

Syracuse University

**SURFACE**

---

Electrical Engineering and Computer Science -  
Technical Reports

College of Engineering and Computer Science

---

4-11-2011

## The Common Information for N Dependent Random Variables

Wei Liu

*Syracuse University*

Ge Xu

*Syracuse University*

Follow this and additional works at: [https://surface.syr.edu/eecs\\_techreports](https://surface.syr.edu/eecs_techreports)



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Liu, Wei and Xu, Ge, "The Common Information for N Dependent Random Variables" (2011). *Electrical Engineering and Computer Science - Technical Reports*. 49.

[https://surface.syr.edu/eecs\\_techreports/49](https://surface.syr.edu/eecs_techreports/49)

This Report is brought to you for free and open access by the College of Engineering and Computer Science at SURFACE. It has been accepted for inclusion in Electrical Engineering and Computer Science - Technical Reports by an authorized administrator of SURFACE. For more information, please contact [surface@syr.edu](mailto:surface@syr.edu).



# Department of Electrical Engineering and Computer Science

## Technical Report

SYR-EECS-2011-05

April 11, 2011

### The Common Information for $N$ Dependent Random Variables

Wei Liu      wliu28@syr.edu

Ge Xu        gexu@syr.edu

Biao Chen    bichen@syr.edu

**ABSTRACT:** This paper generalizes Wyner's definition of common information of a pair of random variables to that of  $N$  random variables. We prove coding theorems that show the same operational meanings for the common information of two random variables generalize to that of  $N$  random variables. As a byproduct of our proof, we show that the Gray-Wyner source coding network can be generalized to  $N$  source sequences with  $N$  decoders. We also establish a monotone property of Wyner's common information which is in contrast to other notions of the common information, specifically Shannon's mutual information and Gács and Körner's common randomness. Examples about the computation of Wyner's common information of  $N$  random variables are also given.

**KEYWORDS:** Wyner's common information, Gray-Wyner source coding network, distribution approximation, circularly symmetric binary source

Syracuse University - Department of EECS,  
4-206 CST, Syracuse, NY 13244  
(P) 315.443.2652 (F) 315.443.2583  
<http://ecs.syr.edu>

# The Common Information of $N$ Dependent Random Variables

Wei Liu

Department of EECS  
Syracuse University  
Email: wliu28@syr.edu

Ge Xu

Department of EECS  
Syracuse University  
Email: gexu@syr.edu

Biao Chen

Department of EECS  
Syracuse University  
Email: bichen@syr.edu

**Abstract**—This paper generalizes Wyner’s definition of common information of a pair of random variables to that of  $N$  random variables. We prove coding theorems that show the same operational meanings for the common information of two random variables generalize to that of  $N$  random variables. As a byproduct of our proof, we show that the Gray-Wyner source coding network can be generalized to  $N$  source sequences with  $N$  decoders. We also establish a monotone property of Wyner’s common information which is in contrast to other notions of the common information, specifically Shannon’s mutual information and Gács and Körner’s common randomness. Examples about the computation of Wyner’s common information of  $N$  random variables are also given.

## I. INTRODUCTION

Consider a pair of dependent random variables  $X$  and  $Y$  with joint distribution  $P(x, y)$ . Characterizing the common information between  $X$  and  $Y$  has been a topic of research interest in the past decades [1]–[5]. There have been three classical notions reported in the literature.

**Shannon’s [6] mutual information**  $I(X; Y)$

Shannon’s mutual information measures how much uncertainty can be reduced with respect to one random variable by observation the other random variable. In the case that  $X$  and  $Y$  are independent, mutual information  $I(X; Y) = 0$ , indicating that observing one variable  $X$  does not give any information about  $Y$  and vice versa. Shannon’s mutual information carries operational meanings that are instrumental in laying the foundation for information theory.

**Gács and Körner’s [1] common randomness**  $K(X, Y)$

Consider a pair of independent and identically distributed random sequences  $X^n, Y^n$  with each pair  $(X_i, Y_i) \sim P(x, y)$ . These two sequences are observed respectively by two nodes, which attempt to map the sequences onto a common message set  $\mathcal{W}$ . Specifically, let  $f_n$  and  $g_n$  be such mappings, i.e.,

$$\begin{aligned} f_n &: \mathcal{X}^n \rightarrow \mathcal{W}, \\ g_n &: \mathcal{Y}^n \rightarrow \mathcal{W}. \end{aligned}$$

Define  $\epsilon_n = Pr(W_1 \neq W_2)$  where  $W_1 = f_n(X^n)$  and  $W_2 = g_n(Y^n)$ . Gács and Körner’s common randomness is defined as

$$K(X, Y) = \lim_{n \rightarrow \infty, \epsilon_n \rightarrow 0} \sup \frac{1}{n} H(W_1).$$

Gács and Körner’s common randomness has found extensive applications in cryptography, i.e., for key generation [7]–[9]. On the other hand, the common randomness notion is rather restrictive as it equals 0 in most cases except for the following special case (or random variable pairs that can be converted

to such distributions through relabeling of realizations, i.e., permutation of joint distribution matrix). Let  $X$  and  $Y$  be  $X = (X', V)$  and  $Y = (Y', V)$ , respectively, where  $X', Y', V$  are independent. Clearly, the common part between  $X$  and  $Y$  is  $V$  and it follows that  $K(X; Y) = H(V)$ . Note that for this example  $I(X; Y) = K(X; Y) = H(V)$ .

**Wyner’s [4] common information**  $C(X, Y)$

Wyner’s common information is defined as

$$C(X, Y) = \min_{X \rightarrow W \rightarrow Y} I(XY; W). \quad (1)$$

Thus the hidden (or auxiliary) variable  $W$  induces a Markov chain  $X - W - Y$ , or, equivalently, a conditional independence structure of  $X, Y$  being independent given  $W$ . Wyner gave two operational meanings for the above definition. The first approach is shown in Fig. 1. The encoder observes a pair of sequences  $(X^n, Y^n)$ , and map them to three messages  $W_0, W_1, W_2$ , taking values in alphabets of respective sizes  $2^{nR_0}, 2^{nR_1}, 2^{nR_2}$ . Decoder 1, upon receiving  $(W_0, W_1)$ , needs to reproduce  $X^n$  reliably while decoder 2, upon receiving  $(W_0, W_2)$ , needs to reproduce  $Y^n$  reliably. Let  $C_1$  be the infimum of all admissible  $R_0$  for the system in Fig. 1 such that the total rate  $R_0 + R_1 + R_2 \approx H(X, Y)$ .

