}_{X_1,X_2}(\tilde{e}) \in \mathcal{I}_{X,1} \times \gamma_{X,1},\\
\tilde{e}_{0,1} &= \mathbb{E}_{X_1|X_2}(\tilde{e}|X_2) - \mathbb{E}_{X_1,X_2}(\tilde{e}) \in \gamma_{X,1} \otimes \mathcal{I}_{X,1},\\
\tilde{e}_{0,0} &= \mathbb{E}_{X_1,X_2}(\tilde{e}) \in \gamma_{X,1} \otimes \gamma_{X,1}.
\end{aligned}$$

It is straightforward to show that each of the $\tilde{e}_{i,j}$'s, $i, j \in \{0,1\}$, belong to their corresponding subspaces. For instance, $\tilde{e}_{0,1}$ is constant in X_1, and is a 0 mean function of X_2 (i.e. $\mathbb{E}_{X_2}\left(\tilde{e}_{0,1}(x_1, X_2)\right) = 0, x_1 \in \mathcal{X}$), so $\tilde{e}_{0,1} \in \gamma_{X,1} \otimes \mathcal{I}_{X,1}$.

Lastly, we derive an expression for $\mathbf{P}_{\mathbf{i}}$ using Proposition 2:

Proposition 3. *For arbitrary $e : \mathcal{X}^n \to \{0,1\}$, let $\tilde{e}$ be the corresponding real-valued function, and let $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$ be the decomposition in the form of Equation (3). The variance of each component in the decomposition is given by the following recursive formula $\mathbf{P}_{\mathbf{i}} = \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})) - \sum_{\mathbf{j}<\mathbf{i}} \mathbf{P}_{\mathbf{j}}, \forall \mathbf{i} \in \{0,1\}^n$, where $\mathbf{P}_{\underline{0}} \triangleq 0$.*

Proof. Please refer to the appendix. □

Corollary 1. *For an arbitrary $e : \mathcal{X}^n \to \{0,1\}$ with corresponding real function $\tilde{e}$, and decomposition $\tilde{e} = \sum_{\mathbf{i}} \tilde{e}_{\mathbf{i}}$. Let the variance of $\tilde{e}$ be denoted by $\mathbf{P}$. Then, $\mathbf{P} = \sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}$.*

The corollary is a special case of Proposition 3, where we have taken $\mathbf{i}$ to be the all ones vector.

IV. Correlation Preservation in Arbitrary Encoders

Our objective is to bound the correlation preserving properties of general n-length encoding functions. As a first step, we derive bounds on the correlation between the outputs of two arbitrary Boolean functions (i.e. functions whose output is a binary scalar). For pedagogical reasons we present the results of this section in two parts. First, we consider binary input alphabets, and derive bounds on the probability of agreement of Boolean functions. Then, we extend these results to the case of non-binary input alphabets.

A. Binary Input Alphabets

We proceed with presenting the main result of this section. Let (X, Y) be a pair of binary DMS's. Consider two arbitrary Boolean functions $e : \{0,1\}^n \to \{0,1\}$ and $f : \{0,1\}^n \to \{0,1\}$. The following theorem provides an upper-bound on the probability of equality between the functions $e(X^n)$ and $f(Y^n)$.

Theorem 1. *Let $\epsilon \triangleq P(X \neq Y)$, the following bound holds:*

$$\begin{aligned}
&2\sqrt{\sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}}\sqrt{\sum_{\mathbf{i}} \mathbf{Q}_{\mathbf{i}}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \leq P(e(X^n) \neq f(Y^n))\\
&\leq 1 - 2\sqrt{\sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}}\sqrt{\sum_{\mathbf{i}} \mathbf{Q}_{\mathbf{i}}} + 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}},
\end{aligned}$$

where $N_{\mathbf{i}} \triangleq w_H(\mathbf{i})$, $\mathbf{P}_{\mathbf{i}}$ is the variance of $\tilde{e}_{\mathbf{i}}$, and $\underline{\tilde{e}}$ is the real function corresponding to $\underline{e}$, and $\mathbf{Q}_{\mathbf{i}}$ is the variance of $\tilde{f}_{\mathbf{i}}$, and finally, $C_{\mathbf{i}} \triangleq (1-2\epsilon)^{N_{\mathbf{i}}}$.

Proof. Please refer to the appendix. □

Remark 5. *The value $C_{\mathbf{i}} = (1-2\epsilon)^{N_{\mathbf{i}}}$ is decreasing with $N_{\mathbf{i}}$. So, in order to increase $P(e(X^n) \neq f(Y^n))$, most of the variance $\mathbf{P_i}$ should be distributed on $\tilde{e}_{\mathbf{i}}$ which have lower $N_{\mathbf{i}}$ (i.e. operate on smaller blocks). Particularly, the lower bound is minimized by setting*

$$\frac{\mathbf{P_i}}{\sqrt{Var(X)}} = \frac{\mathbf{Q_i}}{\sqrt{Var(Y)}} = \begin{cases} 1 & \mathbf{i} = \mathbf{i}_1, \\ 0 & otherwise. \end{cases} \tag{8}$$

This recovers the result in [1]. More precisely, if we replace the dependency spectrum in Theorem 1 by the values in Equation (8)*, we get*:

$$2(1-C_{\mathbf{i}_1})\sqrt{Var(X)Var(Y)} \leq P(e(X^n) \neq f(Y^n)).$$

This is the bound given in Theorem 2 in [1], where $cos(\theta) = C_{\mathbf{i}_1}$.[3]

Remark 6. *For fixed $\mathbf{P_i}$, the lower-bound is minimized by taking $\tilde{e}$, and $\tilde{f}$ to be the same functions.*

Corollary 2. *We can relax the bound in Theorem 1 as follows*:

$$2\sum_{\mathbf{i}}(1-C_{\mathbf{i}})\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \leq P(e(X^n) \neq f(Y^n)) \leq 1-2\sum_{\mathbf{i}}(1-C_{\mathbf{i}})\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}$$

Proof.

$$\begin{aligned} \sigma &\geq 2\sqrt{\sum_{\mathbf{i}}\mathbf{P_i}}\sqrt{\sum_{\mathbf{i}}\mathbf{Q_i}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\ &\overset{(a)}{\Rightarrow} \sigma \geq 2\sum_{\mathbf{i}}\mathbf{P}_i^{\frac{1}{2}}\mathbf{Q}_i^{\frac{1}{2}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\ &\Rightarrow \sigma \geq 2\sum_{\mathbf{i}}(1-C_{\mathbf{i}})\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}. \end{aligned}$$

In (a) we have used the Cauchy-Schwarz inequality. □

B. Arbitrary Input Alphabets

So far, we have only considered Boolean functions with binary input alphabets. Next, we extend Theorem 1; and derive a new bound on the correlation between the outputs of Boolean functions with arbitrary (finite) input alphabets. Similar to the previous part, let (X, Y) be a pair of DMS's with joint distribution $P_{X,Y}$. Assume that the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. Consider two arbitrary Boolean functions $e : \mathcal{X}^n \to \{0, 1\}$ and $f : \mathcal{Y}^n \to \{0, 1\}$. We prove the following extension of Theorem 1.

Theorem 2. *Let $\psi \triangleq \sup(E(h(X)g(Y))$, where the supremum is taken over all single-letter functions $h : \mathcal{X} \to \mathbb{R}$, and $g : \mathcal{Y} \to \mathbb{R}$ such that $h(X)$ and $g(Y)$ have unit variance and zero mean. the following bound holds*:

$$\begin{aligned} &2\sqrt{\sum_{\mathbf{i}}\mathbf{P_i}}\sqrt{\sum_{\mathbf{i}}\mathbf{Q_i}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \leq P(e(X^n) \neq f(Y^n)) \\ &\leq 1 - 2\sqrt{\sum_{\mathbf{i}}\mathbf{P_i}}\sqrt{\sum_{\mathbf{i}}\mathbf{Q_i}} + 2\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}, \end{aligned}$$

[3]For a definition of θ please refer to [1].

where 1) $C_{\mathbf{i}} \triangleq \psi^{N_{\mathbf{i}}}$, 2) $\mathbf{P_i}$ is the variance of $\tilde{e}_{\mathbf{i}}$, 3) $\underline{\tilde{e}}$ is the real function corresponding to $\underline{e}$, 4) $\mathbf{Q_i}$ is the variance of $\tilde{f}_{\mathbf{i}}$, and 5) $N_{\mathbf{i}} \triangleq w_H(\mathbf{i})$.

Proof. Pleas refer to the Appendix. □

Remark 7. *In Lemma 9 which was used in the proof of Theorem 1 in the Appendix, it was shown that for binary random variables X and Y, with $P(X \neq Y) = \epsilon$, we have $\psi = 1 - 2\epsilon$. So, the bounds in Theorem 1 and Theorem 2 are the same for binary inputs.*

Remark 8. *The value of ψ is in the interval $[0, 1]$. ψ is equal to one if and only if $X = Y$. The proof is straightforward and follows from the Cauchy-Schwarz inequality.*

C. Discontinuity of the Output Correlation at Asymptoticly Large Effective Lengths

In [11], it was shown that an extension of the Berger-Tung achievable region with common components for the distributed source coding problem is discontinuous in the source distribution. We argue that this is a widespread phenomenon in current coding strategies in multi-terminal communications and that it is an artifact of the discontinuity in the correlation between the outputs of functions with asymptotically large effective lengths.

Lemma 4. *Let (X^n, Y^n) be a sequence pairs of i.i.d. binary random variables and let $(\tilde{e}_{\mathbf{i}}^n)_{\mathbf{i}}$ and $(\tilde{f}_{\mathbf{i}}^n)_{\mathbf{i}\in\{0,1\}^n}$ be the real decomposition vector corresponding to the Boolean function $e^n(X^n)$ and $f^n(Y^n)$, respectively, where $(e^n(X^n), f^n(Y^n))_{n\in\mathbf{N}}$ is a sequence of pairs of Boolean functions such that:*

$$\frac{\mathbf{P}_{\mathbf{i}}^n}{\sqrt{Var(X)}} = \frac{\mathbf{Q}_{\mathbf{i}}^n}{\sqrt{Var(Y)}} = \begin{cases} 1 & \mathbf{i} = \underline{1}, \\ 0 & otherwise. \end{cases} \tag{9}$$

Then, if $\epsilon = P(X \neq Y) \neq 0$,

$$2\sqrt{\mathbf{P}_{\underline{1}}}\sqrt{\mathbf{Q}_{\underline{1}}} \leq \liminf_{n\to\infty} P(e^n(X^n) \neq f^n(Y^n)), \tag{10}$$

whereas for $\epsilon = 0$ and $e^n(\cdot) = f^n(\cdot)$, we have $P(e^n(X^n) \neq f^n(Y^n)) = 0$.

The proof of Equation (10) follows directly from Theorem 1. For $\epsilon = 0$, note that $P(X = Y) = 1$, so, given that the two Boolean function are the same, their outputs are equal with probability one.

Lemma 4 shows that functions with asymptotically large effective lengths produce outputs whose correlation is discontinuous as a function of the input distribution. In the next section, we show that single letter coding strategies produce functions with asymptotically large effective lengths. We use the discontinuity in correlation proved in Lemma 4 to show that these coding strategies are sub-optimal in various communication scenarios.

V. Correlation in Single Letter Coding Ensembles

In this section, we investigate the coding ensembles used in multi-terminal communication schemes. Deriving computable characterizations of the optimal achievable performance limits of communication networks has been a topic of significant

interest in multi-terminal information theory. One of the main instruments used in the information theoretic analysis of communication schemes is the concentration of measure properties which manifests when considering SLCE. These coding ensembles are used in schemes such as Shannon's point-to-point (PtP) source coding scheme, the Berger-Tung coding scheme for distributed source coding [25], the Zhang-Berger multiple-descriptions coding scheme [26], the Cover-El Gamal-Salehi coding scheme [6] for transmission of correlated sources over the multiple-access channel, and the Salehi-Kurtas scheme [27] for the transmission of sources over the interference channel. As a first step, it is shown that SLCEs produce encoding functions which have most of their variance either on the single-letter components or on the components with asymptotically large blocklengths. This along with Theorem 2 are used to prove that such schemes are inefficient in preserving correlation. In the next step, we provide several examples of multi-terminal scenarios where in order to achieve the optimal performance, the correlation between the outputs of the encoding functions used in different terminals must satisfy specific lower bounds which cannot be satisfied using SLCEs.

A. Single Letter Coding Ensembles

Traditionally, the probabilistic method has been used to investigate the performance of coding schemes in multi-terminal communications. The coding scheme is designed by providing a stochastic rule which chooses the encoding function from the set of all possible encoding functions. A 'good' coding scheme is one which produces 'good' encoding functions with high probability. A coding ensemble consists of a probability distribution on the set of all encoding functions:

Definition 7. *For a fixed $t \in \mathbb{N}$, let $(r_k^i)_{i\in\mathbb{N}}, k \in [1,t]$ be sequences of natural numbers which go to infinity as $i \to \infty$. Define sets of encoding functions $\mathcal{E}_k^i = \{\underline{e}_k^i : \mathcal{X}^i \to \{0,1\}^{r_k^i}\}, k \in [1,t], i \in \mathbb{N}$. A coding ensemble $\mathbb{S}$ is characterized by a sequence of probability measures $P_{\mathbb{S},i}(\underline{e}_1^i, \underline{e}_2^i, \cdots, \underline{e}_t^i), i \in \mathbb{N}$ on the set of encoding functions $\mathcal{E}_1^i \times \mathcal{E}_2^i \times \cdots \mathcal{E}_t^i$. The variable i is called the blocklength.*

Remark 9. *Whenever the choice of the coding ensemble and the blocklength is clear, we denote the distribution $P_{\mathcal{S},i}$ by $P_{\underline{E}_1,\underline{E}_1,\cdots,\underline{E}_t}(\underline{e}_1, \underline{e}_2, \cdots, \underline{e}_t)$.*

In a multi-terminal scenario with t encoders, for a given blocklength n, the coding ensemble chooses t encoding functions $\underline{E}_k^n, k \in [1,t]$ randomly and based on the joint distribution $P_{\underline{E}_1,\underline{E}_1,\cdots,\underline{E}_t}(\underline{e}_1, \underline{e}_2, \cdots, \underline{e}_t)$. Let X_k^n be the input of the kth encoder. The output of the encoder is the binary sequence $\underline{E}_k^n(X_k^n)$. The length of the output sequence is r_k^n. As an example, in Shannon's point-to-point coding ensemble, the probability distribution $P_{\underline{E}_1}(\underline{e}_1)$ on the set of encoding functions is determined by using single-letter distributions to assign probabilities to the corresponding codebooks. Shannon's method of assigning probabilities to encoding functions leads to specific properties which are shared among the coding ensembles used in many multi-terminal communication scenarios. These properties are described below. For simplicity of notation, in presenting these conditions we write $P_{\underline{E}}(\underline{e})$ instead of $P_{\underline{E}_k}(\underline{e}_k)$ when the notation does not cause ambiguity:

Definition 8. *The coding ensemble characterized by $P_{\mathbb{S},i}, i \in \mathbb{N}$ is called an SLCE if the following properties hold for every $k \in [1,t]$. Fix $k \in [1,t]$, let $\underline{E} = \underline{E}_k = (E_1, E_2, \cdots, E_n)$, then*[4]

*1) **Asymptotically Independent Codewords:** $\exists \delta_X > 0$ such that $\forall x^n, \exists \mathcal{B}_n(x^n) \subset \mathcal{X}^n$ such that the following holds:*

$$Pr(X^n \in \mathcal{B}_n(x^n)) \le 2^{-n\delta_X}, \quad \text{and}$$
$$\forall \tilde{x}^n \notin \mathcal{B}_n(x^n), e^r, \tilde{e}^r \in \{0,1\}^r :$$
$$(1 - 2^{-n\delta_X}) P_{\underline{E}(x^n)}(e^r) P_{\underline{E}(\tilde{x}^n)}(\tilde{e}^r) < P_{\underline{E}(x^n),\underline{E}(\tilde{x}^n)}(e^r, \tilde{e}^r)$$
$$< (1 + 2^{-n\delta_X}) P_{\underline{E}(x^n)}(e^r) P_{\underline{E}(\tilde{x}^n)}(\tilde{e}^r).$$

*2) **Asymptotically Independent Output Bits:** $\forall \delta > 0, \exists m \in \mathbb{N}$ such that $\forall n > m, \forall x^n \in \{0,1\}^n, \upsilon \in \{0,1\}, \forall i \in [1,n]$: $|P(E_i(X^n) = \upsilon | X^n = x^n) - P(E_i(X^n) = \upsilon | X_i = x_i)| < \delta$.*

*3) **Typicality Encoding:** $\forall \pi \in S_n : P_{\underline{E}}(\underline{E}) = P_{\underline{E}}(\underline{E}_\pi)$, where $\underline{E}_\pi(X^n) = \pi^{-1}(\underline{E}(\pi(X^n)))$, where S_n is the symmetric group of length n.*

The properties of SLCE codebooks can be explained as follows:

1) **Asymptotically Independent Codewords:** Take an arbitrary vector x^n. The condition requires that the codewords $E(x^n)$ and $E(\tilde{x})^n, \tilde{x}^n \in \{0,1\}^r$ be independently generated except for the set of vectors $\tilde{x}^n \in \mathcal{B}_n(x^n)$, where the probability of the set $\mathcal{B}_n(x^n)$ goes to 0 exponentially fast as $n \to \infty$.