The second approach is shown in Fig. 2. A common input  $W$ , uniformly distributed on  $\mathcal{W} = \{1, \dots, 2^{nR_0}\}$  is given to two separate processors which are otherwise independent of each other. These processors (random variable generators) generating independent and identically distributed sequences according to  $q_1(X|W)$  and  $q_2(Y|W)$  respectively. The output sequences of the two processors are denoted by  $\tilde{X}^n$  and  $\tilde{Y}^n$  respectively. Thus the joint distribution of the output sequences is,

$$Q(\tilde{X}^n, \tilde{Y}^n) = \sum_{w \in \mathcal{W}} \frac{1}{|\mathcal{W}|} q_1(X^n|W) q_2(Y^n|W). \quad (2)$$

Define  $C_2$  of  $(X, Y)$  to be infimum of rate  $R_0$  for the common input such that  $q(\tilde{X}^n, \tilde{Y}^n)$  close to  $p(X^n, Y^n)$ , where the closeness is defined using the average divergence of the two distributions

$$D_n(P, Q) = \frac{1}{n} \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} P(x^n, y^n) \log \frac{P(x^n, y^n)}{Q(x^n, y^n)}. \quad (3)$$

Wyner proved that

$$C_1 = C_2 = C(X, Y). \quad (4)$$

It was observed in [4] that

$$K(X, Y) \leq I(X; Y) \leq C(X, Y). \quad (5)$$

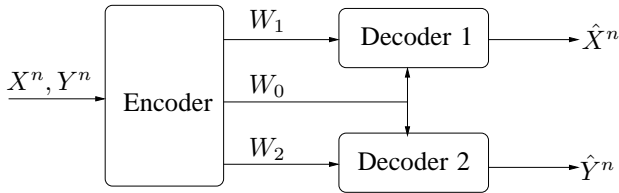


Fig. 1. Source coding over a simple network.

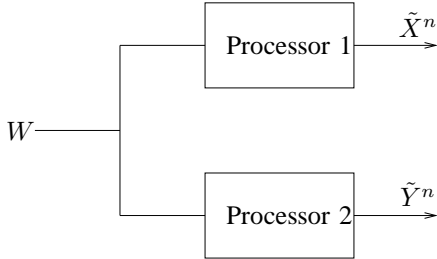


Fig. 2. Random variable generators.

Wyner [4] and Witsenhausen [5] also provide several examples on how to calculate the common information  $C(X, Y)$ . For the example of  $X = (X', V)$  and  $Y = (Y', V)$  with  $(X', Y', V)$  mutually independent,  $C(X, Y) = I(X; Y) = K(X, Y) = H(V)$ .

Generalizing of mutual information to  $N$  random variables was first reported in [10]. The generalization comes from the observation that for a pair of random variables, Shannon's information measures is consistent with the Venn diagram for set operation and a comprehensive treatment was available in [11], [12]. Gács and Körner's common randomness was recently generalized to multiple random variables by Tyagi, Narayan and Gupta in [13], which extends the encoding process in the definition of common randomness to that of  $N$  terminals.

In this paper, we generalize Wyner's common information of a pair of random variables to that of  $N$  dependent variables. We show that the operational meaning defined in both approaches are still valid. Moreover, we establish some monotone property of such generalization which contrast to the notion of 'common' information. Specifically, we show that the common information does *not* decrease as the number of variables increases while keeping the same marginal distribution. This is different from the other two notions of common information. Examples on evaluating  $C(X_1, X_2, \dots, X_N)$  are given for circularly symmetric binary sources and the asymptotic results are also studied.

The rest of this paper is organized as follows. Section II gives the problem formulation and main results. Section III gives some examples and discussions. Section IV concludes the paper.

## II. PROBLEM STATEMENT AND MAIN RESULTS

Let  $X_1, X_2, \dots, X_N$  be random variables that take values on the finite alphabet sets  $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N$  with joint distribution  $P(x_1, x_2, \dots, x_N)$ . Our generalization of Wyner's

common information is to define a similar measure for  $N$  random variables by preserving the conditional independence structure through the introduction of an auxiliary random variable. Specifically, we define

$$C(X_1, X_2, \dots, X_N) \triangleq \inf I(X_1, X_2, \dots, X_N; W), \quad (6)$$

where the infimum is taken over all the joint distributions of  $(X_1, X_2, \dots, X_N, W)$  such that

$$\sum_w P(x_1, x_2, \dots, x_n, w) = P(x_1, x_2, \dots, x_n), \quad (7)$$

$$P(x_1, \dots, x_n | w) = \prod_{i=1}^n P(x_i | w). \quad (8)$$

Thus the marginal distribution of  $(X_1, X_2, \dots, X_N)$  is  $P(x_1, x_2, \dots, x_N)$  and  $(X_1, \dots, X_N)$  are conditionally independent given  $W$ .

We now give formal definitions of  $C_1$  and  $C_2$  for  $N$  random variables. Consider  $N$  length- $n$  independent and identically distributed source sequences  $(x_1^n, x_2^n, \dots, x_N^n)$  with  $(X_{1i}, X_{2i}, \dots, X_{Ni}) \sim p(x_1, x_2, \dots, x_N)$ , i.e.,

$$P^{(n)}(x_1^n, x_2^n, \dots, x_N^n) = \prod_{i=1}^n P(x_{1i}, x_{2i}, \dots, x_{Ni}). \quad (9)$$

For the Gray-Wyner source coding network, we start with the definition of encoder-decoders.

*Definition 1:* A  $(n, \mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_N)$  code consists of the following:

- An encoder mapping

$$f: \mathcal{X}_1^n \times \mathcal{X}_2^n \times \dots \times \mathcal{X}_N^n \rightarrow \mathcal{M}_0 \times \mathcal{M}_1 \times \dots \times \mathcal{M}_N,$$

where  $\mathcal{M}_i = \{1, 2, \dots, 2^{nR_i}\}$ .

- $N$  decoders  $g_i$ , for  $i = 1, 2, \dots, N$ ,

$$g_i: \mathcal{M}_i \times \mathcal{M}_0 \rightarrow \mathcal{X}_i^n. \quad (10)$$

The probability of error is defined as

$$P_e^{(n)} = \Pr\{(\hat{X}_1^n \hat{X}_2^n \dots \hat{X}_N^n) \neq (X_1^n, X_2^n, \dots, X_N^n)\}, \quad (11)$$

where  $\hat{X}_i^n = g_i(M_i, M_0)$  for  $i = 1, \dots, N$ .

*Definition 2:* A number  $R_0$  is said to be *achievable* if for any  $\epsilon > 0$ , we can find an  $n$  sufficiently large such that there exists a  $(n, \mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_N)$  code with

$$\mathcal{M}_0 \leq 2^{nR_0} \quad (12)$$

$$P_e^{(n)} \leq \epsilon, \quad (13)$$

$$\frac{1}{n} \sum_{i=0}^N \log \mathcal{M}_i \leq H(X_1, X_2, \dots, X_N) + \epsilon. \quad (14)$$

As with the case for two random variables,  $C_1$  is defined as the infimum of all achievable  $R_0$ .

For the second approach of approximating joint distribution, we again start with the following definition.

*Definition 3:* An  $(n, \mathcal{M}, \Delta)$  generator consists of the following:

- a message set  $\mathcal{W} \in \{1, 2, \dots, 2^{nR}\}$ ;

- for all  $w \in \mathcal{W}$  and  $N$  conditional probability distributions  $q_i^{(n)}(x_i^n|w)$ , for  $i = 1, 2, \dots, N$ , define the probability distribution on  $\mathcal{X}_1^n \times \mathcal{X}_2^n \times \dots \times \mathcal{X}_N^n$

$$Q^{(n)}(X_1^n, X_2^n, \dots, X_N^n) = \sum_{w \in \mathcal{W}} \frac{1}{\mathcal{M}} \prod_{i=1}^N q_i^{(n)}(x_i^n|w). \quad (15)$$

Thus the  $N$  processors serve as random number generators each generating independent and identically distributed (i.i.d.) sequence  $\hat{X}_i^n$  according to  $q(x_i|w)$  and the output of the processors follow joint distribution defined in (9). Let

$$\Delta = D_n(P^{(n)}; Q^{(n)}) = \frac{1}{n} \sum_{x_i^n \in \mathcal{X}_i^n, i=1,2,\dots,N} P^{(n)} \log \frac{P^{(n)}}{Q^{(n)}}, \quad (16)$$

where  $P^{(n)}$  and  $Q^{(n)}$  are defined as in (9) and (15) respectively.

*Definition 4:* A number  $R$  is said to be achievable if for all  $\epsilon > 0$ , we can find an  $n$  sufficiently large such that there exists a  $(n, \mathcal{M}, \Delta)$  generator with  $\mathcal{M} \leq 2^{nR}$  and  $\Delta \leq \epsilon$ .

We define  $C_2$  as the infimum of all achievable  $R$ .

The main result of this paper is the following theorem.

*Theorem 1:*

$$C_1 = C_2 = C(X_1, X_2, \dots, X_N). \quad (17)$$

The proof of Theorem 1 is given in the Appendix. Thus both  $C_1$  and  $C_2$  admit single letter characterization which coincides with  $C(X_1, \dots, X_N)$ .

### III. EXAMPLES AND DISCUSSIONS

We start with the following example. Let  $X = (X', U, V)$ ,  $Y = (Y', V, W)$  and  $Z = (Z', W, U)$  where the random variables  $X', Y', Z', U, V, W$  are mutually independent. It is easy to show that for this example

$$I(X; Y; Z) = K(X, Y, Z) = 0,$$

whereas

$$C(X, Y, Z) = H(UVW).$$

On the other hand,

$$\begin{aligned} C(X, Y) &= H(V), \\ C(X, Z) &= H(U), \\ C(Y, Z) &= H(W). \end{aligned}$$

What is interesting is that the inclusion of an additional variable increases the common information. This is somewhat surprising: if the information is *common* it ought to be non-increasing when more random variables are included. Indeed, we can prove the following general result:

*Lemma 1:* Let  $(X_1, \dots, X_N) \sim p(x_1, \dots, x_N)$ . For any two sets  $A, B$  that satisfy  $A \subseteq B \subseteq \mathcal{N} = \{1, 2, \dots, N\}$ ,

$$C(\mathbf{X}_A) \leq C(\mathbf{X}_B), \quad (18)$$

where  $\mathbf{X}_A = \{X_i, i \in A\}$  and  $\mathbf{X}_B = \{X_i, i \in B\}$ .

*Proof:* Let  $W'$  be the  $W$  that achieves  $C(\mathbf{X}_B)$ , i.e.,  $I(W'; \mathbf{X}_B) = \inf I(W; \mathbf{X}_B)$ . But  $A \subseteq B$ , thus  $\mathbf{X}_B$  conditionally independent given  $W'$  implies that  $\mathbf{X}_A$  is conditionally independent given  $W'$ . Thus

$$\begin{aligned} I(\mathbf{X}_B; W') &\geq I(\mathbf{X}_A; W') \\ &\geq \inf I(\mathbf{X}_A; W) \end{aligned}$$

where the infimum is taken over all  $W$  such that  $\mathbf{X}_A$  is independent given  $W$ .

This monotone property perhaps suggests that the name common information, while meaningful for pair of variables, no longer suits the generalization to  $N$  variables. We comment here that Gács and Körner's common randomness follows a different monotone property

$$K(\mathbf{X}_A) \geq K(\mathbf{X}_B)$$

while there is no definitive inequality relationship for mutual information.

As a consequence, we have for any  $N$  random variables

$$C(X_1, X_2, \dots, X_N) \geq K(X_1, X_2, \dots, X_N).$$

We now examine another example in which Wyner's common information increases as the number of the observations increases. Moreover the common information eventually converges and the asymptote suggests that the notion of common information may have potential application in certain inference problem.