An interpretation for this property is that codewords are chosen pairwise independently as the blocklength goes to infinity. For instance, let us investigate the property in the conventional Shannon code ensembles, where codewords are chosen pairwise independently. In order for $E(x^n)$ and $E(\tilde{x})^n$ to be correlated, they must be mapped to the same codeword. This requires that $\tilde{x}^n \in \mathcal{B}_n(x^n)$, where $\mathcal{B}_n(x^n)$ is the set of all vectors $\tilde{x}^n$ which are jointly typical with x^n with respect to the distribution $P_{X,\tilde{X}}$, where $P_{Y,X,\tilde{X}} = P_Y P_{X|Y} P_{\tilde{X}|Y}$.

2) **Asymptotically Independent Output Bits:** The property requires that the joint distribution of the input sequence and the output sequence of the encoding function averaged over all possible encoding functions approaches a product distribution in variational distance as $n \to \infty$ (i.e. the output bits 'look' independent.). It is well known that the property holds for conventional Shannon coding ensembles (e.g. [28]).

3) **Typicality Encoding:** The explanation for the third condition is that the probability that a vector x^n is mapped to y^n depends only on their joint type and is equal to the probability that the permuted sequence $\pi(x^n)$ is mapped to $\pi(y^n)$. As an example typicality encoding satisfies this condition.

B. Examples of Single-Letter Coding Ensembles

In the following examples, we show that the coding ensembles used in Shannon's point-to-point source coding

[4]Recall that the ith component of the vector of encoding functions $\underline{E}$ is denoted by E_i.

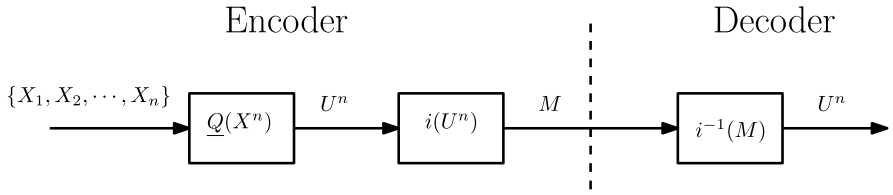

Fig. 2. Point-to-point source coding example

scheme [29] and the Cover-El Gamal-Salehi (CES) [6] scheme for the transmission of correlated sources over the multiple access channel are SLCEs.

1) Point-to-Point Source Coding: Consider the PtP source coding problem depicted in Figure 2. A discrete memoryless source X is fed to an encoder. The encoder uses the mapping $\underline{E} : \mathcal{X}^n \to \mathcal{U}^n$ to compress the source sequence. The codebook is defined as the image of $\underline{E}$. The codebook is indexed by the bijection $i : Im(\underline{E}) \to [1, |Im(\underline{E})|]$. The index $M \triangleq i(E(X^n))$ is sent to the decoder. The decoder reconstructs the compressed sequence $U^n \triangleq i^{-1}(M) = \underline{E}(X^n)$. The efficiency of the reconstruction is evaluated based on the separable distortion criteria $d_n : \mathcal{X}^n \times \mathcal{U}^n \to [0, \infty)$, where separability property means that $d_n(x^n, u^n) = \sum_{i\in[1,n]} d_1(x_i, u_i)$. We assume that the alphabets $\mathcal{X}$ and $\mathcal{U}$ are both binary. The rate of transmission is defined as $R \triangleq \frac{1}{n}\log|Im(\underline{E})|$, and the average distortion is defined as $\frac{1}{n}\mathbb{E}(d_n(X^n, U^n))$. The goal is to choose $\underline{E}$ such that the rate-distortion tradeoff is optimized. Note that the choice of the bijection 'i' does not affect the performance of the coding scheme. It is well-known that for a source X and distortion criteria $d_1 : \{0,1\} \times \{0,1\} \to [0, \infty)$, the rate-distortion pair $(R, D) = \big(r, \mathbb{E}_{X,U}(d_1(X, U))\big)$ is achievable for all $r > I(U; X)$ and conditional distributions $P_{U|X}$. The conventional proof [29] uses SLCE's to construct the coding scheme. In order to verify the properties of the SLCE's in the coding ensemble in [29], we provide an outline of the scheme. Fix $n \in \mathbb{N}$, and $\epsilon > 0$. Define $P_U(u) = \mathbb{E}_X\{P_{U|X}(u|X))\}$. In [29], a randomly generated encoding function is constructed with the aid of a set of vectors called the codebook, and typicality encoding. The codebook $\mathcal{C}$ is constructed as follows. Let $A_\epsilon^n(U) \triangleq \{u^n \big| |\frac{1}{n}w_H(u^n) - P_U(1)| < \epsilon\}$ be the set of n-length binary vectors which are ϵ-typical with respect to P_U. The codebook $\mathcal{C}$ is constructed by choosing $\lceil 2^{nR}\rceil$ vectors from $A_\epsilon^n(U)$ randomly and uniformly. For an arbitrary sequence $x^n \in \{0,1\}^n$, define $A_\epsilon^n(U|x^n)$ as the set of vectors in $\mathcal{C}$ which are jointly ϵ-typical with x^n with respect to $P_{U|X}$. The vector $\underline{E}(x^n)$ is chosen randomly and uniformly from $A_\epsilon^n(U|x^n) \cap \mathcal{C}$.

Remark 10. *The codebook generation process could be altered in the following way: instead of choosing the codewords randomly and uniformly from the set of typical sequences $A_\epsilon^n(U)$, the encoder can produce each codeword independent of the others and with the distribution $P_{U^n}(u^n) = \Pi_{i\in[1,n]}P_U(u_i)$. However, the discussion that follows remains unchanged regardless of which of these codebook generation methods are used.*

Lemma 5. *The ensemble described above is a SLCE.*

Proof. Please refer to the Appendix. □

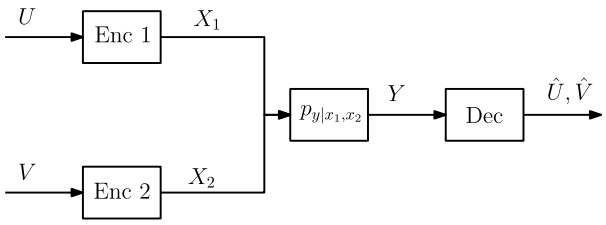

Fig. 3. Transmission of Sources over MAC

2) Transmission of Correlated Sources Over the Multiple Access Channel (CS-MAC): Consider the problem of the lossless transmission of the sources U and V over a MAC depicted in Figure 3. The largest known transmissible region for this problem is achieved using the CES scheme. The following lemma gives the transmissible region using the CES scheme in absence of common components (i.e. when there is no random variable W such that (a) $H(W) > 0$, (b) $W = f(U) = g(V)$.)

Lemma 6. *[6] The sources U, and V are transmissible over a CS-MAC with channel input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, and output alphabet $\mathcal{Y}$, and channel transition probability $p(y|x_1, x_2)$, if there exists a probability mass function $p(x_1|u)p(x_2|v)$ such that*:

$$H(U|V) < I(X_1; Y|X_2, V),$$
$$H(V|U) < I(X_2; Y|X_1, U),$$
$$H(U, V) < I(X_1, X_2; Y),$$

where $p(u, v, x_1, x_2) = p(u, v)p(x_1|u)p(x_2|v)$.

Similar to the previous example achievability is proved by providing a coding ensemble which specifies a probability distribution $P_{\mathbb{S}}(\mathbf{e}_1, \mathbf{e}_2)$ on the set of pairs of encoding functions at the two transmitters. The CES scheme generates the encoding functions independently (i.e. $P_{\mathbb{S}}(\mathbf{e}_1, \mathbf{e}_2) = P_{\mathbb{S}}(\mathbf{e}_1)P_{\mathbb{S}}(\mathbf{e}_2)$). Each of the encoding functions is generated by a method similar to the previous example. Hence, the marginals $P_{\mathbb{S}}(\mathbf{e}_i), i \in \{1, 2\}$ each satisfy the conditions in Definition 8. So, the coding ensemble is a SLCE.

C. Bounds on Output Correlation for SLCEs: The 1-∞ Law

Our objective is to analyze the correlation preserving properties of SLCE's. For a randomly generated encoding function $\underline{E} = (E_1, E_2, \cdots, E_n)$, denote the decomposition of the real function corresponding to the kth element into the form in Equation (3) as $\tilde{E}_k = \sum_{\mathbf{i}} \tilde{E}_{k,\mathbf{i}}, k \in [1, n]$. Let $\mathbf{P}_{k,\mathbf{i}}$ be the variance of $\tilde{E}_{k,\mathbf{i}}$. For a fixed $m \in \mathbb{N}$, we are interested in the quantity $\sum_{\mathbf{i}:N_{\mathbf{i}}\le m, \mathbf{i}\neq 00\cdots 01} \mathbf{P}_{k,\mathbf{i}}$ which is the total variance allocated to components of the decomposition

which operate on at most m elements of the input except for the single-letter component. From Theorem 1 we know that if $\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq 00\cdots 01}\mathbf{P}_{k,\mathbf{i}}$ is small, then the encoding function preserves less correlation. The following proposition shows that the probability $P_{\mathcal{S}}(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k}\mathbf{P}_{k,\mathbf{i}}\geq\gamma)$ is independent of the index k. This is due to property 3) in the Definition 8 of SLCE's.

Proposition 4. *$P_{\mathcal{S}}(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq 00\cdots 01}\mathbf{P}_{k,\mathbf{i}}\geq\gamma)$ is constant as a function of k.*

Proof. Please refer to the Appendix. □

The next theorem shows that most of the variance in the components of the decomposition of $\tilde{E}_k$ is concentrated in the single-letter component $\tilde{E}_{k,\mathbf{i}_k}$ and the large effective length components of the decomposition. We refer to this as the 1-∞ law. The proof of the theorem is provided in the Appendix.

Theorem 3. *For any $k\in\mathbb{N}, m\in\mathbb{N}, \gamma>0$, $P_{\mathcal{S}}(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k}\mathbf{P}_{k,\mathbf{i}}\geq\gamma)\to 0$, as $n\to\infty$, where, $\mathbf{i}_k$ is the kth standard basis element.*

Remark 11. *Theorem 3 shows that SLCE's distribute most of the variance of $\tilde{E}_k$ on $\tilde{E}_{k,\mathbf{i}}$'s which operate on large blocks. Hence, the encoders generated using such ensembles have high expected effective-lengths. This along with Theorem 1 gives an upper bound on the correlation preserving properties of SLCE's. This is stated in the following theorem.*

Theorem 4. *Let (X,Y) be a pair of DMS's, with $P(X=Y)=1-\epsilon$. Also, assume that the pair of BBE's $\underline{E},\underline{F}$ are produced using SLCE's. Define $E\triangleq E_1$, and $F\triangleq F_1$. Then,*

$$\forall\delta>0: P_{\mathcal{S}}\left(P_{X^n,Y^n}\left(E(X^n)\neq F(Y^n)\right)>\zeta\right)\to 1,$$

as $n\to\infty$, where $\zeta=2\mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}}-2(1-2\epsilon)\mathbf{P}_{\mathbf{i}_1}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}_1}^{\frac{1}{2}}-\delta$, $\mathbf{P}_{\mathbf{i}}\triangleq Var(\tilde{E}_{\mathbf{i}})$, $\mathbf{Q}_{\mathbf{i}}\triangleq Var(\tilde{F}_{\mathbf{i}})$, $\mathbf{P}\triangleq Var(\tilde{E})$, and $\mathbf{Q}\triangleq Var(\tilde{F})$.

The proof is provided in the Appendix.

Remark 12. *Note that in this theorem we consider a pair of BBEs produced using SLCEs. The bound is presented as function of the dependency spectra of the two BBEs. The two SLCEs can have arbitrary correlation. As an example, E and F can be taken to be either independent or exactly equal to each other.*

Remark 13. *The previous theorem gives a bound on the correlation preserving properties on SLCE's. The theorem shows that in order to increase correlation in these schemes the encoder needs to put more variance on the element $\tilde{E}_{k,\mathbf{i}_k}, k\in[1,n]$. This would require more correlation between the input and output of the encoder, which itself would require more rate. As an example consider the extreme case where $Var(\tilde{E}_k)=Var(\tilde{E}_{k,\mathbf{i}_k})$, which requires $E_k(X^n)=X_k$. This means that in order to achieve maximum correlation, the encoder must use uncoded transmission.*

Remark 14. *In the case when $X=Y$, there is common-information [8] available at the encoders. If the encoders use the same encoding function E, their outputs would be equal. Whereas from theorem 4, for any non-zero ϵ, the output correlation is bounded away from 0 (except when doing uncoded transmission). So, the correlation between the outputs of SLCE's is discontinuous as a function of ϵ.*

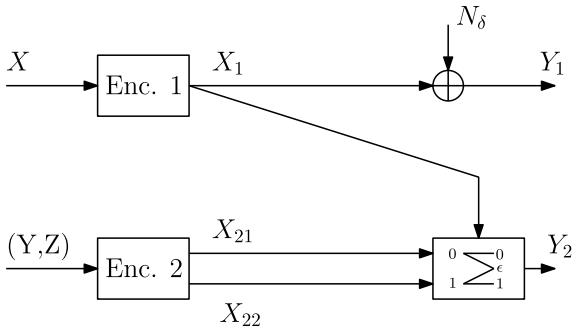

Fig. 4. An CS-IC example where SLCEs are suboptiomal.

VI. Multi-Terminal Communication Examples

In this section, we provide two examples of multi-terminal communication problems where SLCEs have suboptimal performance. We use the discontinuity mentioned in the previous section to show the sub-optimality of SLCEs.