Consider first the example of three binary random variables  $X_1, X_2, X_3$  with joint distribution

$$P(x_1, x_2, x_3) = \begin{cases} \frac{1}{2} - \frac{3}{4}a_0 & \text{if } x_1 = x_2 = x_3 \\ \frac{1}{4}a_0 & \text{otherwise} \end{cases} \quad (19)$$

where the parameter  $a_0$  satisfies  $0 \leq a_0 \leq \frac{1}{2}$ .

It can be easily verified that

$$Pr\{X_i = 0\} = \frac{1}{2}, \quad (20)$$

for  $i = 1, 2, 3$  and that for  $1 \leq i, j \leq 3, i \neq j$ ,

$$Pr(X_i = x_i, X_j = x_j) = \frac{1}{2}(1 - a_0)\delta_{x_i, x_j} + \frac{1}{2}a_0(1 - \delta_{x_i, x_j}), \quad (21)$$

where  $\delta_{a,b} = 1$  if  $a = b$  and 0 otherwise.

Thus, each pair of  $(X_i, X_j)$ ,  $i \neq j$ , can be viewed as a doubly symmetric binary source as defined in [4]. We refer to this set of exchangeable binary sources circularly symmetric binary source. For such circularly symmetric binary source  $(X_1, X_2, X_3)$  with joint distribution given in (19) and random variables  $(X_1, X_2, X_3, W)$  that satisfy (7) and (8), we have the following lemma.

*Lemma 2:*

$$H(X_1|W) + H(X_2|W) + H(X_3|W) \leq 3h(a_1), \quad (22)$$

where  $a_1 = \frac{1}{2} - \frac{1}{2}(1 - 2a_0)^{\frac{1}{2}}$ .

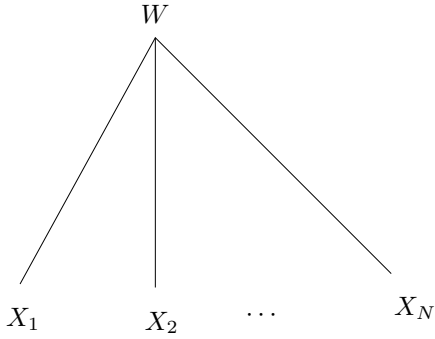


Fig. 3. A simple Bayesian graph model.

This lemma is a direct consequence of Wyner's result on doubly symmetric binary source [4]. Therefore, we have,

$$\begin{aligned}
& I(X_1 X_2 X_3; W) \\
&= H(X_1 X_2 X_3) - H(X_1 X_2 X_3 | W), \\
&= H(X_1 X_2 X_3) - \sum_{i=1}^3 H(X_i | W), \\
&\geq H(X_1 X_2 X_3) - 3h(a_1), \\
&= 1 + h(a_0) + a_0 + (1 - a_0)h\left(\frac{a_0}{2(1 - a_0)}\right) - 2h(a_1), \tag{23}
\end{aligned}$$

This lower bound can indeed be achieved by choosing the following random variables. Let  $W$  be a random variable with  $p_W(0) = p_W(1) = 1/2$ , i.e., a Bernoulli(1/2) random variable. Let each  $X_i$  be the output of a binary symmetric channel (BSC) with crossover probability  $a_1$  with  $W$  as input. The channels share the common input  $W$  but are otherwise independent of each other. This is illustrated in the simple Bayesian graph model in Fig. 3 with  $N = 3$  where each link represents a BSC with crossover probability  $a_1$ .

Thus, the common information of this circularly symmetric binary source is,

$$\begin{aligned}
C(X_1, X_2, X_3) &= 1 + a_0 + h(a_0) + \\
&\quad (1 - a_0)h\left(\frac{a_0}{2(1 - a_0)}\right) - 3h(a_1), \tag{24}
\end{aligned}$$

Notice that any pair of  $(X_i, X_j)$  is a doubly symmetric binary source [4], therefore,

$$C(X, Y) = 1 + h(a_0) - 2h(a_1).$$

It is straightforward to check that

$$C(X, Y, Z) > C(X, Y)$$

when  $0 < a_0 < \frac{1}{2}$ . This is also shown numerically in Fig. 4.

We now study the generalization of above example to arbitrary  $N$  and in particular the asymptotic value of the common information for the circularly symmetric binary sources.

Consider  $N$  binary random variables  $X_1, X_2, \dots, X_N$  with joint distribution  $p(X_1, X_2, \dots, X_N)$  generated by an underlying Bayesian graph model as in Fig. 3, where  $W$  is a

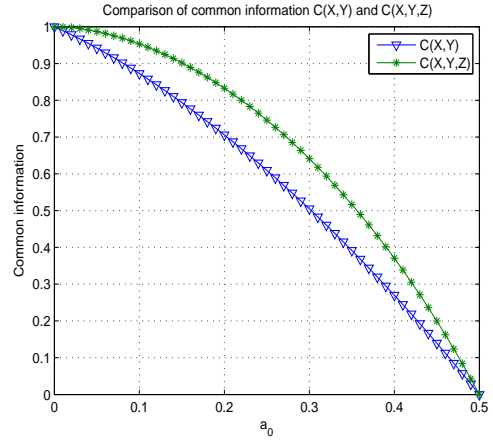


Fig. 4. Comparison of common information.

Bernoulli(1/2) random variable and each  $X_i$ ,  $i = 1, 2, \dots, N$ , is the output of a BSC with crossover probability  $a_1$  ( $0 \leq a_1 \leq \frac{1}{2}$ ) with a common input  $W$ . Hence, for  $x_1, x_2, \dots, x_N \in \{0, 1\}$ ,

$$P(x_1, x_2, \dots, x_N) = \sum_{w \in \{0,1\}} \frac{1}{2} \prod_{i=1}^N P_i(x_i | w), \tag{25}$$

where for each  $i = 1, 2, \dots, N$ ,  $p_i(x_i | w) = (1 - a_1)$  if  $x_i = w$  and  $a_1$  otherwise.

Similarly, we have,

$$\sum_{i=1}^N H(X_i | W) \leq N h(a_1), \tag{26}$$

for any random variable  $W$  that satisfies (7) and (8).

Therefore,  $C(X_1, X_2, \dots, X_N)$  can be lower bounded by

$$C(X_1, X_2, \dots, X_N) \geq H(X_1, X_2, \dots, X_N) - N h(a_1). \tag{27}$$

On the other hand, the above lower bound is achievable by exactly the same  $W$  in the above Bayesian model. Hence, we have,

$$C(X_1, X_2, \dots, X_N) = H(X_1, X_2, \dots, X_N) - N h(a_1), \tag{28}$$

where  $H(X_1, X_2, \dots, X_N)$  can be calculated from (25).

Now consider the above model but with increasing  $N$ . For any  $\epsilon$  and  $a_1 < 1/2$ , it is clear that

$$H(W | X_1, X_2, \dots, X_N) < \epsilon$$

for  $N$  sufficiently large. This can be established by the Fano's inequality as one can estimate  $W$  with arbitrary reliability given  $X_1, \dots, X_N$  for sufficiently large  $N$ . Therefore,

$$\begin{aligned}
& C(X_1, X_2, \dots, X_N) \\
&= H(X_1, X_2, \dots, X_N) - N h(a_1), \\
&= H(X_1, X_2, \dots, X_N, W) - N h(a_1) \\
&\quad - H(W | X_1, X_2, \dots, X_N), \\
&\geq H(W) - \epsilon, \tag{29}
\end{aligned}$$

where the last step is from the fact that  $H(X_1, X_2, \dots, X_N|W) = Nh(a_1)$ . On the other hand,

$$C(X_1, \dots, X_N) \leq H(W)$$

for any  $N$ . Thus, for  $a_1 < 1/2$ ,

$$\lim_{N \rightarrow \infty} C(X_1, X_2, \dots, X_N) = H(W) = 1$$

If  $a_1 = 1/2$ , then  $X_1, \dots, X_N$  are mutually independent hence  $C(X_1, \dots, X_N) = 0$ .

#### IV. CONCLUSIONS

This paper generalized Wyner's common information, defined for a pair of random variables, to that of  $N$  dependent random variables. We showed that it is the minimum common information rate  $R_0$  needed for  $N$  separate decoders to recover their intended sources losslessly while keeping the total rate close to the entropy bound. It is also equivalently to the smallest rate of the common input to  $N$  independent processors (random number generators), such that the output distribution is approximately the same as the given joint distribution. It was shown that such generalization leads to the phenomenon of 'common' information non-decreasing as the number of sources increases.

For the example of circularly symmetric binary sources, we show that common information not only increases as  $N$  grows, but eventually converges to the entropy of  $W$  that achieves  $C(X_1, \dots, X_N)$ .

#### APPENDIX

In this appendix, we give the proof of Theorem 1. First, as with [4], we define a quantity  $\Gamma(\delta_1, \delta_2)$  which plays an important role in the proof.

Let  $(X_1, X_2, \dots, X_N) \sim P(x_1, x_2, \dots, x_N)$  where  $X_1, \dots, X_N$  take values in finite alphabet  $\mathcal{X}_1, \dots, \mathcal{X}_N$ . Let  $(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N, W)$  be a  $(N+1)$ -tuple of random variables where  $\hat{X}_1 \in \mathcal{X}_1, \hat{X}_2 \in \mathcal{X}_2, \dots, \hat{X}_N \in \mathcal{X}_N$  and  $W \in \mathcal{W}$ , a finite set. Denote the marginal distribution of  $(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N)$  by

$$Q(x_1, x_2, \dots, x_N) = Pr(\hat{X}_1 = x_1, \hat{X}_2 = x_2, \dots, \hat{X}_N = x_N), \quad (30)$$

for  $x_i \in \mathcal{X}_i, i = 1, 2, \dots, N$ .

For any  $\delta_1, \delta_2 \geq 0$ , define

$$\Gamma(\delta_1, \delta_2) = \sup H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N|W), \quad (31)$$

where the supremum is taken over all  $(N+1)$ -tuples  $(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N, W)$  that satisfy

$$D(P; Q) = \sum_{x,y} P(x_1, x_2, \dots, x_N) \log \frac{P(x_1, x_2, \dots, x_N)}{Q(x_1, x_2, \dots, x_N)} \leq \delta_1, \quad (32)$$

and

$$\sum_{i=1}^N H(\hat{X}_i|W) - H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N|W) \leq \delta_2. \quad (33)$$

It follows that  $C(X_1, X_2, \dots, X_N)$  as defined in Theorem 1, is equivalent to

$$C(X_1, X_2, \dots, X_N) = H(X_1, X_2, \dots, X_N) - \Gamma(0, 0). \quad (34)$$

The following lemma gives some properties of  $\Gamma(\delta_1, \delta_2)$ .

*Lemma 3:* 1) For all  $\delta_1, \delta_2 \geq 0$ , there exists a  $(N+1)$ -tuple  $(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N, W)$  such that (32) and (33) are satisfied and

$$\Gamma(\delta_1, \delta_2) = H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N|W). \quad (35)$$

Moreover, for  $\delta_1, \delta_2 = 0$ ,

$$|\mathcal{W}| \leq \prod_{i=1}^N |\mathcal{X}_i|. \quad (36)$$

2)  $\Gamma(\delta_1, \delta_2)$  is a concave function of  $(\delta_1, \delta_2)$  and it is continuous for all  $\delta_1, \delta_2 \geq 0$ .

3) For  $\delta \geq 0$ , define  $\Gamma_1(\delta) = \Gamma(0, \delta)$  and  $\Gamma_2(\delta) = \Gamma(\delta, 0)$ , then  $\Gamma_1(\delta)$  and  $\Gamma_2(\delta)$  are concave and continuous for  $\delta \geq 0$ .

The proof of Lemma 1 follows similarly as the proof of Theorem 4.4 in [4].