A. Transmission of Correlated Sources Over the Interference Channel

Consider the problem of transmission of correlated sources over the interference channel (CS-IC) described in [27]. We examine the specific CS-IC setup shown in Figure 4. We are restricting our attention to bandwidth expansion factor equal to one. Here, the sources X and Y are Bernoulli random variables with parameters α_X and α_Y, and Z is a q-ary random variable with distribution P_Z. X and Z are independent. Y and Z are also independent. Finally, X and Y are correlated, and $P(X\neq Y)=\epsilon$. The random variable N_δ is Bernoulli with parameter δ. Decoder one reconstructs X and decoder two reconstructs Z losslessly. The first transmitter transmits the binary input X_1, and the second transmitter transmits the pair of inputs (X_{21},X_{22}), where X_{21} is q-ary and X_{22} is binary. Receiver 1 receives $Y_1=X_1\oplus_2 N_\delta$, and receiver 2 receives Y_2 which is given below:

$$Y_2=\begin{cases}X_{21}, & \text{if } X_{22}=X_1,\\ e, & otherwise.\end{cases}\tag{11}$$

So, the second channel outputs X_{21} noiselessly if the second encoder 'guesses' the first encoder's output correctly (i.e. $X_{22}=X_1$), otherwise an erasure is produced. The following proposition gives a set of sufficient conditions for the transmission of correlated sources over this interference channel:

Proposition 5. *The sources X and Z are transmissible if there exist $\epsilon,\gamma,d>0$, and $n\in\mathbb{N}$ such that:*

$$H(X)\leq(1-h_b(\delta))\left(1-\frac{h_b(\gamma+d)}{1-h_b(\delta)}\right)+O\left(\frac{1+\sqrt{nV+k\mathcal{V}(d)}Q^{-1}(\gamma)}{\sqrt{n}}\right),$$
$$H(Z)\leq\left((1-\epsilon)^k\log q\right)\left(1-\frac{h_b(\gamma+d)}{1-h_b(\delta)}\right),$$

where $h_b(\cdot)$ is the binary entropy function, $V = \delta(1-\delta)log_2(\frac{1-\delta}{\delta})$ is the channel dispersion, and $\mathcal{V}(d)$ is the rate-dispersion function as in [30], and $Q(.)$ is the Gaussian complementary cumulative distribution function.

For a fixed n, ϵ and γ, we denote the set of pairs $(H(X), H(Z))$ which satisfy the bounds by $S(n, \epsilon, \gamma)$.

Proof. First we provide an outline of the coding strategy. Fix $n, m \in \mathbb{N}, d, \gamma \in \mathbb{R}$, where $n \ll m$. Let $k = n\left(1 - \frac{h_b(d+\gamma)}{1-h_b(\delta)}\right)^{-1}$. The encoders send km bits of the compressed input at each block of transmission. The first encoder transmits its source in two steps. First, it uses a fixed blocklength source-channel code [30] with parameters (k, n, d, γ). The code maps k-length blocks of the source to n-length blocks of the channel input, and the average distortion resulting from the code is less than $d + \gamma$. In this step, the encoder transmits the source in m blocks of length k. A total of nm channel uses are needed (note that n<k). In the second step, the encoder uses a large blocklength code to correct the errors in the previous step. The code has rate close to $\frac{h_b(\gamma+d)}{1-h_b(\delta)}$, and its input length is equal to km.

The second encoder only transmits messages in the first step of transmission. It uses the same fixed blocklength code as the first encoder and the source sequence Y^k to estimate the outcome of the first encoder. It sends this estimate of the first encoder's output on X_{22}^n. Since $P(X^k = Y^k) = (1-\epsilon)^k$, we conclude that X_1^k and X_{22}^k are equal at least with probability $(1-\epsilon)^k$. The encoder sends the source Z using X_{21} over the resulting q-ary erasure channel which has probability of erasure at most $(1-\epsilon)^k$. The following provides a detailed descriptions of the coding strategy:

Codebook Generation: Fix n, ϵ, d. Let $k = n\left(1 - \frac{h_b(d+\gamma)}{1-h_b(\delta)}\right)^{-1}$. Let C_k be the optimal source-channel code with parameters (k, n, d, γ) for the point-to-point transmission of a binary source over the binary symmetric channel, as described in [30]. The code transmits k-length blocks of the source using n-length blocks of the channel input; and guarantees that the resulting distortion at each block is less than d with probability $(1-\epsilon)$ (i.e $P(d_H(X^n, \hat{X}^n) > d) \leq \gamma$, where $\hat{X}$ is the reconstruction of the binary source X at the decoder). In [30], it is shown that the parameters of the code satisfy:

$$n(1-h_b(\delta)) - k\,(H(X) - h_b(\alpha_X * d)) = \sqrt{nV + k\mathcal{V}(d)}\,Q^{-1}(\gamma) + O(log(n)).$$

Since $P(d_H(X^n, \hat{X}^n) > d) \leq \gamma$, it is straightforward to show that the average distortion is less than or equal to $\gamma + d$. Also, construct a family of good channel codes $C'_m, m \in \mathbb{N}$ for the binary symmetric channel with rate $R_m = 1 - h_b(\delta) - \lambda_m$, where $\lambda_m \to 0$ as $m \to \infty$. Next, construct a family of good channel codes $C''_m, m \in \mathbb{N}$ for the q-ary erasure channel with rate $R_m = (1-\epsilon)^k log(q) - \lambda_m$. Finally, randomly and uniformly bin the space of binary vectors of length kn with rate $R' = h_n(d+\gamma)$. More precisely, generate a binning function $B : \{0,1\}^{km} \to \{0,1\}^{kmR'}$, by mapping any vector $\mathbf{i}$ to a value chosen uniformly from $\{0,1\}^{kmR'}$.

Encoding: Fix m. At each block the encoders transmit km symbols of the source input. Let the source sequences be denoted by $X(1:k, 1:m), Y(1:k, 1:m), Z(1:k, 1:m)$, where we have broken the source vectors into m blocks of length k. In this notation $X(i, j)$ is the ith element of the jth block, and $X(1:k, j)$, $j \in [1, m]$ is the jth block.

Step 1: Encoder 1 uses the code C_k to transmit each of the blocks $X(1:k, i), i \in [1, m]$ to the decoder. The second encoder finds the output of the code C_k when $Y(1:k, i)$ is fed to the code, and transmits the output vector on $X_{22}(1:n, i)$. The encoder uses an interleaving method similar to the one in [12] to transmit Z. For the sequence $Z(1:k, 1:m)$, it finds the output of C''_{km} for this input and transmits it on $X_{21}(1:n, 1:m)$.

Step 2: The first encoder transmits $B(X(1:k, 1:m))$ to the decoder losslessly using $C'_{kmR'}$.

Decoding: In the first step, the first decoder reconstructs $X(1:k, 1:m)$ with average distortion at most $\gamma + d$. In the second step, using the bin number $B(X(1:k, 1:m))$ it can losslessly reconstruct the source, since $C'_{kmR'}$ is a good channel code. Decoder 2 also recovers $Z(1:k, 1:m)$ losslessly using $Y_2(1:k, 1:m)$ since C''_{km} is a good channel code.

The conditions for successful transmission is given as follows:

$$n(1-h_b(\delta)) - k\,(H(X) - h_b(\alpha_X * d)) \geq \sqrt{nV + k\mathcal{V}(d)}\,Q^{-1}(\gamma) + O(log(n)),$$
$$n(1-\epsilon)^k log(q) \geq kH(Z).$$

Simplifying these conditions by replacing $k = n\left(1 - \frac{h_b(d+\gamma)}{1-h_b(\delta)}\right)^{-1}$ proves the proposition. □

The bound provided in Proposition 5 is not calculable without the exact characterization of the $O(\frac{log(n)}{n})$ term. However, we use this bound to prove the sub-optimality of SLCEs. First, we argue that the transmissible region is 'continuous' as a function of ϵ. Note that for $\epsilon = 0$, sources with parameters $(H(X), H(Z)) = (1 - h_b(\delta), \log q)$ are transmissible. The region in Proposition 5 is continuous in the sense that as ϵ approaches 0, the pairs $(H(X), H(Z))$ in the neighborhood of $(1 - h_b(\delta), \log q)$ satisfy the bounds given in the proposition (i.e. the corresponding sources are transmissible).

Proposition 6. *For all $\lambda > 0$, there exist $\epsilon_0, \gamma_0 > 0$, and $n_0 \in \mathbb{N}$ such that*:

$$\forall \epsilon < \epsilon_0 : (1 - h_b(\delta) - \lambda, \log q - \lambda) \in S(n_0, \epsilon, \gamma_0).$$

Proof. Follows directly from Proposition 5. □

For an arbitrary encoding scheme operating on blocks of length n, let the encoding functions be as follows: $X_1^n = \underline{e}_1(X^n)$, and $(X_{21}^n, X_{22}^n) = (\underline{e}_{21}(Y^n, Z^n), \underline{e}_{22}(Y^n, Z^n))$. The following lemma gives an outer bound on $H(Z)$ as a function of the correlation between the outputs of $\underline{e}_1$ and $\underline{e}_{21}$.

Lemma 7. *For a coding scheme with encoding functions $\underline{e}_1(X^n), \underline{e}_{21}(Y^n, Z^n), \underline{e}_{22}(Y^n, Z^n)$, the following holds*:

$$H(Z) \leq \frac{1}{n}\sum_{i=1}^{n} P(e_{1,i}(X^n) = e_{22,i}(Y^n, Z^n)) + 1. \qquad (12)$$

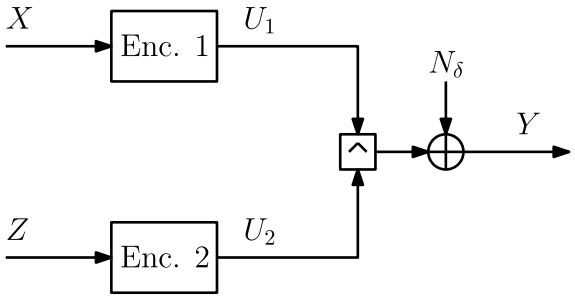

Fig. 5. A CS-MAC example where SLCEs are suboptiomal.

Proof. Since Z^n is reconstructed losslessly at the decoder, by Fano's inequality the following holds:

$$\begin{aligned} H(Z^n) &\approx I(Y_2^n; Z^n) \overset{(a)}{=} I(E^n, Y_2^n; Z^n) \\ &= I(E^n; Z^n) + I(Y_2^n; Z^n|E^n) \\ &\overset{(b)}{\leq} H(E^n) + \sum_{i=1}^{n} P(e_{1,i}(X^n) = e_{22,i}(Y^n, Z^n)) \log q \\ &\overset{(c)}{\leq} n + \sum_{i=1}^{n} P(e_{1,i}(X^n) = e_{22,i}(Y^n, Z^n)) \log q, \end{aligned}$$

where in (a) we have defined E^n as the indicator function of the event that $Y_2 = e$, in (b) we have used Equation (11) and in (c) we have used the fact that E^n is a binary vector. □

Using Theorem 4, we show that if the encoding functions are generated using SLCEs, $P(e_{1,i}(X^n) = e_{22,i}(Y^n, Z^n))$ is discontinuous in ϵ. The next proposition shows that SLCEs are sub-optimal:

Proposition 7. *There exists $\lambda > 0$, and $q \in \mathbb{N}$, such that sources with $(H(X), H(Z)) = (1 - h_b(\delta) - \lambda, \log q - \lambda)$ are not transmissible using SLCEs.*

Proof. Let $X_1^n = \underline{E}_1(X^n)$, and $X_{22^n} = \underline{E}_{22,z^n}(Y^n), z^n \in \{0,1\}^n$ be the encoding functions used in the two encoders to generate X_1 and X_{22}. If $H(Z) \approx \log(q)$, from (12), we must have $P(E_{1,j}(X^n) = E_{22,z^n,j}(Y^n)) \approx 1$ for almost all of the indices $j \in [1,n]$. From Theorem 4, this requires $\mathbf{P}_{j,\mathbf{i}_j} \approx 1$, which requires uncoded transmission (i.e. $X_1^n = \underline{E}(X^n) \approx X^n$). However, uncoded transmission contradicts the lossless reconstruction of the source at the first decoder. □

The proof is not restricted to any particular scheme, rather it shows that any SLCE would have sub-optimal performance.

B. Transmission of Correlated Sources Over the Multiple Access Channel (CS-MAC)

We examine the CS-MAC setup shown in Figure 5. Again, we restrict our attention to bandwidth expansion factor equal to one. Here, the source X is a q-ary source. The source Z is defined as $Z = X \oplus_q N_\epsilon$, where N_ϵ is a q-ary random variable with

$$P(N_\epsilon = i) = \begin{cases} 1-\epsilon, & \text{if } i = 0, \\ \frac{\epsilon}{q-1} & \text{if } i \in \{1, 2, \cdots, q-1\}, \end{cases}$$

and

$$P(N_\delta = i) = \begin{cases} 1-\delta, & \text{if } i = 0, \\ \frac{\delta}{q-1} & \text{if } i \in \{1, 2, \cdots, q-1\}, \end{cases}$$

The output is:

$$Y = U_1 \wedge U_2 \oplus_q N_\delta = \begin{cases} U_2 \oplus_q N_\delta, & \text{if } U_1 = U_2, \\ N_\delta & \text{if } U_1 \neq U_2, \end{cases}$$

where U_1 and U_2 are the outputs of Encoder 1 and Encoder 2, respectively. The goal is to transmit both sources X and Z losslessly to the decoder.

In this setup, there are two strategies available to the encoders. The first strategy is for both encoders to transmit the sources simultaneously. In this case, the encoders must have equal outputs. Otherwise, the decoder receives the noise N_δ. So, in this strategy, the encoders must 'guess' each other's outputs. The second strategy is to make a binary symmetric channel with noise δ for one of the encoders, while the other encoder does not transmit any messages. For example, in order to create such a channel for Encoder 1, encoder two transmits a constant sequence $U_2^n = (j, j, \cdots, j), j \in [1, q-1]$. Then, Encoder 2 can transmit a binary codeword using alphabet $\{0, j\}$. The rates of transmission for this strategy is:

$$\begin{aligned} R_{s,1} &= \max_{p(U_1)} I(U_1; Y) \\ &= h\left(\frac{1}{2}\left(1 - \frac{(q-2)\delta}{q-1}\right), \frac{1}{2}\left(1 - \frac{(q-2)\delta}{q-1}\right), \frac{\delta}{q-1}, \cdots, \frac{\delta}{q-1}\right) \\ &\quad - h\left(1-\delta, \frac{\delta}{q-1}, \cdots, \frac{\delta}{q-1}\right), \\ R_{s,2} &= 0. \end{aligned}$$

The following Proposition gives a condition under which the sources are transmissible:

Proposition 8. *There exists positive reals $\lambda_\epsilon, \epsilon \in (0, \frac{1}{2}]$, with $\lim_{\epsilon \to 0} \lambda_\epsilon = 0$, such that the sources X and Y are transmissible if the following condition is satisfied*:

$$H(X) \leq \log q - H(N_\delta) - \lambda_\epsilon.$$

Proof. The ideas in this proof are similar to the ones in Proposition 5. We provide an outline of the proof here. There are two steps for the transmission of the sources. First, the first strategy described above is used to transmit at a rate close to $\log q - H(N_\delta)$. In this step, the encoders use a finite blocklength code to maximize their probability of agreement. In the second step, the encoders use the second strategy described above to correct the errors from the first step. The errors in the first step vanish as $\epsilon \to 0$, since the sources become equal with probability going to one. So, the rate of transmission approaches the rate of the first step which is close to $\log q - H(N_\delta)$. We provide a more detailed summary of the proof: Fix n. Both encoders use an finite blocklength source-channel code for the q-ary symmetric channel with noise N_δ to transmit the sources. Let the blocklength of this code be equal to n, and the rate be equal to $\log q - H(N_\delta) + O(\frac{1}{\sqrt{n}})$. From the problem statement $P(X^n = Z^n) = P(N_\epsilon^n = 0^n) = (1-\epsilon)^n$. Since U_1^n is a function of X^n, a and U_2^n is a function of Z^n, we conclude

that $P(U_1^n = U_2^n) \geq (1-\epsilon)^n$. The encoders then take turns to send refinements to the decoder. This is done using the second strategy described above. The rate required for this part of the transmission is $\frac{(1-\epsilon)^n}{R_{s,1}} + \frac{H(N_\epsilon)}{R_{s,1}}$. Note that $\frac{(1-\epsilon)^n}{R_{s,1}} + \frac{H(N_\epsilon)}{R_{s,1}}$ goes to 0 as $\epsilon \to 0$. This completes the proof. □

The following lemma provides an upper-bound to the entropy of X as a function of ϵ and the correlation between U_1 and U_2.

Lemma 8. *For a coding scheme with encoding functions $U_1^n = \underline{e}_1(X^n), U_2^n = \underline{e}_2(Z^n)$, the following holds*:

$$H(X) \leq \frac{1}{n}\sum_{i=1}^{n} P(U_{1,i} = U_{2,i})\,(\log q - H(N_\delta)) + 1. \quad (13)$$

Proof. Similar to the proof of Lemma 7, we use Fano's inequality to prove the lemma. Since X^n is reconstructed losslessly at the decoder, by Fano's inequality the following holds:

$$\begin{aligned}
H(X^n) &\approx I(U_1^n U_2^n; Y^n)\\
&\leq \sum_{i=1}^{n} I(U_{1,i}U_{2,i}; Y_i) \overset{(a)}{=} \sum_{i=1}^{n} I(E_i, U_{1,i}U_{2,i}; Y_n)\\
&= \sum_{i=1}^{n} I(E_i; Y_i) + I(U_{1,i}U_{2,i}; Y_i|E_i)\\
&\overset{(b)}{\leq} H(E_i) + \sum_{i=1}^{n} P(U_{1,i} = U_{2,i})(Y^n, Z^n))\,(\log q - H(N_\delta))\\
&\overset{(c)}{\leq} n + \sum_{i=1}^{n} P(U_{1,i} = U_{2,i})\,(\log q - H(N_\delta)),
\end{aligned}$$

where in (a) we have defined E_i as the indicator function of the event that $U_{1,i} = U_{2_i}, i \in [1,n]$, in (b) we have used that $I(U_{1,i}U_{2,i}; Y_i|E_i) = P(E_i = 0)\cdot 0 + P(E_i = 1)I(U_{1,i}; Y|U_{1,i} = U_{2,i})$ and in (c) we have used the fact that E^n is binary.

□

Since for SLCE's $P(U_{1,i} = U_{2,i})$ is bounded away from 1 for $\epsilon \neq 0$, we conclude that there exists q and N_δ such that $\frac{1}{n}\sum_{i=1}^{n} P(U_{1,i} = U_{2,i})\,(\log q - H(N_\delta)) + 1 \leq \log q - H(N_\delta)$. So, SLCE's are suboptimal in this example as well.

VII. Conclusion

We derived a new bound on the maximum correlation between Boolean functions operating on pairs of sequences of random variable. The bound was presented as a function of the dependency spectrum of the functions. We developed a new mathematical apparatus for analyzing Boolean functions, provided formulas for decomposing the Boolean function into additive components, and for calculating the dependency spectrum of these functions. The new bound may find applications in security, control and information theory.

Next, we characterized a set of properties which are shared between the SLCEs used in the literature. We showed that ensembles which have these properties produce encoding functions which are inefficient in preserving correlation. We derived a probabilistic upper-bound on the correlation between the outputs of random encoders generated using SLCEs. We showed that the correlation between the outputs of such encoders is discontinuous with respect to the input distribution. We used this discontinuity to show that all SLCSs are sub-optimal in two specific multi-terminal communications problem involving the transmission of correlated source.

Appendix

A. Proof of Lemma 2

Proof. By definition, any element of $\mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$ satisfies the conditions in the proposition. Conversely, we show that any function satisfying the conditions (1) and (2) is in the tensor product. Let $\tilde{f} = \sum_{\mathbf{j}} \tilde{f}_{\mathbf{j}}$, $\tilde{f}_{\mathbf{j}} \in \mathcal{G}_{j_1} \otimes \mathcal{G}_{j_2} \otimes \cdots \otimes \mathcal{G}_{j_n}$ be an arbitrary function satisfying conditions (1) and (2). Assume $i_k = 1$ for some $k \in [1,n]$. Then:

$$\begin{aligned}
0 &\overset{(2)}{=} \mathbb{E}_{X^n|X_{\sim \mathbf{i}_k}}\Big(\sum_{\mathbf{j}} \tilde{f}_{\mathbf{j}}|X_{\sim \mathbf{i}_k}\Big) \overset{(a)}{=} \sum_{\mathbf{j}} \mathbb{E}_{X^n|X_{\sim \mathbf{i}_k}}(\tilde{f}_{\mathbf{j}}|X_{\sim \mathbf{i}_k})\\
&\overset{(1)}{=} \sum_{\mathbf{j}:j_k=0} \mathbb{E}_{X^n|X_{\sim \mathbf{i}_k}}(\tilde{f}_{\mathbf{j}}|X_{\sim \mathbf{i}_k}) \overset{(2)}{=} \sum_{\mathbf{j}:j_k=0} \tilde{f}_{\mathbf{j}},
\end{aligned}$$

where we have used linearity of expectation in (a), and the last two equalities use the fact that $\tilde{f}_{\mathbf{j}} \in \mathcal{G}_{j_1} \otimes \mathcal{G}_{j_2} \otimes \cdots \otimes \mathcal{G}_{j_n}$ which means it satisfies properties (1) and (2). So far we have shown that $\tilde{f} = \sum_{\mathbf{j} \geq \mathbf{i}} \tilde{f}_{\mathbf{j}}$. Recall that $\mathbf{i}$ is given in the statement of the proposition. Now assume $i_{k'} = 0$. Then:

$$\begin{aligned}
\sum_{\mathbf{j}\geq\mathbf{i}} \tilde{f}_{\mathbf{j}} = \tilde{f} &\overset{(1)}{=} \mathbb{E}_{X^n|X_{\sim \mathbf{i}_{k'}}}\Big(\sum_{\mathbf{j}\geq\mathbf{i}} \tilde{f}_{\mathbf{j}}|X_{\sim \mathbf{i}_{k'}}\Big)\\
&= \sum_{\mathbf{j}\geq\mathbf{i}} \mathbb{E}_{X^n|X_{\sim \mathbf{i}_{k'}}}(\tilde{f}_{\mathbf{j}}|X_{\sim \mathbf{i}_{k'}}) = \sum_{\mathbf{j}\geq\mathbf{i}:j_{k'}=0} \tilde{f}_{\mathbf{j}} \Rightarrow \sum_{\mathbf{j}\geq\mathbf{i}:j_{k'}=1} \tilde{f}_{\mathbf{j}} = 0.
\end{aligned}$$

So, $\tilde{f} = \sum_{\mathbf{i}\geq\mathbf{j}\geq\mathbf{i}} \tilde{f}_{\mathbf{j}} = \tilde{f}_{\mathbf{i}}$. By assumption we have $\tilde{f}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$. □

B. Proof of Lemma 3

Proof. 1) For two n-length binary vectors $\mathbf{i}$, and $\mathbf{j}$, we write $\mathbf{i} \leq \mathbf{j}$ if $i_k \leq j_k, \forall k \in [1,n]$. The set $\{0,1\}^n$ equipped with $\leq$ is a well-founded set (i.e. any subset of $\{0,1\}^n$ has at least one minimal element). The following presents the principle of Noetherian induction on well-founded sets:

Proposition 9 (Principle of Noetherian Induction [31])**.** *Let $(A, \preccurlyeq)$ be a well-founded set. To prove the property $P(x)$ is true for all elements x in A, it is sufficient to prove the following*

*1) **Induction Basis:** $P(x)$ is true for all minimal elements in A.*

*2) **Induction Step:** For any non-minimal element x in A, if $P(y)$ is true for all minimal y such that $y \prec x$, then it is true for x.*

We will use Noetherian induction to prove the result. Let $\mathbf{i}_j, j \in [1,n]$ be the jth element of the standard basis. Then $\tilde{e}_{\mathbf{i}_j} = \mathbb{E}_{X^n|X_j}(\tilde{e}|X_j)$. By the smoothing property of expectation, $\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}_j}) = \mathbb{E}_{X^n}(\tilde{e}) = 0$. Assume that $\forall \mathbf{j} < \mathbf{i}$,

$\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}) = 0$. Then,

$$\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}) = \mathbb{E}_{X^n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \tilde{e}_{\mathbf{j}}\right) = \mathbb{E}_{X^n}(\tilde{e}) - \sum_{\mathbf{j}<\mathbf{i}} \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}) = 0 - \sum_{\mathbf{j}<\mathbf{i}} 0 = 0.$$

2) This statement is also proved by induction. $\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})$ is a function of $X_{\mathbf{i}}$, so by induction $\tilde{e}_{\mathbf{i}} = \mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \tilde{e}_{\mathbf{k}}$ is also a function of $X_{\mathbf{i}}$.

3) Let $\mathbf{i}_k, k \in [1, n]$ be defined as the kth element of the standard basis, and take $j, j' \in [1, n]$, $j \neq j'$. We have:

$$\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}_j}\tilde{e}_{\mathbf{i}_{j'}}) = \mathbb{E}_{X^n}(\mathbb{E}_{X^n|X_j}(\tilde{e}|X_j)\mathbb{E}_{X^n|X_{j'}}(\tilde{e}|X_{j'})) \overset{(a)}{=} \mathbb{E}_{X^n}(\mathbb{E}_{X^n|X_j}(\tilde{e}|X_j))\mathbb{E}_{X^n}(\mathbb{E}_{X^n|X_{j'}}(\tilde{e}|X_{j'})) \overset{(b)}{=} \mathbb{E}^2_{X^n}(\tilde{e}) = 0,$$

where we have used the memoryless property of the source in (a) and (b) results from the smoothing property of expectation. We extend the argument by Noetherian induction. Fix $\mathbf{i}, \mathbf{k}$. Assume that $\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\tilde{e}_{\mathbf{j}'}) = \mathbb{1}(\mathbf{j} = \mathbf{j}')\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}), \forall \mathbf{j} < \mathbf{i}, \mathbf{j}' \leq \mathbf{k}$, and $\forall \mathbf{j} \leq \mathbf{i}, \mathbf{j}' \leq \mathbf{k}$. Then, we have

$$\begin{aligned}
&\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{k}}) \\
&= \mathbb{E}_{X^n}\left(\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \tilde{e}_{\mathbf{j}}\right)\left(\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}}) - \sum_{\mathbf{j}'<\mathbf{k}} \tilde{e}_{\mathbf{j}'}\right)\right) \\
&= \mathbb{E}_{X_n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) - \sum_{\mathbf{j}<\mathbf{i}} \mathbb{E}_{X^n}\left(\tilde{e}_{\mathbf{j}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) \\
&- \sum_{\mathbf{j}'<\mathbf{k}} \mathbb{E}_{X^n}\left(\tilde{e}_{\mathbf{j}'}\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) + \sum_{\mathbf{j}<\mathbf{i},\mathbf{j}'<\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\tilde{e}_{\mathbf{j}'}).
\end{aligned}$$

The second and third terms in the above expression can be simplified as follows. First, note that:

$$\tilde{e}_{\mathbf{i}} = \mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \tilde{e}_{\mathbf{j}} \Rightarrow \sum_{\mathbf{j}\leq\mathbf{i}} \tilde{e}_{\mathbf{j}} = \mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}). \quad (14)$$

Our goal is to simplify $\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\mathbb{E}_{X^n|X_{\mathbf{j}'}}(\tilde{e}|X_{\mathbf{j}'}))$. We proceed by considering two different cases:

Case 1: $\mathbf{i} \not\leq \mathbf{k}$ and $\mathbf{k} \not\leq \mathbf{i}$:

Let $\mathbf{j} < \mathbf{i}$:

$$\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})) \overset{(14)}{=} \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\sum_{\mathbf{l}\leq\mathbf{k}} \tilde{e}_{\mathbf{l}})) = \sum_{\mathbf{l}\leq\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}}\tilde{e}_{\mathbf{l}}) = \sum_{\mathbf{l}\leq\mathbf{k}} \mathbb{1}(\mathbf{j} = \mathbf{l})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) = \mathbb{1}(\mathbf{j} \leq \mathbf{k})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}).$$

By the same arguments, for $\mathbf{j}' \leq \mathbf{k}$:

$$\mathbb{E}_{X^n}\left(\tilde{e}_{\mathbf{j}'}\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) = \mathbb{1}(\mathbf{j}' \leq \mathbf{i})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}'}).$$

Replacing the terms in the original equality we get:

$$\begin{aligned}
\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{k}}) &= \mathbb{E}_{X^n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) \quad (15)\\
&- \sum_{\mathbf{j}<\mathbf{i}} \mathbb{1}(\mathbf{j} \leq \mathbf{k})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&- \sum_{\mathbf{j}'\leq\mathbf{k}} \mathbb{1}(\mathbf{j}' \leq \mathbf{i})\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}'} + \sum_{\mathbf{j}<\mathbf{i},\mathbf{j}'<\mathbf{k}} \mathbb{1}(\mathbf{j} = \mathbf{j}')\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}).
\end{aligned}$$

Note that:

$$\sum_{\mathbf{j}<\mathbf{i}} \mathbb{1}(\mathbf{j} \leq \mathbf{k})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) = \sum_{\mathbf{j}<\mathbf{i},\mathbf{j}\leq\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) = \sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}})$$
$$\sum_{\mathbf{j}'\leq\mathbf{k}} \mathbb{1}(\mathbf{j}' \leq \mathbf{i})\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}'} = \sum_{\mathbf{j}'\leq\mathbf{k},\mathbf{j}'\leq\mathbf{i}} \mathbb{E}_{X^n}(\tilde{e}_{\mathbf{j}'} = \sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}})$$
$$\sum_{\mathbf{j}<\mathbf{i},\mathbf{j}'<\mathbf{k}} \mathbb{1}(\mathbf{j} = \mathbf{j}')\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}} = \sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}})$$

Replacing the terms in (15), we have:

$$\begin{aligned}
\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{k}}) &= \mathbb{E}_{X^n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) - \sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&\overset{(a)}{=} \mathbb{E}_{X^n}(\mathbb{E}^2_{X^n|X_{\mathbf{i}\wedge\mathbf{k}}}(\tilde{e}(X^n)|X_{\mathbf{i}\wedge\mathbf{k}})) - \sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&\overset{(b)}{=} \mathbb{E}_{X^n}(\mathbb{E}^2_{X^n|X_{\mathbf{i}\wedge\mathbf{k}}}(\tilde{e}(X^n)|X_{\mathbf{i}\wedge\mathbf{k}})) - \mathbb{E}_{X^n}\left((\sum_{\mathbf{j}\leq\mathbf{i}\wedge\mathbf{k}} \tilde{e}_{\mathbf{j}})^2\right) \overset{(14)}{=} 0,
\end{aligned}$$

where in (b) we have used that $\tilde{e}_{\mathbf{i}}$'s are uncorrelated, and (a) is proved below:

$$\begin{aligned}
&\mathbb{E}_{X^n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) \\
&= \sum_{x_{\mathbf{i}\wedge\mathbf{k}}} P(x_{\mathbf{i}\wedge\mathbf{k}})\left(\left(\sum_{x_{|\mathbf{i}-\mathbf{k}|^+}} P(x_{|\mathbf{i}-\mathbf{k}|^+})\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) \times \right. \\
&\left.\left(\sum_{x_{|\mathbf{k}-\mathbf{i}|^+}} P(x_{|\mathbf{k}-\mathbf{i}|^+})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right)\right) \\
&= \sum_{x_{\mathbf{i}\wedge\mathbf{k}}} P(x_{\mathbf{i}\wedge\mathbf{k}})\mathbb{E}^2_{X^n|X_{\mathbf{i}\wedge\mathbf{k}}}(\tilde{e}|x_{\mathbf{i}\wedge\mathbf{k}}) \\
&= \mathbb{E}_{X^n}(\mathbb{E}^2_{X^n|X_{\mathbf{i}\wedge\mathbf{k}}}(\tilde{e}(X^n)|X_{\mathbf{i}\wedge\mathbf{k}})).
\end{aligned}$$

Case 2: Assume $\mathbf{i} \leq \mathbf{k}$:

$$\begin{aligned}
&\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\tilde{e}_{\mathbf{k}}) \\
&= \mathbb{E}_{X^n}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}})\right) - \sum_{\mathbf{j}<\mathbf{i}} \mathbb{1}(\mathbf{j} \leq \mathbf{k})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&- \sum_{\mathbf{j}'<\mathbf{k}} \mathbb{1}(\mathbf{j}' \leq \mathbf{i})\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}'}) + \sum_{\mathbf{j}<\mathbf{i},\mathbf{j}'<\mathbf{k}} \mathbb{1}(\mathbf{j} = \mathbf{j}')\mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&\overset{(a)}{=} \mathbb{E}_{X^n}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})) - \sum_{\mathbf{j}<\mathbf{i}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&- \sum_{\mathbf{j}'\leq\mathbf{i}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}'}) + \sum_{\mathbf{j}\leq\mathbf{i}} \mathbb{E}_{X^n}(\tilde{e}^2_{\mathbf{j}}) \\
&= 0,
\end{aligned}$$

where in (a) we have used (a) proved above.