A. *Proof of  $C_1 = C(X_1, X_2, \dots, X_N)$ .*

In this section, we prove the first part of Theorem 1, that is  $C_1 = C(X_1, X_2, \dots, X_N)$ . We first prove the converse part, that is for any  $R_0$  that is achievable for the Gray-Wyner source coding network, we have,

*Theorem 2 (Converse):*

$$C_1 \geq C(X_1, X_2, \dots, X_N). \quad (37)$$

To prove the converse, first let  $(f, g_i), i = 1, 2, \dots, N$  be any  $(n, \mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_N)$  code that satisfies (12), (13) and (14).

Then, we have,

$$\log \mathcal{M}_0 \geq H(M_0), \quad (38)$$

$$\geq I(X_1^n X_2^n \dots X_N^n; M_0), \quad (39)$$

$$= H(X_1^n X_2^n \dots X_N^n) - H(X_1^n X_2^n \dots X_N^n | M_0), \quad (40)$$

$$= nH(X_1 X_2 \dots X_N) - \sum_{j=1}^n H(X_{1j} X_{2j} \dots X_{Nj} | W_j), \quad (41)$$

where  $W_j \triangleq (M_0, X_1^{j-1}, X_2^{j-1}, \dots, X_N^{j-1})$  and  $X_i^{j-1} = (X_{i1}, X_{i2}, \dots, X_{i,j-1})$  for  $i = 1, 2, \dots, N$ .

Notice that, the  $(N+1)$ -tuple  $(X_{1j}, X_{2j}, \dots, X_{Nj}, W_j)$  satisfies condition (32) and (33) with  $\delta_1 = 0$  and

$$\delta_2^{(j)} = \sum_{i=1}^N H(X_{i,j} | W_j) - H(X_{1j}, X_{2j}, \dots, X_{Nj} | W_j). \quad (42)$$

Hence, by the definition of  $\Gamma(\delta_1, \delta_2)$ , we have

$$H(X_{1j} X_{2j} \dots X_{Nj} | W_j) \leq \Gamma_1(\delta_2^{(j)}). \quad (43)$$

Substitute (43) into (41), we get,

$$\log \mathcal{M}_0 \geq nH(X_1 X_2 \dots X_N) - \sum_{j=1}^n \Gamma_1(\delta_2^{(j)}), \quad (44)$$

$$\geq nH(X_1 X_2 \dots X_N) - n\Gamma_1\left(\frac{1}{n} \sum_{j=1}^n \delta_2^{(j)}\right). \quad (45)$$

where the last step is from the concavity of  $\Gamma_1(\cdot)$  function. Now define

$$\eta = \frac{1}{n} \sum_{j=1}^n \delta_2^{(j)}. \quad (46)$$

The following lemma gives an upper bound on  $\eta$ .

*Lemma 4:* For any  $(n, \mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_N)$  code that satisfies (12), (13) and (14), we have

$$\eta \leq (N+1)\epsilon. \quad (47)$$

*Proof:*

By Fano's inequality, we have, for  $i = 1, 2, \dots, N$ ,

$$H(X_i^n | M_0 M_i) \leq n\epsilon. \quad (48)$$

Hence, we have, for  $i = 1, 2, \dots, N$ ,

$$\log \mathcal{M}_i \geq H(M_i), \quad (49)$$

$$\geq H(M_i | M_0), \quad (50)$$

$$= H(X_i^n M_i | M_0) - H(X_i^n | M_i M_0), \quad (51)$$

$$\geq H(X_i^n M_i | M_0) - n\epsilon, \quad (52)$$

$$= H(X^n | M_0) - n\epsilon. \quad (53)$$

Then, we get,

$$\sum_{i=1}^N \log \mathcal{M}_i \geq \sum_{i=1}^N H(X_i^n | M_0) - n\epsilon'. \quad (54)$$

where  $\epsilon' = N\epsilon$ . Together with (41), we get,

$$\begin{aligned} \sum_{i=0}^N \log \mathcal{M}_i &\geq nH(X_1 X_2 \cdots X_N) \\ &\quad - \sum_{j=1}^n H(X_{1j} X_{2j} \cdots X_{Nj} | W_j) \\ &\quad + \sum_{i=1}^N H(X_i^n | M_0) - n\epsilon'. \end{aligned} \quad (55)$$

Together with (14), we get,

$$\sum_{i=1}^N H(X_i^n | M_0) - \sum_{j=1}^n H(X_{1j} X_{2j} \cdots X_{Nj} | W_j) \leq n\epsilon''. \quad (56)$$

where  $\epsilon'' = (N+1)\epsilon$ . On the other hand, we have,

$$\begin{aligned} &\sum_{i=1}^N H(X_i^n | M_0) \\ &= \sum_{i=1}^N \sum_{j=1}^n H(X_{ij} | X_i^{j-1} M_0), \end{aligned} \quad (57)$$

$$\geq \sum_{i=1}^N \sum_{j=1}^n H(X_{ij} | X_1^{j-1}, X_2^{j-1}, \dots, X_N^{j-1}, M_0), \quad (58)$$

$$= \sum_{i=1}^N \sum_{j=1}^n H(X_{ij} | W_j). \quad (59)$$

Combine (56) and (59), we have,

$$\sum_{j=1}^n \left[ \sum_{i=1}^N H(X_{ij} | W_j) - H(X_{1j} X_{2j} \cdots X_{Nj} | W_j) \right] \leq n\epsilon''. \quad (60)$$

Hence, we have,

$$\frac{1}{n} \sum_{j=1}^n \delta_2^{(j)} \leq \epsilon''. \quad (61)$$

This completes the proof of Lemma 4.  $\square$

Now, from Lemma 4 and (45), we get,

$$R_0 \geq \frac{1}{n} \log \mathcal{M}_0 \geq H(X_1, X_2, \dots, X_N) - \Gamma_1(\eta). \quad (62)$$

Together with the continuity of  $\Gamma_1(\cdot)$ , we have, as  $n \rightarrow \infty$ ,

$$R_0 \geq H(X_1, X_2, \dots, X_N) - \Gamma_1(0), \quad (63)$$

$$= C(X_1, X_2, \dots, X_N). \quad (64)$$

This completes the proof of converse part.  $\square$

We now prove the achievability part, that is, let the joint distribution  $P(x_1, x_2, \dots, x_N)$  be given, we have,

*Theorem 3 (Achievability):*

$$C_1 \leq C(X_1, X_2, \dots, X_N). \quad (65)$$

Our proof mainly involves generalizing Gray-Wyner source coding network [14] to that of  $N$  sources. The system model we considered here is the same as Fig. 1 described in section II except that definition 2 is replaced by,

*Definition 5:* A rate tuple  $(R_0, R_1, \dots, R_N)$  is said to be *achievable* if for all  $\epsilon > 0$ , we can find an  $n$  sufficiently large such that there exists a  $(n, 2^{nR_0}, 2^{nR_1}, \dots, 2^{nR_N})$  code with

$$P_e^{(n)} \leq \epsilon. \quad (66)$$

Our purpose is to find all achievable rate tuples  $(R_0, R_1, \dots, R_N)$ . The rate region of this source coding problem is summarized in the following theorem.