Case 3: When $\mathbf{k} \leq \mathbf{i}$ the proof is similar to case 2.

4) Clearly when $|\mathbf{i}| = 1$, the claim holds. Assume it is true for all $\mathbf{j}$ such that $|\mathbf{j}| < |\mathbf{i}|$. Take $\mathbf{i} \in \{0, 1\}^n$ and $t \in [1, n], i_t = 1$

arbitrarily. We first prove the claim for $\mathbf{k} = \mathbf{i} - \mathbf{i}_t$:

$$\begin{aligned}
&\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{i}}|X_{\mathbf{k}}) = \mathbb{E}_{X^n|X_{\mathbf{k}}}\left(\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)|X_{\mathbf{k}}\right)\\
&= \mathbb{E}_{X^n|X_{\mathbf{k}}}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})|X_{\mathbf{k}}\right) - \sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}})\\
&\overset{(a)}{=} \mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}}) - \sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}})\\
&\overset{(b)}{=} \sum_{\mathbf{j}\le\mathbf{i}-\mathbf{i}_t}\tilde{e}_{\mathbf{j}} - \sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}})\\
&\overset{(c)}{=} \sum_{\mathbf{j}\le\mathbf{i}-\mathbf{i}_t}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}}) - \sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}})\\
&= \sum_{s\neq t}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{i}-\mathbf{i}_s}|X_{\mathbf{k}})\\
&\overset{(d)}{=} \sum_{s\neq t}\mathbb{E}_{X^n|X_{\mathbf{k}-\mathbf{i}_s}}(\tilde{e}_{\mathbf{i}-\mathbf{i}_s}|X_{\mathbf{k}-\mathbf{i}_s}) \overset{(e)}{=} 0,
\end{aligned}$$

where in (a) we have used $\mathbf{i} > \mathbf{k}$, (b) follows from equation (14), also (c) follows from $\mathbf{j} < \mathbf{k}$, (e) uses $\mathbf{k} \wedge (\mathbf{i} - \mathbf{i}_s) = \mathbf{k} - \mathbf{i}_s$, and finally, (d) uses the induction assumption. Now we extend the result to general $\mathbf{k} < \mathbf{i}$. Fix $\mathbf{k}$. Assume the claim is true for all $\mathbf{j}$ such that $\mathbf{k} < \mathbf{j} < \mathbf{i}$ (i.e $\forall \mathbf{k} < \mathbf{j} < \mathbf{i}$, $\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{X_{\mathbf{j}}|X_{\mathbf{k}}}) = 0$). We have:

$$\begin{aligned}
&\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{i}}|X_{\mathbf{k}})\\
&= \mathbb{E}_{X^n|X_{\mathbf{k}}}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}}\right)\\
&= \mathbb{E}_{X^n|X_{\mathbf{k}}}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})|X_{\mathbf{k}}\right) - \sum_{\mathbf{j}\le\mathbf{k}}\mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}_{\mathbf{j}}|X_{\mathbf{k}})\\
&= \mathbb{E}_{X^n|X_{\mathbf{k}}}(\tilde{e}|X_{\mathbf{k}}) - \sum_{\mathbf{j}\le\mathbf{k}}\tilde{e}_{\mathbf{j}} \overset{(14)}{=} 0.
\end{aligned}$$

□

C. Proof of Proposition 3

Proof.

$$\begin{aligned}
&\mathbf{P}_{\mathbf{i}} = Var_{X_{\mathbf{i}}}(\tilde{e}_{\mathbf{i}}(X^n)) = \mathbb{E}_{X_{\mathbf{i}}}(\tilde{e}^2_{\mathbf{i}}(X^n)) - \mathbb{E}^2_{X_{\mathbf{i}}}(\tilde{e}_{\mathbf{i}}(X^n))\\
&\overset{(a)}{=} \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right) - 0\\
&= \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\tilde{e}_{\mathbf{j}}\right)\\
&+ \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)\\
&\overset{(b)}{=} \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}_{X^n|X_{\mathbf{i}}}(\sum_{\mathbf{l}\le\mathbf{i}}\tilde{e}_{\mathbf{l}}|X_{\mathbf{i}})\tilde{e}_{\mathbf{j}}\right)\\
&+ \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)
\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{=} \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{i}}}\left(\sum_{\mathbf{l}\le\mathbf{i}}\mathbb{E}_{X^n|X_{\mathbf{i}}}(\tilde{e}_{\mathbf{l}}|X_{\mathbf{i}})\tilde{e}_{\mathbf{j}}\right)\\
&+ \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)\\
&\overset{(d)}{=} \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{i}}}\left(\sum_{\mathbf{l}<\mathbf{i}}\tilde{e}_{\mathbf{l}}\tilde{e}_{\mathbf{j}}\right)\\
&+ \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)\\
&\overset{(e)}{=} \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\sum_{\mathbf{l}<\mathbf{i}}\mathbb{1}(\mathbf{j}=\mathbf{l})\mathbb{E}_{X_{\mathbf{i}}}\left(\tilde{e}_{\mathbf{l}}\tilde{e}_{\mathbf{j}}\right)\\
&+ \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)\\
&= \mathbb{E}_{X_{\mathbf{i}}}\left(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})\right) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{j}}}(\tilde{e}^2_{\mathbf{j}}) + \mathbb{E}_{X_{\mathbf{i}}}\left(\left(\sum_{\mathbf{j}<\mathbf{i}}\tilde{e}_{\mathbf{j}}\right)^2\right)\\
&= \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{j}}}(\tilde{e}^2_{\mathbf{j}}) + \sum_{\mathbf{j}<\mathbf{i}}\sum_{\mathbf{k}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{i}}}(\tilde{e}_{\mathbf{j}}\tilde{e}_{\mathbf{k}})\\
&\overset{(f)}{=} \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})) - 2\sum_{\mathbf{j}<\mathbf{i}}\mathbb{E}_{X_{\mathbf{j}}}(\tilde{e}^2_{\mathbf{j}})\\
&+ \sum_{\mathbf{j}<\mathbf{i}}\sum_{\mathbf{k}<\mathbf{i}}\mathbb{1}(\mathbf{j}=\mathbf{k})\mathbb{E}_{X_{\mathbf{i}}}(\tilde{e}^2_{\mathbf{j}}) = \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{e}|X_{\mathbf{i}})) - \sum_{\mathbf{j}<\mathbf{i}}\mathbf{P}_{\mathbf{j}},
\end{aligned}$$

□

where (a) follows from condition 1) in Lemma 3, b) follows from the decomposition in Equation (14) in the appendix, (c) uses linearity of expectation, (d) holds from condition 2) in Lemma 3, and in (e) and (f) we have used condition 1) in Lemma 3.

D. Proof of Theorem 1

Proof. This proof builds upon the results in [1]. The proof involves three main steps. In the first two steps we prove the lower bound. First, we bound the Pearson correlation [32] between the real-valued functions $\tilde{e}$, and $\tilde{f}$. In the second step, we relate the correlation to the probability that the two functions are equal and derive the necessary bounds. Finally, in the third step we use the lower bound proved in the first two steps to derive the upper bound.

Step 1: Let $s \triangleq P_X(e(X^n) = 1)$, $r \triangleq P_Y(f(Y^n) = 1)$. From Remark 1, the expectation of both functions is 0. So, the Pearson correlation is given by

$$\frac{\mathbb{E}_{X^n,Y^n}(\tilde{e}\,\tilde{f})}{(rs(1-s)(1-r))^{\frac{1}{2}}}.$$

Our goal is to bound this value. We have:

$$\mathbb{E}_{X^n,Y^n}(\tilde{e}\,\tilde{f}) \overset{(a)}{=} \mathbb{E}_{X^n,Y^n}\left((\sum_{\mathbf{i}\in\{0,1\}^n}\tilde{e}_{\mathbf{i}})(\sum_{\mathbf{k}\in\{0,1\}^n}\tilde{f}_{\mathbf{k}})\right)$$

$$\overset{(b)}{=} \sum_{\mathbf{i}\in\{0,1\}^n} \sum_{\mathbf{k}\in\{0,1\}^n} \mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{k}}). \tag{16}$$

In (a) we have used Definition 5, and in (b) we use linearity of expectation. Using the fact that $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1} \otimes \mathcal{G}_{i_2} \otimes \cdots \otimes \mathcal{G}_{i_n}$ and Definition 4, we have:

$$\tilde{e}_{\mathbf{i}} = c_{\mathbf{i}} \prod_{t:i_t=1} \tilde{h}(X_t),\ \tilde{f}_{\mathbf{k}} = d_{\mathbf{k}} \prod_{t:k_t=1} \tilde{g}(Y_t). \tag{17}$$

where,

$$\tilde{h}(X) = \begin{cases} 1-q, & \text{if } X=1, \\ -q. & \text{if } X=0, \end{cases}, \qquad \tilde{g}(Y) = \begin{cases} 1-r, & \text{if } Y=1, \\ -r. & \text{if } Y=0, \end{cases} \tag{18}$$

We replace $\tilde{e}_{\mathbf{i}}$ and $\tilde{f}_{\mathbf{k}}$ in (16):

$$\mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{k}}) \overset{(17)}{=} \mathbb{E}_{X^n,Y^n}\left(\left(c_{\mathbf{i}} \prod_{t:i_t=1} \tilde{h}(X_t)\right)\left(d_{\mathbf{k}} \prod_{s:k_s=1} \tilde{g}(Y_s)\right)\right) \tag{19}$$

$$\begin{aligned}
&\overset{(a)}{=} c_{\mathbf{i}}d_{\mathbf{k}}\mathbb{E}_{X^n,Y^n}\left(\prod_{t:i_t=1} \tilde{h}(X_t) \prod_{s:k_s=1} \tilde{g}(Y_s)\right) \\
&\overset{(b)}{=} c_{\mathbf{i}}d_{\mathbf{k}}\mathbb{E}_{X^n,Y^n}\left(\prod_{t:i_t=1,k_t=1} \tilde{h}(X_t)\tilde{g}(Y_{k_t})\right)\mathbb{E}_{X^n}\left(\prod_{t:i_t=1,k_t=0} \tilde{h}(X_t)\right) \\
&\quad \times \mathbb{E}_{Y^n}\left(\prod_{t:i_t=0,k_t=1} \tilde{g}(Y_{k_t})\right) \\
&\overset{(c)}{=} \mathbb{1}(\mathbf{i}=\mathbf{k})c_{\mathbf{i}}d_{\mathbf{k}} \prod_{t:i_t=1} \mathbb{E}_{X^n,Y^n}\left(\tilde{h}(X_t)\tilde{g}(Y_t)\right) \\
&\overset{(d)}{\leq} \mathbb{1}(\mathbf{i}=\mathbf{k})c_{\mathbf{i}}d_{\mathbf{k}}(1-2\epsilon)^{N_{\mathbf{i}}} \prod_{t:i_t=1} \mathbb{E}^{\frac{1}{2}}_{X^n}\left(\tilde{e}^2(X_t)\right)\mathbb{E}^{\frac{1}{2}}_{Y^n}\left(\tilde{g}^2(Y)\right) \\
&\overset{(e)}{=} \mathbb{1}(\mathbf{i}=\mathbf{k})(1-2\epsilon)^{N_{\mathbf{i}}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} = \mathbb{1}(\mathbf{i}=\mathbf{k})C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}.
\end{aligned} \tag{20}$$

(a) follows from linearity of expectation. In (b) we have used the fact that in a pair of DMS's, X_i and Y_j are independent for $i \neq j$. (c) holds since from Lemma 3, $\mathbb{E}(\tilde{e}_i) = \mathbb{E}(\tilde{f}_i) = 0, \forall i \in [1,n]$. We prove (d) in Lemma 9 below. In (e) we have used proposition 1.

Lemma 9. *Let $g(X)$ and $h(Y)$ be two arbitrary zero-mean, real valued functions, then*:

$$\mathbb{E}_{X,Y}(g(X)h(Y)) \leq (1-2\epsilon)\mathbb{E}^{\frac{1}{2}}_X(g^2(X))\mathbb{E}^{\frac{1}{2}}_Y(h^2(Y)).$$

Proof. This is a well-known result [33]. A proof is provided here for completeness: Let the functions be given as follows:

$$g(X) = \begin{cases} \alpha & \text{if } X=0 \\ \beta & \text{if } X=1. \end{cases}, \qquad h(Y) = \begin{cases} \gamma & \text{if } Y=0 \\ \delta & \text{if } Y=1. \end{cases}$$

Also, let $P(X=1)=p$, and $P(Y=1)=q$. The zero-mean condition enforces the following equalities:

$$\alpha(1-p)+\beta p = 0 \Rightarrow \beta = \frac{-(1-p)\alpha}{p},$$
$$\gamma(1-q)+\delta q = 0 \Rightarrow \delta = \frac{-(1-q)\gamma}{q}.$$

Next, we calculate the joint distribution of P_{XY}. Let $P_{i,j} \triangleq P(X=i,Y=j), i,j \in \{0,1\}$. We have the following:

$$\begin{aligned}
P_{0,0}+P_{0,1} &= P(X=0) = 1-p, \\
P_{0,0}+P_{1,0} &= P(Y=0) = 1-q, \\
P_{0,0}+P_{1,1} &= P(X=Y) = 1-\epsilon, \\
P_{0,0}+P_{0,1}+P_{1,0}+P_{1,1} &= 1.
\end{aligned}$$

Solving the system of equations yields:

$$P_{0,0} = 1 - \frac{p+q+\epsilon}{2}, \qquad P_{0,1} = \frac{q+\epsilon-p}{2}, \tag{21}$$
$$P_{1,0} = \frac{p+\epsilon-q}{2}, \qquad P_{1,1} = \frac{p+q-\epsilon}{2}. \tag{22}$$

With the following constraint on the variables:

$$\begin{aligned}
p+\epsilon &\geq q, \qquad p+q \geq \epsilon, \\
q+\epsilon &\geq p, \qquad p+q+\epsilon \leq 2.
\end{aligned}$$

We have:

$$\frac{\mathbb{E}_{X,Y}(gh)}{\mathbb{E}^{\frac{1}{2}}_X(g^2)\mathbb{E}^{\frac{1}{2}}_Y(h^2)} \tag{23}$$

$$\begin{aligned}
&= \frac{\alpha\gamma\left(P_{0,0} - P_{0,1}\frac{(1-q)}{q} - P_{1,0}\frac{(1-p)}{p} + P_{1,1}\frac{(1-q)(1-p)}{pq}\right)}{\alpha\gamma\left(\left((1-p)+\frac{(1-p)^2}{p}\right)^{\frac{1}{2}}\left((1-q)+\frac{(1-q)^2}{q}\right)^{\frac{1}{2}}\right)} \\
&= \frac{P_{0,0} - P_{0,1}\frac{(1-q)}{q} - P_{1,0}\frac{(1-p)}{p} + P_{1,1}\frac{(1-q)(1-p)}{pq}}{(\frac{1-p}{p})^{\frac{1}{2}}(\frac{1-q}{q})^{\frac{1}{2}}} \\
&= \frac{P_{0,0}pq - P_{0,1}(1-q)p}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&\quad - \frac{P_{1,0}(1-p)q - P_{1,1}(1-q)(1-p)}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&\overset{(22)}{=} \frac{(1-\frac{p+q+\epsilon}{2})pq - (\frac{q+\epsilon-p}{2})(1-q)p}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&\quad + \frac{-(\frac{p+\epsilon-q}{2})(1-p)q + (\frac{p+q-\epsilon}{2})(1-q)(1-p)}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&= \frac{pq + (\frac{p+q}{2})\left((1-p)(1-p) - pq\right)}{(pq(1-p)(1-q))^{\frac{1}{2}}} + \\
&\quad + \frac{(\frac{q-p}{2})\left(q(1-p) - p(1-q)\right))}{(pq(1-p)(1-q))^{\frac{1}{2}}} + \\
&\quad \frac{\frac{\epsilon}{2}(pq + p(1-q) + q(1-p) + (1-p)(1-q))}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&= \frac{pq + \frac{p+q}{2}(1-p-q) - \frac{p-q}{2}(q-p) - \frac{\epsilon}{2}}{(pq(1-p)(1-q))^{\frac{1}{2}}} \\
&= \frac{p+q-2pq-\epsilon}{2\left(pq(1-p)(1-q)\right)^{\frac{1}{2}}}.
\end{aligned} \tag{24}$$

We calculate the optimum point by taking partial derivatives:

$$\begin{aligned}
&\frac{\delta}{\delta p}\frac{\mathbb{E}_{X,Y}(gh)}{\mathbb{E}_X^{\frac{1}{2}}(g^2)\mathbb{E}_Y^{\frac{1}{2}}(h^2)} = 0 \Rightarrow \\
&2(1-2q)(pq(1-p)(1-q))^{\frac{1}{2}} \\
&-\frac{(1-2p)}{\sqrt{p(1-p)}}\sqrt{q(1-q)}(p+q-2pq-\epsilon) = 0 \\
&\overset{(a)}{\Rightarrow} 2(1-2q)p(1-p) - (1-2p)(p+q-2pq-\epsilon) = 0 \\
&\Rightarrow 2p(1-p)(1-2q) - p(1-2p)(1-2q) \\
&-(1-2p)q + (1-2p)\epsilon = 0 \\
&\Rightarrow p(1-2q) - (1-2p)q + (1-2p)\epsilon = 0 \\
&\Rightarrow p - q + (1-2p)\epsilon = 0. \qquad (25)
\end{aligned}$$

Where in (a) we have used $p, q \notin \{0,1\}$ to multiply by $\sqrt{pq(1-p)(1-q)}$. Taking the partial derivative with respect to q, by similar calculations we get:

$$\frac{\delta}{\delta q}\frac{\mathbb{E}_{X,Y}(gh)}{\mathbb{E}_X^{\frac{1}{2}}(g^2)\mathbb{E}_Y^{\frac{1}{2}}(h^2)} = 0 \rightarrow q - p + (1-2q)\epsilon. \qquad (26)$$

In order for (25) and (26) to be satisfied simultaneously, we must have $\epsilon = 0$, $p = q$, or $\epsilon = p+q = 1$, or $p = q = \frac{1}{2}$. For $\epsilon \notin \{0,1\}$, we must have $p = q = \frac{1}{2}$ in which case the value in (24) is:

$$\frac{\mathbb{E}_{X,Y}(gh)}{\mathbb{E}_X^{\frac{1}{2}}(g^2)\mathbb{E}_Y^{\frac{1}{2}}(h^2)} = 1 - 2\epsilon.$$

This completes the proof of the Lemma. □

Using equations (16) and (20) we get:

$$\mathbb{E}_X(\tilde{e}\tilde{f}) \leq \sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}.$$

Step 2: We use the results from step one to derive a bound on $P(e \neq f)$. Define $a \triangleq P(e(X^n) = 1, f(Y^n) = 1)$, $b \triangleq P(e(X^n) = 0, f(Y^n) = 1)$, $c \triangleq P(e(X^n) = 1, f(Y^n) = 0)$, and $d \triangleq P(e(X^n) = 0, f(Y^n) = 0)$, then

$$\begin{aligned}
&\mathbb{E}_{X^n,Y^N}(\tilde{e}(X^n)\tilde{f}(Y^n)) \\
&\quad = a(1-s)(1-r) - bs(1-r) - c(1-s)r + dsr, \qquad (27)
\end{aligned}$$

We write this equation in terms of $\sigma \triangleq P(f \neq g)$, s, and r using the following relations:

1) $a + c = s$, 2) $b + d = 1 - s$,
3) $a + b = r$, 4) $c + d = 1 - r$, 5) $b + c = \sigma$.

Solving the above we get:

$$a = \frac{s + r - \sigma}{2}, \qquad b = \frac{r + \sigma - s}{2}, \qquad (28)$$

$$c = \frac{s - r + \sigma}{2}, \qquad d = 1 - \frac{s + r + \sigma}{2}. \qquad (29)$$

We replace a, b, c, and d in (27) by their values in (29):

$$\begin{aligned}
&\frac{\sigma}{2} \geq (\frac{s+r}{2})(1-s)(1-r) + (\frac{s-r}{2})s(1-r) \\
&+ (\frac{r-s}{2})(1-s)r + sr(1-\frac{s+r}{2}) - \sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\
&\Rightarrow \sigma \geq s + r - 2rs - 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\
&\Rightarrow \sigma \geq (\sqrt{s(1-r)} - \sqrt{r(1-s)})^2 + 2\sqrt{s(1-s)r(1-r)} \\
&- 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\
&\Rightarrow \sigma \geq 2\sqrt{s(1-s)r(1-r)} - 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}
\end{aligned}$$

On the other hand $\mathbb{E}_X(\tilde{e}^2) = s(1-s) = \sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}$, where the last equality follows from the fact that $\tilde{e}_{\mathbf{i}}$'s are uncorrelated. This proves the lower bound. Next we use the lower bound to derive the upper bound.

Step 3: The upper-bound can be derived by considering the function $h(Y^n)$ to be the complement of $f(Y^n)$ (i.e. $h(Y^n) \triangleq 1 \oplus_2 f(Y^n)$.) In this case $P(h(Y^n) = 1) = P(f(Y^n) = 0) = 1 - r$. The corresponding real function for $h(Y^n)$ is:

$$\begin{aligned}
\tilde{h}(Y^n) &= \begin{cases} r & \text{if } h(Y^n) = 1, \\ -(1-r) & \text{if } h(Y^n) = 0, \end{cases} \\
&= \begin{cases} r & \text{if } f(Y^n) = 0, \\ -(1-r) & \text{if } f(Y^n) = 1, \end{cases} \Rightarrow \tilde{h}(Y^n) = -\tilde{f}(Y^n).
\end{aligned}$$

So, $\tilde{h}(Y^n) = -\sum_{\mathbf{i}} \tilde{f}_{\mathbf{i}}$. Using the same method as in the previous step, we have:

$$\begin{aligned}
&\mathbb{E}_{X^n,Y^n}(\tilde{e}\tilde{h}) = -\mathbb{E}_{X^n,Y^n}(\tilde{e}\tilde{f}) \leq \sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\
&\Rightarrow P(e(X^n) \neq h(Y^n)) \geq 2\sqrt{\sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}}\sqrt{\sum_{\mathbf{i}} \mathbf{Q}_{\mathbf{i}}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}
\end{aligned}$$

On the other hand $P(e(X^n) \neq h(Y^n)) = P(e(X^n) \neq 1 \oplus f(Y^n)) = P(e(X^n) = f(Y^n)) = 1 - P(e(X^n) \neq f(Y^n))$. So,

$$\begin{aligned}
&1 - P(e(X^n) \neq f(Y^n)) \geq 2\sqrt{\sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}}\sqrt{\sum_{\mathbf{i}} \mathbf{Q}_{\mathbf{i}}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \\
&\Rightarrow P(e(X^n) \neq f(Y^n)) \leq \\
&1 - 2\sqrt{\sum_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}}\sqrt{\sum_{\mathbf{i}} \mathbf{Q}_{\mathbf{i}}} + 2\sum_{\mathbf{i}} C_{\mathbf{i}} \mathbf{P}_{\mathbf{i}}^{\frac{1}{2}} \mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}}.
\end{aligned}$$

This completes the proof. □

E. Proof of Theorem 2

Proof. The proof of Theorem 2 follows similar steps as the proof of Theorem 1. The only difference is in the proof of step 1.

Step 1: Let $q \triangleq P_X(e(X^n)=1)$, $r \triangleq P_Y(f(Y^n)=1)$. We have:

$$\mathbb{E}_{X^n,Y^n}(\tilde{e}\tilde{f}) \overset{(a)}{=} \mathbb{E}_{X^n,Y^n}\left(\left(\sum_{\mathbf{i}\in\{0,1\}^n}\tilde{e}_{\mathbf{i}}\right)\left(\sum_{\mathbf{k}\in\{0,1\}^n}\tilde{f}_{\mathbf{k}}\right)\right) \overset{(b)}{=} \sum_{\mathbf{i}\in\{0,1\}^n}\sum_{\mathbf{k}\in\{0,1\}^n}\mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{k}}). \tag{30}$$

In (a) we have used Definition 5, and in (b) we use linearity of expectation. Using the fact that $\tilde{e}_{\mathbf{i}} \in \mathcal{G}_{i_1}\otimes\mathcal{G}_{i_2}\otimes\cdots\otimes\mathcal{G}_{i_n}$ and Lemma 1, we have:

$$\tilde{e}_{\mathbf{i}}(X^n) = \sum_{\forall t\in\tau: l_t\in[1,|\mathcal{X}|-1]} c_{\mathbf{i},(l_t)_{t\in\tau}}\prod_{t\in\tau}\tilde{h}_{l_t}(X_t),$$
$$\tilde{f}_{\mathbf{i}}(Y^n) = \sum_{\forall t\in\tau: l_t\in[1,|\mathcal{Y}|-1]} d_{\mathbf{i},(l_t)_{t\in\tau}}\prod_{t\in\tau}\tilde{g}_{l_t}(Y_t), \tag{31}$$

where $c_{\mathbf{i},(\mathbf{l_t})_{\mathbf{t}\in\tau}} \in \mathbb{R}$, and $\tilde{h}_l(X), l\in\{1,2,\cdots,|\mathcal{X}|-1\}$, and $\tilde{g}_l(Y), l\in\{1,2,\cdots,|\mathcal{Y}|-1\}$ are a basis for $\mathcal{I}_{X,1}$ and $\mathcal{I}_{Y,1}$, respectively. We have:

$$\begin{aligned}\mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{k}}) &\overset{(a)}{=} \mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{i}})\mathbb{1}(\mathbf{i}=\mathbf{k})\\ &\overset{(b)}{=} \mathbb{1}(\mathbf{i}=\mathbf{k})\mathbb{E}_{X^n}(\tilde{e}_{\mathbf{i}}\mathbb{E}_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n))\\ &\overset{(c)}{\leq} \mathbb{1}(\mathbf{i}=\mathbf{k})\mathbb{E}^{\frac{1}{2}}_{X^n}(\tilde{e}^2_{\mathbf{i}})\mathbb{E}^{\frac{1}{2}}_{X^n}(\mathbb{E}^2_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n))\\ &= \mathbb{1}(\mathbf{i}=\mathbf{k})\mathbf{P}^{\frac{1}{2}}_{\mathbf{i}}\mathbb{E}^{\frac{1}{2}}_{X^n}(\mathbb{E}^2_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n)),\end{aligned} \tag{32}$$

(a) follows by the same arguments as the ones in step 1 of the proof of Theorem 1, (b) follows from the law of total expectation and the fact that $e_{\mathbf{i}}$ is a function of X^n. In (c) we have used the Cauchy-Schwarz inequality. It only remains to find bounds on $\mathbb{E}_{X^n}(\mathbb{E}^2_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n))$ which are functions of $\mathbf{Q_i}$, ψ, and $N_{\mathbf{i}}$. Let $(i_1,i_2,\cdots,i_{N_{\mathbf{i}}})$ be the indices for which the elements of $\mathbf{i}$ are equal to one. Note that:

$$\begin{aligned}&\mathbb{E}_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n) = \mathbb{E}_{Y_{\mathbf{i}}|X_{\mathbf{i}}}(\tilde{f}_{\mathbf{i}}|X_{\mathbf{i}})\\ &= \mathbb{E}_{Y_{i_{N_{\mathbf{i}}}}|X_{i_{N_{\mathbf{i}}}}}\left(\mathbb{E}_{Y_{\mathbf{i}-\mathbf{i}_{N_{\mathbf{i}}}}|X_{\mathbf{i}-\mathbf{i}_{N_{\mathbf{i}}}}}\left(\tilde{f}_{\mathbf{i}}|X_{\mathbf{i}-\mathbf{i}_{N_{\mathbf{i}}}}\right)|X_{i_{N_{\mathbf{i}}}}\right)\\ &= \mathbb{E}_{Y_{i_{N_{\mathbf{i}}}}|X_{i_{N_{\mathbf{i}}}}}\left(\mathbb{E}_{Y_{i_{N_{\mathbf{i}}-1}}|X_{i_{N_{\mathbf{i}}-1}}}\left(\cdots\left(\mathbb{E}_{Y_{i_1}|X_{i_1}}\left(\tilde{f}_{\mathbf{i}}|X_{i_1}\right)|X_{i_2}\right)\cdots\right)|X_{i_{N_{\mathbf{i}}}}\right),\end{aligned} \tag{33}$$

where the first equality follows from the fact that $f_{\mathbf{i}}$ is a function of $Y_{\mathbf{i}}$. The rest of the equalities follow from the discrete and memoryless properties of the input. For ease of notation define the following projection operators for $1\leq i\leq n$:

$$\Pi_{X_i}: \mathcal{I}_{Y,i}\to\mathcal{I}_{X,i},\quad h(Y_i)\mapsto \mathbb{E}_{Y_i|X_i}(h(Y_i)).$$

Π_{X_i} can be interpreted as the projector of zero-mean functions of the random variable Y_i onto zero-mean functions of the random variable X_i. We can rewrite Equation (33) as follows:

$$\mathbb{E}_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n) = \Pi_{X_{i_{N_{\mathbf{i}}}}}\circ\Pi_{X_{i_{N_{\mathbf{i}}-1}}}\circ\cdots\circ\Pi_{X_{i_1}}(f_{\mathbf{i}}). \tag{34}$$