*Theorem 4:* For the source coding model described above, a rate tuple  $(R_0, R_1, \dots, R_N)$  is achievable if and only if the following conditions are satisfied,

$$R_0 \geq I(X_1, X_2, \dots, X_N; W), \quad (67)$$

$$R_i \geq H(X_i | W), \quad (68)$$

for  $i = 1, 2, \dots, N$ , and for some  $W \sim P(w | x_1, x_2, \dots, x_N)$ , where  $W \in \mathcal{W}$  and  $|\mathcal{W}| \leq \prod_{i=1}^N |\mathcal{X}_i| + 2$ .

*Proof of Theorem 4 (Sketch):*

For the achievability part, we want to show that for any rate tuple  $(R_0, R_1, \dots, R_N)$  that satisfies above conditions, we can construct a  $(n, 2^{nR_0}, 2^{nR_1}, \dots, 2^{nR_N})$  code such that the decoding error  $P_e^{(n)} \rightarrow 0$  as codeword length  $n \rightarrow \infty$ .

*Codeword Generation:* for any given distributions  $P(x_1, x_2, \dots, x_N)$  and  $P(w | x_1, x_2, \dots, x_N)$ , we calculate the marginal distribution  $P(w)$ .

1) Codebook  $C_0$ : we first randomly generate  $2^{nR_0}$  sequences  $w^n$  i.i.d.  $\sim P(w)$ , and index them by  $m_0 \in \{1, 2, \dots, 2^{nR_0}\}$ .



- 2) Codebook  $C(X_i)$ : for each  $i = 1, 2, \dots, N$ , for each  $x_i^n \in \mathcal{X}_i^n$ , randomly put them into  $2^{nR_i}$  bins and index them bins by  $m_i \in \{1, 2, \dots, 2^{nR_i}\}$ .

*Encoding:*

- 1) for each source sequences  $(x_1^n, x_2^n, \dots, x_N^n)$ , encoder  $f_0$  finds a  $w^n(m_0) \in C_0$  such that  $(x_1^n, x_2^n, \dots, x_N^n, w^n(m_0)) \in T_\epsilon^n$ , where  $T_\epsilon^n$  is the jointly typical set as defined in [15], and send the index  $m_0$  to the decoder. If there is no more than one  $w^n$ , choose the sequence  $w^n$  with the smallest index; if there exist no such sequence, choose sequence  $w^n(1)$ ,
- 2) for  $i = 1, 2, \dots, N$ , encoder  $f_i$  sends the bin index  $m_i$  of sequence  $x_i^n$ .

*Decoding:* for  $i = 1, 2, \dots, N$ , decoder  $i$  looks at bin  $m_i$  for codebook  $C(X_i)$  and finds the sequence  $\hat{x}_i^n$  such that  $(\hat{x}_i^n, w^n(m_0)) \in T_\epsilon^n$ . If there is more than one or none such sequence, declare an error.

*Error analysis:* Assuming  $m_i, i = 0, 1, \dots, N$  are the chosen indices for encoding  $(x_1^n, x_2^n, \dots, x_N^n)$ . There are three error events.

- 1)  $E_1: (x_1^n, x_2^n, \dots, x_N^n, w^n(m_0)) \notin T_\epsilon^n$  for all  $m_0 \in \{1, 2, \dots, 2^{nR_0}\}$ .
- 2)  $E_2: (x_i^n, w^n(m_0)) \notin T_\epsilon^n$  for each  $i$ .
- 3)  $E_3: \text{for some } i, \text{ there exists } \tilde{x}_i^n \neq x_i^n \text{ in bin } m_i \text{ of codebook } C(X_i) \text{ such that } (\tilde{x}_i^n, w^n(m_0)) \in T_\epsilon^n$ .

Hence,

$$P_e^{(n)} \leq P(E_1) + P(E_2|E_1^c) + P(E_3|E_1^c, E_2^c). \quad (69)$$

By some standard argument, we can get, as  $n \rightarrow \infty$ ,

- 1)  $P(E_1) \rightarrow 0$  if

$$R_0 \geq I(X_1, X_2, \dots, X_N; W) + \epsilon, \quad (70)$$

- 2)  $P(E_2|E_1^c) \rightarrow 0$ ,
- 3)  $P(E_3|E_1^c, E_2^c) \rightarrow 0$  if for each  $i = 1, 2, \dots, N$ ,

$$R_i \geq H(X_i|W) + \epsilon. \quad (71)$$

This completes the achievability proof.

For the converse part, we want to show that for any achievable rate tuple  $(R_0, R_1, \dots, R_N)$ , it should satisfy (67) and (68).