We find bounds on $\mathbb{E}_{X^n}(\mathbb{E}^2_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n))$ as follows:

$$\begin{aligned}&\mathbb{E}_{X^n}(\mathbb{E}^2_{Y^n|X^n}(\tilde{f}_{\mathbf{i}}|X^n))\\ &= \mathbb{E}_{X^n}\left(\left(\Pi_{X_{i_{N_{\mathbf{i}}}}}\circ\Pi_{X_{i_{N_{\mathbf{i}}-1}}}\circ\cdots\circ\Pi_{X_{i_1}}(f_{\mathbf{i}})\right)^2\right)\\ &\overset{(a)}{\leq} \mathbf{Q_i}||\Pi_{X_{i_{N_{\mathbf{i}}}}}\circ\Pi_{X_{i_{N_{\mathbf{i}}-1}}}\circ\cdots\circ\Pi_{X_{i_1}}||\\ &\overset{(b)}{=} \mathbf{Q_i}||\Pi_{X_{i_{N_{\mathbf{i}}}}}||\cdot||\Pi_{X_{i_{N_{\mathbf{i}}-1}}}||\cdots||\Pi_{X_{i_1}}||\\ &\overset{(c)}{=} \mathbf{Q_i}||\Pi_{X_1}||^n,\end{aligned} \tag{35}$$

where in (a) the operation norm is defined as $||\Pi|| = \sup_e \mathbb{E}(\Pi^2(e))$ where the supremum is taken over all zero-mean functions e with unit variance. (b) follows from the discrete memoryless property of the inputs. Finally, (c) holds since the source elements are identically distributed. On the other hand, we have:

$$\begin{aligned}\psi &= sup_{h,g\in\mathcal{L}}\mathbb{E}_{X_1,Y_1}(h(X_1)g(Y_1))\\ &= sup_{h,g\in\mathcal{L}}\mathbb{E}_{X_1}(h(X_1)\mathbb{E}_{Y_1|X_1}(g(Y_1)|X_1))\\ &\overset{(a)}{=} sup_{g\in\mathcal{L}}\mathbb{E}^{\frac{1}{2}}_{X_1}(h^2(X_1)\mathbb{E}^{\frac{1}{2}}_{X_1}(\mathbb{E}^2_{Y_1|X_1}(g(Y_1)|X_1))\\ &\overset{(b)}{=} sup_{g\in\mathcal{L}}\mathbb{E}^{\frac{1}{2}}_{X_1}(\mathbb{E}^2_{Y_1|X_1}(g(Y_1)|X_1))\\ &\overset{(c)}{=} ||\Pi_{X_1}||,\end{aligned} \tag{36}$$

where $\mathcal{L}$ is the set of all pairs of functions $g(X)$ and $h(Y)$ with zero mean which have unit variance. (a) follows from the Cauchy-Schwarz inequality and the fact that equality is satisfied by taking $g(X_1) = c\mathbb{E}_{Y_1|X_1}(h(Y_1)|X_1)$ where the constant c is chosen properly, so that $g(X_1)$ has unit variance. The quality (b) holds since $h(X_1)$ has unit variance, and (c) holds by the definition of operator norm. Combining equations (32), (35), (36) we have:

$$\mathbb{E}_{X^n,Y^n}(\tilde{e}_{\mathbf{i}}\tilde{f}_{\mathbf{k}}) \leq \mathbb{1}(\mathbf{i}=\mathbf{k})\psi^{N_{\mathbf{i}}}\mathbf{P}^{\frac{1}{2}}_{\mathbf{i}}\mathbf{Q}^{\frac{1}{2}}_{\mathbf{i}}.$$

The rest of the proof follows by the exact same arguments as in steps 2 and 3 in the proof of Theorem 1. □

F. Proof of Lemma 5

Proof. We provide an outline of the proof that the three properties in Definition 8 are satisfied:

1) As a reminder, the set $B_n(X^n)$ is the set of sequences $\tilde{x}^n$ which may be mapped to the same output sequence u^n as the output of x^n. In this coding scheme $B_n(x^n)$ is as follows:

$$B_n(x^n) = \{\tilde{x}^n | \exists u^n : (x^n,u^n), (\tilde{x}^n,u^n)\in A^n_\epsilon(X,U)\}.$$

Following the notation in [3], let $\mathcal{V}(x^n)$ be the set of all conditional types of sequences $\tilde{x}^n$ given x^n, and let $T_\upsilon(x^n)$ be the set of all sequences $\tilde{x}^n$ which have the conditional type $\upsilon\in\mathcal{V}(x^n)$ with respect to the sequence x^n. Then:

$$|B_n(x^n)| = \sum_{\upsilon\in\mathcal{V}(x^n)}|B(x^n)\bigcap T_\upsilon(x^n)|.$$

Note that $|B(x^n)\bigcap T_\upsilon(x^n)|\neq 0$ if and only if there exists a joint conditional type $\tilde{\upsilon}_{U,\tilde{X}|x^n}$ such that $|\tilde{P}_{U,X}-P_{U,X}|<\epsilon$ and $|\tilde{P}_{U,\tilde{X}}-P_{UX}|<\epsilon$ and $\tilde{P}_{U|X}=\upsilon$, where $\tilde{P}_{U,X,\tilde{X}}$ is the

joint type of the sequences in $\tilde{v}_{U,\tilde{X}|x^n}$ with x^n. As a result we have:

$$|B_n(x^n)| = \sum_{\substack{v\in\mathcal{V}(x^n)\\ \exists \tilde{v}_{U,\tilde{X}|x^n}:\\ |\tilde{P}_{U,X}-P_{U,X}|<\epsilon,|\tilde{P}_{U,\tilde{X}}-P_{UX}|<\epsilon}} |B(x^n)\bigcap T_v(x^n)|.$$

By standard type analysis arguments we conclude that:

$$|B_n(x^n)| \le 2^{\max n(H(\tilde{X}|X)+\delta_n)},$$

where the maximum is taken over all distributions $P_{U,X,\tilde{X}}$ such that $P_{U,X}=P_{U,\tilde{X}}$, and δ_n is a sequence of positive numbers which converges to 0 as $n\to\infty$. Since all of the sequences in $B_n(X^n)$ are typical we have:

$$P(\tilde{X}^n\in B_n(x^n)) \approx \frac{|B_n(x^n)|}{|A_\epsilon^n(x^n)|} \approx 2^{-n(I(\tilde{X};X)-\delta)} \triangleq 2^{-n\delta_X}.$$

Next we show that for $\tilde{x}^n\notin B_n(x^n)$:

$$\begin{aligned}&(1-2^{-n\delta_X})P(\underline{E}(x^n))P(\underline{E}(\tilde{x}^n))\\ &< P(\underline{E}(x^n),\underline{E}(\tilde{x}^n)) < (1+2^{-n\delta_X})P(\underline{E}(x^n))P(\underline{E}(\tilde{x}^n)).\end{aligned} \tag{37}$$

Note that by our construction x^n is mapped to a sequence in $\mathcal{C}\bigcap A_\epsilon^n(U|x^n)$ randomly and uniformly. So:

$$\begin{aligned}&P(\underline{E}(x^n)=c^n|\mathcal{C})\\ &=\frac{\mathbb{1}\left(c^n\in\mathcal{C}\bigcap A_\epsilon^n(U|x^n)\right)}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)|}=\frac{\mathbb{1}\left(c^n\in\mathcal{C}\bigcap A_\epsilon^n(U|x^n)\right)}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)-\{c^n\}|+1}\\ &\Rightarrow P(\underline{E}(x^n)=c^n)\\ &=P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n)\right)\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)-\{c^n\}|+1}\right).\end{aligned}$$

By a similar argument for $\tilde{x}^n\notin B_n(x^n)$ we have:

$$\begin{aligned}&P(\underline{E}(x^n)=c^n,\underline{E}(\tilde{x}^n)=\tilde{c}^n)=\\ &P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n),\tilde{c}^n\in\mathcal{C}\cap A_\epsilon^n(U|\tilde{x}^n)\right)\times\\ &\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)-\{c^n\}|+1}\frac{1}{|\mathcal{C}\bigcap A_{\epsilon^n(U|\tilde{x}^n)}-\{\tilde{c}^n\}|+1}\right).\end{aligned}$$

We show the following bound:

$$\begin{aligned}&P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n),\tilde{c}^n\in\mathcal{C}\cap A_\epsilon^n(U|\tilde{x}^n)\right)\\ &\le P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n)\right)P\left(\tilde{c}^n\in\mathcal{C}\cap A_\epsilon^n(U|\tilde{x}^n)\right)(1+2^{-n\delta_X}).\end{aligned} \tag{38}$$

Assume that $c^n\in A_\epsilon^n(U|x^n)$, and $\tilde{c}^n\in A_\epsilon^n(U|\tilde{x}^n)$, otherwise the two sides are equal to 0. The following equalities hold by the construction algorithm:

$$\begin{aligned}&P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n)\right)=P\left(c^n\in\mathcal{C}\right)=\frac{|\mathcal{C}|}{|A_\epsilon^n(U)|},\\ &P\left(\tilde{c}^n\in\mathcal{C}\cap A_\epsilon^n(U|\tilde{x}^n)\right)=P\left(\tilde{c}^n\in\mathcal{C}\right)=\frac{|\mathcal{C}|}{|A_\epsilon^n(U)|},\\ &P\left(c^n\in\mathcal{C}\cap A_\epsilon^n(U|x^n),\tilde{c}^n\in\mathcal{C}\cap A_\epsilon^n(U|\tilde{x}^n)\right)\\ &=P\left(c^n\in\mathcal{C},\tilde{c}^n\in\mathcal{C}\right)\\ &=\frac{\binom{|\mathcal{C}|}{2}}{\binom{A_\epsilon^n(U)}{2}}.\end{aligned}$$

Using the above it is straightforward to check that the bound in (38) holds. Similarly, it follows that

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)-\{c^n\}|+1}\frac{1}{|\mathcal{C}\bigcap A_{\epsilon^n(U|\tilde{x}^n)}-\{\tilde{c}^n\}|+1}\right)\le\\ &\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{C}\bigcap A_\epsilon^n(U|x^n)-\{c^n\}|+1}\right)\times\\ &\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{C}\bigcap A_{\epsilon^n(U|\tilde{x}^n)}-\{\tilde{c}^n\}|+1}\right)(1+2^{-n\delta_X}).\end{aligned}$$

Multiplying the two bound recovers the right-hand side of the inequality in (37). The left-hand side can be shown by similar arguments.

2) As n becomes large, the ith output element $E_i(X^n)$ is correlated with the input sequence X^n only through the ith input element X_i:

$$\begin{aligned}&\forall\delta>0,\exists n\in\mathbb{N}:m>n\Rightarrow\forall x^m\in\{0,1\}^m,v\in\{0,1\},\\ &|P_{\mathcal{S}}(E_i(X^m)=v|X^m=x^m)-P_{\mathcal{S}}(E_i(X^m)=v|X_i=x_i)|<\delta.\end{aligned}$$

The proof is as follows: For a fixed quantization function $\underline{e}:\{0,1\}^m\to\{0,1\}^m$, $\underline{e}(X^m)$ is a function of X^m. However, without the knowledge that which encoding function is used, $E_i(X^m)$ is related to X^m only through X_i. In other words, averaged over all encoding functions, the effects of the rest of the elements diminishes. We provide a proof of this statement below:

First, we are required to provide some definitions relating to the joint type of pairs of sequences. For binary strings u^m,x^m, define $N(a,b|u^m,x^m)\triangleq|\{j|u_j=a,x_j=b\}|$, that is the number of indices j for which the value of the pair (u_j,x_j) is (a,b). For $s,t\in\{0,1\}$, define $l_{s,t}\triangleq N(s,t|u^m,x^m)$, the vector $(l_{0,0},l_{0,1},l_{1,0},l_{1,1})$ is called the joint type of (u^m,x^m). For fixed x^m The set of sequences $T_{l_{0,0},l_{0,1},l_{1,0},l_{1,1}}=\{u^m|N(s,t|u^m,x^m)=l_{s,t},s,t\in\{0,1\}\}$, is the set of vectors which have joint type $(l_{0,0},l_{0,1},l_{1,0},l_{1,1})$ with the sequence x^m. Fix $m,\epsilon>0$, and define $\mathcal{L}_{\epsilon,n}\triangleq\{(l_{0,0},l_{0,1},l_{1,0},l_{1,1}):|\frac{l_{s,t}}{m}-P_{U,X}(s,t)|<\epsilon,\forall s,t\}$. Then for the conditional typical set $A_\epsilon^n(U|x^m)$ defined above we can write

$$A_\epsilon^n(U|x^m)=\bigcup_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}} T_{l_{0,0},l_{0,1},l_{1,0},l_{1,1}}.$$

The type of x^m, denoted by (l_0,l_1) is defined in a similar manner. Since $E_i(X^m)$ are chosen uniformly from the set $A_\epsilon^n(U|x^m)$, we have:

$$\begin{aligned}&P_{\mathcal{S}}(E_i(X^m)=v|X^m=x^m)=\frac{|\{u^m|u_1=v,u^m\in A_\epsilon^m(U|x^m)\}|}{|\{u^m|u^m\in A_\epsilon^m(U|x^m)\}|}\\ &=\frac{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}|\{u^m|u_1=v,u^m\in T_{l_{0,0},l_{0,1},l_{1,0},l_{1,1}}\}|}{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}|\{u^m|u^m\in T_{l_{0,0},l_{0,1},l_{1,0},l_{1,1}}\}|}\\ &=\frac{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}\binom{l_{x_1}-1}{l_{u_1,x_1}-1}\binom{l_{\tilde{x}_1}}{l_{u_1,\tilde{x}_1}}}{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}\binom{l_{x_1}}{l_{u_1,x_1}}\binom{l_{\tilde{x}_1}}{l_{u_1,\tilde{x}_1}}}\\ &=\frac{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}\frac{(l_{x_1}-1)!}{(l_{u_1,x_1}-1)!(l_{x_1}-l_{u_1,x_1})!}\frac{l_{\tilde{x}_1}!}{l_{u_1,\tilde{x}_1}!(l_{\tilde{x}_1}-l_{u_1,\tilde{x}_1})!}}{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}}\frac{l_{x_1}!}{l_{u_1,x_1}!(l_{x_1}-l_{u_1,x_1})!}\frac{l_{\tilde{x}_1}!}{l_{u_1,\tilde{x}_1}!(l_{\tilde{x}_1}-l_{u_1,\tilde{x}_1})!}}\end{aligned}$$

$$\overset{(a)}{=} \frac{\sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}} l_{u_1,x_1} \frac{1}{l_{u_1,x_1}!(l_{x_1}-l_{u_1,x_1})!} \frac{1}{l_{u_1,\bar{x}_1}!(l_{\bar{x}_1}-l_{u_1,\bar{x}_1})!}}{l_{x_1} \sum_{(l_{0,0},l_{0,1},l_{1,0},l_{1,1})\in\mathcal{L}_{\epsilon,n}} \frac{1}{l_{u_1,x_1}!(l_{x_1}-l_{u_1,x_1})!} \frac{1}{l_{u_1,\bar{x}_1}!(l_{\bar{x}_1}-l_{u_1,\bar{x}_1})!}}$$

$$\overset{(b)}{\Rightarrow} \frac{P_{U,X}(u_1,x_1)-\epsilon}{P_X(x_1)+\epsilon} \leq P_{\mathbb{S}}(E_i(X^m)=v|X^m=x^m) \leq \frac{P_{U,X}(u_1,x_1)+\epsilon}{P_X(x_1)-\epsilon}$$

$$\Rightarrow \exists m, \epsilon>0 : |P_{\mathbb{S}}(E_i(X^m)=v|X^m=x^m)-P_{U|X}(u_1|x_1)| \leq \delta.$$

In (a), we use the fact that for fixed x^m, $(l_{x_1}, l_{\bar{x}_1})$ is fixed to simplify the numerators. In (b) we have used that for jointly typical ϵ-sequences (u^m, x^m), $l_{u_1,x_1} \in [n(P_{U,X}(u_1,x_1)-\epsilon), n(P_{U,X}(u_1,x_1)+\epsilon)]$, and $l_{x_1} \in [n(P_X(x_1)-\epsilon), n(P_X(x_1)+\epsilon)]$.

3) The encoder is insensitive to permutations. Due to typicality encoding the probability that a vector x^n is mapped to y^n depends only on their joint type and is equal to the probability that $\pi(x^n)$ is mapped to $\pi(y^n)$. □

G. Proof of Proposition 4

Proof.

Fix $k, k' \in \mathbb{N}$. Define the permutation $\pi_{k\to k'} \in S_n$ as the permutation which switches the kth and k'th elements and fixes all other elements. Also, let $\mathcal{E}$ be the set of all mappings $e : \{0,1\}^n \to \{0,1\}^n$.

$$P_{\mathbb{S}}\left(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k} \mathbf{P}_{k,\mathbf{i}} > \gamma\right) = \sum_{\underline{e}\in\mathcal{E}} P_{\mathbb{S}}\left(\underline{e}\right) \mathbb{1}\left(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k|\underline{e}} \mathbf{P}_{k,\mathbf{i}} > \gamma\right)$$

$$\overset{(a)}{=} \sum_{\underline{e}\in\mathcal{E}} P_{\mathbb{S}}\left(\underline{e}_{\pi_{k\to k'}}\right) \mathbb{1}\left(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k} \mathbf{P}_{k,\mathbf{i}} > \gamma\,|\underline{e}\right)$$

$$\overset{(b)}{=} \sum_{\underline{g}\in\mathcal{E}} P_{\mathbb{S}}\left(\underline{g}\right) \mathbb{1}\left(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_k} \mathbf{P}_{\pi_{k\to k'}k,\pi_{k\to k'}\mathbf{i}} > \gamma\,|\underline{g}\right)$$

$$= \sum_{\underline{g}\in\mathcal{E}} P_{\mathbb{S}}\left(\underline{g}\right) \mathbb{1}\left(\sum_{\underline{l}:N_{\underline{l}}\leq m,\underline{l}\neq\pi_{k\to k'}\mathbf{i}_k} \mathbf{P}_{k',\underline{l}} > \gamma\,|\underline{g}\right)$$

$$= P_{\mathbb{S}}\left(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_{k'}} \mathbf{P}_{k',\mathbf{i}} > \gamma\right),$$

where in (a) we have used property 3) in Definition 8, and in (b) we have defined $\underline{g} \triangleq \underline{e}_{\pi_{k\to k'}}$ and used $\pi^2_{k\to k'} = 1$. □

H. Proof of Theorem 3

Proof.

From Proposition 4, it is enough to show the theorem holds for $k = 1$. For ease of notation we drop the subscript k for the rest of the proof and denote $\mathbf{P}_{1,\mathbf{i}}$ by $\mathbf{P}_{\mathbf{i}}$. By the Markov inequality, we have the following:

$$P_{\mathbb{S}}\Big(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbf{P}_{\mathbf{i}} \geq \gamma\Big) \leq \frac{\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbb{E}_{\mathbb{S}}(\mathbf{P}_{\mathbf{i}})}{\gamma}. \tag{39}$$

So, we need to show that $\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbb{E}_{\mathbb{S}}(\mathbf{P}_{\mathbf{i}})$ goes to 0 for all fixed m. We first prove the following claim.

Claim 1. *Fix* $\mathbf{i}$*, the following holds*:

$$\mathbb{E}_{\tilde{E},X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{E}|X_{\mathbf{i}})) = \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{\tilde{E},X^n|X_{\mathbf{i}}}(\tilde{E}|X_{\mathbf{i}})) + O(e^{-n\delta_X}).$$

Proof.

$$\mathbb{E}_{\tilde{E},X_{\mathbf{i}}}(\mathbb{E}^2_{X^n|X_{\mathbf{i}}}(\tilde{E}|X_{\mathbf{i}})) = \sum_{x_{\mathbf{i}},\tilde{e}} P(x_{\mathbf{i}})P(\tilde{e})\Big(\sum_{x_{\sim\mathbf{i}}} P(x_{\sim\mathbf{i}})\tilde{e}(x^n)\Big)^2$$

$$= \sum_{x_{\mathbf{i}},\tilde{e}} P(x_{\mathbf{i}})P(\tilde{e}) \sum_{x_{\sim\mathbf{i}}} \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}}} P(x_{\sim\mathbf{i}})P(y_{\sim\mathbf{i}})\tilde{e}(x^n)\tilde{e}(y^n)$$

$$= \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}}} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n)\tilde{E}(y^n))$$

$$= \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\in B_n(x^n)} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n)\tilde{E}(y^n))$$

$$+ \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\notin B_n(x^n)} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n)\tilde{E}(y^n))$$

$$\overset{(a)}{\leq} \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\in B_n(x^n)} P(y_{\sim\mathbf{i}})$$

$$+ \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\notin B_n(x^n)} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n)\tilde{E}(y^n))$$

$$= P(Y^n \in B_n(X^n)|Y_{\mathbf{i}} = X_{\mathbf{i}})$$

$$+ \sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\notin B_n(x^n)} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n)\tilde{E}(y^n))$$

$$\overset{(b)}{=} O(e^{-n\delta_X}) +$$

$$\sum_{x^n} P(x^n) \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}},y^n\notin B_n(x^n)} P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n))\mathbb{E}_{\tilde{E}}(\tilde{E}(y^n))$$

$$\leq O(e^{-n\delta_X}) + P(Y^n \in B_n(X^n)|Y_{\mathbf{i}} = X_{\mathbf{i}})$$

$$+ \sum_{x_{\mathbf{i}}} P(x_{\mathbf{i}}) \sum_{x_{\sim\mathbf{i}}} \sum_{y^n:y_{\mathbf{i}}=x_{\mathbf{i}}} P(x_{\sim\mathbf{i}})P(y_{\sim\mathbf{i}})\mathbb{E}_{\tilde{E}}(\tilde{E}(x^n))\mathbb{E}_{\tilde{E}}(\tilde{E}(y^n))$$

$$= O(e^{-n\delta_X}) + \mathbb{E}_{X_{\mathbf{i}}}(\mathbb{E}^2_{\tilde{E},X^n|X_{\mathbf{i}}}(\tilde{E}|X_{\mathbf{i}})).$$

In (a) we use the fact that $\tilde{E} \leq 1$ by definition, in (b) follows from property 1) in Definition 8. □

Define $\bar{E}_{\mathbf{i}} = \mathbb{E}_{\tilde{E}}(\tilde{E}_{\mathbf{i}}) = \mathbb{E}_{\tilde{E}|X_{\mathbf{i}}}(\tilde{E}|X_{\mathbf{i}}) - \sum_{\mathbf{j}<\mathbf{i}} \bar{E}_{\mathbf{j}}$, and also define $\bar{P}_{\mathbf{i}} \triangleq Var(\bar{E}_{\mathbf{i}})$. Using the above claim we have:

$$P_{\mathbb{S}}\Big(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbf{P}_{\mathbf{i}} \geq \gamma\Big) \leq \frac{\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbb{E}_{\mathbb{S}}(\mathbf{P}_{\mathbf{i}})}{\gamma}$$

$$\leq \frac{2^m O(e^{-n\delta_X}) + \sum_{\mathbf{i}:N_{\mathbf{i}}\leq m} \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}}) - \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma}. \tag{40}$$

Using the arguments from the proof of Lemma 3, we can see that the properties stated in that Proposition hold for $\bar{E}_{\mathbf{i}}$ as well. By the same results as in Lemma 3 and Corollary 1, we have that $\sum_{\mathbf{i}\in\{0,1\}^n} \bar{P}_{\mathbf{i}} = \bar{P}_{\underline{1}}$. Following the calculations in (40):

$$P_{\mathbb{S}}\Big(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m,\mathbf{i}\neq\mathbf{i}_1} \mathbf{P}_{\mathbf{i}} \geq \gamma\Big)$$

$$\leq \frac{2^m O(e^{-n\delta_X}) + \sum_{\mathbf{i}:N_{\mathbf{i}}\leq m} \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}}) - \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma}$$

$$\leq \frac{2^m O(e^{-n\delta_X}) + \sum_{\mathbf{i}\in\{0,1\}^n} \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}}) - \mathbb{E}_{\mathbb{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma}$$

$$
\begin{aligned}
&= \frac{2^m O(e^{-n\delta_X}) + \mathbb{E}_{\mathcal{S}}(\sum_{\mathbf{i}\in\{0,1\}^n} \bar{P}_{\mathbf{i}}) - \mathbb{E}_{\mathcal{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma} \\
&= \frac{2^m O(e^{-n\delta_X}) + \mathbb{E}_{X^n}\left(\mathbb{E}^2_{\tilde{E}|X^n}(\tilde{E}(X^n)|X^n)\right) - \mathbb{E}_{\mathcal{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma} \\
&\leq \frac{2^m O(e^{-n\delta_X}) + \mathbb{E}_{\mathcal{S}}(\bar{P}_{\mathbf{i}_1}) + O(\epsilon) - \mathbb{E}_{\mathcal{S}}(\bar{P}_{\mathbf{i}_1})}{\gamma} \\
&= \frac{2^m O(e^{-n\delta_X}) + O(\epsilon)}{\gamma},
\end{aligned}
$$

where in the last inequality we have used the second property in Definition 8. The last line goes to 0 as $n \to \infty$. This completes the proof. □

I. Proof of Theorem 4

Proof. From Theorem 1, we have:

$$\mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}} - 2\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} \leq P(E(X^n) \neq F(Y^n)).$$

From Theorem 3 we have:

$$
\begin{aligned}
&\forall m \in \mathbb{N}, \gamma > 0, P_{\mathcal{S}}\Big(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m, \mathbf{i}\neq\mathbf{i}_1} \mathbf{P}_{\mathbf{i}} < \gamma\Big) \to 1, \\
&P_{\mathcal{S}}\Big(\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m, \mathbf{i}\neq\mathbf{i}_1} \mathbf{Q}_{\mathbf{i}} < \gamma\Big) \to 1. \qquad (41)
\end{aligned}
$$

Note that:

$$
\begin{aligned}
&\sum_{\mathbf{i}:N_{\mathbf{i}}\leq m, \mathbf{i}\neq\mathbf{i}_1} \mathbf{P}_{\mathbf{i}} < \gamma, \sum_{\mathbf{i}:N_{\mathbf{i}}\leq m, \mathbf{i}\neq\mathbf{i}_1} \mathbf{Q}_{\mathbf{i}} < \gamma \Rightarrow \\
&\sum_{\mathbf{i}} C_{\mathbf{i}}\mathbf{P}_{\mathbf{i}}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}}^{\frac{1}{2}} > (1-2\epsilon)(\mathbf{P}_{\mathbf{i}_1} + \gamma)^{\frac{1}{2}}(\mathbf{Q}_{\mathbf{i}_1} + \gamma)^{\frac{1}{2}} \\
&+ (1-2\epsilon)^m \mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}}, \qquad (42)
\end{aligned}
$$

which converges to $(1-2\epsilon)\mathbf{P}_{\mathbf{i}_1}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}_1}^{\frac{1}{2}} + (1-2\epsilon)^m\mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}}$ as $\gamma \to 0$. Also $C_{\mathbf{i}}$ is decreasing in $N_{\mathbf{i}}$ and goes to 0 as $N_{\mathbf{i}} \to \infty$. Choose γ small enough and m large enough such that $(1-2\epsilon)(\mathbf{P}_{\mathbf{i}_1} + \gamma)^{\frac{1}{2}}(\mathbf{Q}_{\mathbf{i}_1} + \gamma)^{\frac{1}{2}} + (1-2\epsilon)^m\mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}} - (1-2\epsilon)\mathbf{P}_{\mathbf{i}_1}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}_1}^{\frac{1}{2}} < \delta$. Then Equations (41) and (42) gives

$$P_{\mathcal{S}}\left(P_{X^n,Y^n}\left(E(X^n) \neq F(Y^n)\right) < \zeta\right) \to 0,$$

where $\zeta = 2\mathbf{P}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}} - 2(1-2\epsilon)\mathbf{P}_{\mathbf{i}_1}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{i}_1}^{\frac{1}{2}} - \delta$. This is equivalent to the statement of the theorem. □

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editor for their comments which led to significant improvements in the manuscript.

REFERENCES

[1] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. Appl. Math.*, vol. 28, no. 1, pp. 100–113, 1975.

[2] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[3] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic, 1981.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Applications of Mathematics). New York, NY, USA: Springer, 1998.

[5] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.

[6] T. M. Cover, A. El Gamal, and M. Salehi, "Multiple access channels with arbitrarily correlated sources," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 6, pp. 648–657, Nov. 1980.

[7] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 3, pp. 306–311, May 1979.

[8] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems Control Inf. Theory*, vol. 2, no. 2, pp. 119–162, 1972.

[9] H. O. Hirschfeld, "A connection between correlation and contingency," *Math. Proc. Cambridge Philos. Soc.*, vol. 31, no. 4, pp. 520–524, 1935.

[10] A. Rényi, "New version of the probabilistic generalization of the large sieve," *Acta Math. Acad. Sci. Hungarica*, vol. 10, nos. 1–2, pp. 217–226, 1959.

[11] A. B. Wagner, B. G. Kelly, and Y. G. Altug, "Distributed rate-distortion with common components," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4035–4057, Jul. 2011.

[12] F. S. Chaharsooghi, A. G. Sahebi, and S. S. Pradhan, "Distributed source coding in absence of common components," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1362–1366.

[13] A. Bogdanov and E. Mossel, "On extracting common random bits from correlated sources," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6351–6355, Oct. 2011.

[14] I. Csiszar and P. Narayan, "Common randomness and secret key generation with a helper," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 344–366, Mar. 2000.

[15] A. Mahajan, A. Nayyar, and D. Teneketzis, "Identifying tractable decentralized control problems on the basis of information structure," in *Proc. 46th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2008, pp. 1440–1449.

[16] G. Pichler, P. Piantanida, and G. Matz, "Dictator functions maximize mutual information," 2016, *arXiv:1604.02109*. [Online]. Available: https://arxiv.org/abs/1604.02109

[17] Y. Geng, A. Gohari, C. Nair, and Y. Yu, "On Marton's inner bound and its optimality for classes of product broadcast channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 22–41, Jan. 2014.

[18] J. Chen and T. Berger, "Robust distributed source coding," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3385–3398, Aug. 2008.

[19] F. Shirani and S. S. Pradhan, "Finite block-length gains in distributed source coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 1702–1706.

[20] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[21] B. Ghazi, P. Kamath, and M. Sudan, "Decidability of non-interactive simulation of joint distributions," in *Proc. IEEE 57th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2016, pp. 545–554.

[22] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*. New York, NY, USA: Academic, 1972.

[23] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, vol. 9. New York, NY, USA: Dover, 1972.

[24] W. Kang and S. Ulukus, "A new data processing inequality and its applications in distributed source and channel coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 56–69, Jan. 2011.

[25] S. Y. Tung, "Multiterminal source coding," Ph.D. dissertation, School Elect. Eng., Cornell Univ., Ithaca, NY, USA, 1978.

[26] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 502–521, Jul. 1987.

[27] M. Salehi and E. Kurtas, "Interference channels with correlated sources," in *Proc. IEEE Int. Symp. Inf. Theory*, Jan. 1993, p. 208.

[28] S. S. Pradhan, "Approximation of test channels in source coding," in *Proc. Conf. Inf. Syst. Sci. (CISS)*, 2004.

[29] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul. 1948.

[30] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.

[31] T. F. N. Baader, *Term Rewriting and All That*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[32] Y. Huang, J. Benesty, J. Chen, and I. Cohen, *Noise Reduction in Speech Processing*. New York, NY, USA: Springer, 2009.

[33] S. Kamath and V. Anantharam, "On non-interactive simulation of joint distributions," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3419–3435, Jun. 2016.