By Fano's inequality, we have

$$H(X_i^n|M_i M_0) \leq n\epsilon. \quad (72)$$

Hence, we have, for  $i = 1, 2, \dots, N$

$$\begin{aligned} & nR_i \\ & \geq H(M_i), \end{aligned} \quad (73)$$

$$\geq H(M_i|M_0), \quad (74)$$

$$\geq H(M_i|M_0) + H(X_i^n|M_i M_0) - n\epsilon, \quad (75)$$

$$= H(X_i^n M_i|M_0) - n\epsilon, \quad (76)$$

$$= H(X_i^n|M_0) - n\epsilon, \quad (77)$$

$$= \sum_{j=1}^n H(X_{ij}|M_0 X_i^{j-1}) - n\epsilon, \quad (78)$$

$$\geq \sum_{j=1}^n H(X_{ij}|M_0, X_1^{j-1}, X_2^{j-1}, \dots, X_N^{j-1}) - n\epsilon \quad (79)$$

and

$$\begin{aligned} & nR_0 \\ & \geq H(M_0), \end{aligned} \quad (80)$$

$$\geq I(M_0; X_1^n, X_2^n, \dots, X_N^n), \quad (81)$$

$$= \sum_{j=1}^n I(M_0; X_{1j} X_{2j} \dots X_{Nj} | X_1^{j-1} X_2^{j-1} \dots X_N^{j-1}) \quad (82)$$

$$= \sum_{j=1}^n I(M_0 X_1^{j-1} X_2^{j-1} \dots X_N^{j-1}; X_{1j} X_{2j} \dots X_{Nj}) \quad (83)$$

Define  $W_j = (M_0, X_1^{j-1}, X_2^{j-1}, \dots, X_N^{j-1})$ , and using a standard time sharing argument, we can get, for  $i = 1, 2, \dots, N$ ,

$$R_i \geq H(X_i|W) - \epsilon, \quad (84)$$

$$R_0 \geq I(X_1 X_2 \dots X_N; W). \quad (85)$$

Let  $n \rightarrow \infty$ , then  $\epsilon \rightarrow 0$ , and this completes the proof of converse. The cardinality bound can be obtained using the technique introduced in [16, Appendix C]. We skip the details. This completes the proof of Theorem 4.  $\square$

Now we proceed to prove Theorem 3. We will show that if  $R_0 > C(X_1, X_2, \dots, X_N)$ , it is achievable for Model I.

Let  $R_0 > C(X_1, X_2, \dots, X_N)$  and any  $\epsilon > 0$  be given and let random variables  $(X_1, X_2, \dots, X_N, W)$  satisfy (7) and (8), such that

$$C(X_1, X_2, \dots, X_N) = I(X_1 X_2 \dots X_N; W). \quad (86)$$

Notice that, the existence of such random variables is guaranteed by Lemma 3. Now define

$$\epsilon_1 = \min\left\{\frac{\epsilon}{N+1}, R_0 - C(X_1, X_2, \dots, X_N)\right\}, \quad (87)$$

and hence  $\epsilon_1 > 0$ . By Theorem 4, there exists a  $(n, \mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_N)$  code with  $P_e^{(n)} \leq \epsilon'$  and  $\epsilon' \leq \epsilon_1$ . Hence,

$$\frac{1}{n} \log \mathcal{M}_0 \leq C(X_1, X_2, \dots, X_N) + \epsilon_1 \leq R_0, \quad (88)$$

$$\frac{1}{n} \log \mathcal{M}_i \leq H(X_i|W) + \epsilon_1. \quad (89)$$

Hence, we have,

$$\sum_{i=0}^N \frac{1}{n} \log \mathcal{M}_i \leq C(X_1, X_2, \dots, X_N) + \sum_{i=1}^N H(X_i|W) + \epsilon, \quad (90)$$

$$\stackrel{(a)}{=} H(X_1, X_2, \dots, X_N) + \epsilon. \quad (91)$$

where (a) is from condition (8). Thus, condition (14) is also satisfied. This implies that  $R_0$  is achievable in Model I, which completes the proof of achievability part. This completes the proof of Theorem 3.  $\square$

**B. Proof of  $C_2 = C_{X_1, X_2, \dots, X_N}$ .**

In this section, we prove the second part of theorem 1, that is  $C_2 = C(X_1, X_2, \dots, X_N)$ . We have the following theorem.

*Theorem 5:*

$$C_2 \geq C(X_1, X_2, \dots, X_N), \quad (92)$$

$$C_2 \leq C(X_1, X_2, \dots, X_N). \quad (93)$$

For the converse part, that is (92), the proof follows almost the same line as in [4, Section 5.2]. For the achievability part, that is (93), the proof follows similarly as in [4, Section 6.2] by applying  $\mathcal{U} = \mathcal{X}_1 \times \mathcal{X}_2, \dots \times \mathcal{X}_N$  in [4, Theorem 6.3]. We omit the details here.

## REFERENCES

- [1] P. Gács and J. Körner, "Common information is much less than mutual information", *Problems Contr. Inform. Theory*, vol. 2, pp. 149-162, 1973.
- [2] R. Ahlswede and J. Körner, "On common information and related characteristics of correlated information sources", *Proc. of the 7th Prague Conference of Information Theory*, 1974.
- [3] H. S. Witsenhausen, "On sequences of pairs of dependent random variables", *SIAM J. Appl. Math.*, vol. 28, pp. 100-113, Jan. 1975.
- [4] A. D. Wyner, "The common information of two dependent random variables", *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163-179, March. 1975.
- [5] H. S. Witsenhausen, "Values and bounds for the common information of two discrete random variables", *SIAM J. Appl. Math.*, vol. 31, no. 2, pp. 313-333, 1976.
- [6] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.
- [7] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography, Part I: Secret sharing", *IEEE Trans. Inf. Theory*, vol. 39, pp. 1121-1132, July 1993.
- [8] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography, Part I: CR capacity", *IEEE Trans. Inf. Theory*, vol. 44, pp. 225-240, Jan. 1998.
- [9] U. M. Maurer, "Secret key agreement by public discussion from common information", *IEEE Trans. Inf. Theory*, vol. 39, pp. 733-742, May 1993.
- [10] G. Hu "On the amount of information", *Teor. Veroyatnost. i Primenen.*, 4: 447-455, 1962 (in Russian).
- [11] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic, New York, 1981.
- [12] R. W. Yeung, *Information Theory and Network Coding*, Springer, 2008.
- [13] H. Tyagi and P. Narayan and P. Gupta, "When is a Function Securely Computable? ", <http://arxiv.org/abs/1007.2945>, 2010.
- [14] R. M. Gray and A. D. Wyner, "Source coding for a simple network", *Bell Syst. Tech. J.*, vol. 58, pp. 1681-1721, Nov. 1974.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
- [16] A. El Gamal and Y. H. Kim, *Lecture notes on network information theory*, <http://arxiv.org/abs/1001.3404>, 2010.