Exact Common Information

Gowtham Ramani Kumar
Electrical Engineering
Stanford University
Email: gowthamr@stanford.edu

Cheuk Ting Li
Electrical Engineering
Stanford University
Email: ctli@stanford.edu

Abbas El Gamal
Electrical Engineering
Stanford University
Email: abbas@stanford.edu

***Abstract*—This paper introduces the notion of exact common information, which is the minimum description length of the common randomness needed for the exact distributed generation of two correlated random variables (X, Y). We introduce the quantity $G(X;Y) = \min_{X \to W \to Y} H(W)$ as a natural bound on the exact common information and study its properties and computation. We then introduce the exact common information rate, which is the minimum description rate of the common randomness for the exact generation of a 2-DMS (X, Y). We give a multiletter characterization for it as the limit $\overline{G}(X;Y) = \lim_{n\to\infty}(1/n)G(X^n;Y^n)$. While in general $\overline{G}(X;Y)$ is greater than or equal to the Wyner common information, we show that they are equal for the Symmetric Binary Erasure Source. We do not know, however, if the exact common information rate has a single letter characterization in general.**

I. Introduction

What is the common information between two correlated random variables or sources? This is a fundamental question in information theory with applications ranging from distributed generation of correlated sources [1] and secret keys [2] to joint source channel coding [3], among others. One of the most studied notions of common information is due to Wyner [1]. Let $(\mathcal{X}\times\mathcal{Y}, p(x,y))$ be a 2-DMS (or correlated sources (X, Y) in short). The Wyner common information $J(X;Y)$ between the sources X and Y is the minimum common randomness rate needed to generate (X, Y) with asymptotically vanishing total variation. Wyner established the single-letter characterization

$$J(X;Y) = \min_{W:\, X\to W\to Y} I(W;X,Y).$$

In this paper we introduce the notion of *exact common information*, which is closely related in its operational definition to the Wyner common information. While the Wyner setup assumes block codes and *approximate* generation of the 2-DMS (X, Y), our setting assumes variable length codes and *exact* generation of (X, Y). As such, the relationship between our setup and Wyner's is akin to that between the zero-error and the lossless source coding problems. In the source coding problem the entropy of the source is the limit on both the zero-error and the lossless compression. Is the limit on the exact common information rate the same as the Wyner common information? We show that they are the same for the Symmetric Binary Erasure Source (SBES) as defined in Section II. We do not, however, know if they are equal in general.

The rest of this paper is organized as follows. In the next section we introduce the exact distributed generation problem and define the exact common information. We introduce the "common-entropy" quantity $G(X;Y) = \min_{X\to W\to Y} H(W)$ as a natural bound on the exact common information and study some of its properties. In Section III, we define the exact common information rate for a 2-DMS. We show that it is equal to the limit $\overline{G}(X;Y) = \lim_{n\to\infty}(1/n)G(X^n;Y^n)$ and that it is in general greater than or equal to the Wyner common information. One of the main results in this paper is to show that $\overline{G}(X;Y) = J(X;Y)$ for the SBES. A consequence of this result is that the quantity $G(X^k;Y^k)$ can be strictly smaller than $kG(X;Y)$, that is, the per-letter common entropy can be reduced by increasing the dimension. We then introduce the notion of approximate common information rate, which relaxes the condition of exact generation to asymptotically vanishing total variation and show that it is equal to the Wyner common information. As computing the quantity $G(X;Y)$ involves solving a non-convex optimization problem, in Section IV we present cardinality bounds on W and use them to find an explicit expression for $G(X;Y)$ when X and Y are binary. Due to space limitation, we do not include many of the proofs. A complete version of this paper is posted on arXiv.

II. Definitions and Properties

Consider the distributed generation setup depicted in Figure 1. Alice and Bob both have access to common randomness W. Alice uses W and her own local randomness to generate X and Bob uses W and his own local randomness to generate Y such that $(X, Y) \sim p_{X,Y}(x, y)$. We wish to find the limit on the least amount of common randomness needed to generate (X, Y) exactly.

More formally, we define a *simulation code* (W, R) for this setup to consist of

- A common random variable $W \sim p_W(w)$. As a measure of the amount of common randomness, we use the per-letter *minimum expected codeword length* R over the set of all variable length *prefix-free* zero-error binary codes $\mathcal{C} \subset \{0,1\}^*$ for W, i.e., $R = \min_{\mathcal{C}} \mathsf{E}(L)$, where L is the codeword length of the code $\mathcal{C}$ for W.
- A stochastic decoder $p_{\hat{X}|W}(x|w)$ for Alice and a stochastic decoder $p_{\hat{Y}|W}(y|w)$ for Bob such that $\hat{X}$ and $\hat{Y}$ are conditionally independent given W.

This work was partially supported by Air Force grant FA9550-10-1-0124. Cheuk Ting Li is supported by a Hong Kong Alumni Stanford Graduate Fellowship.

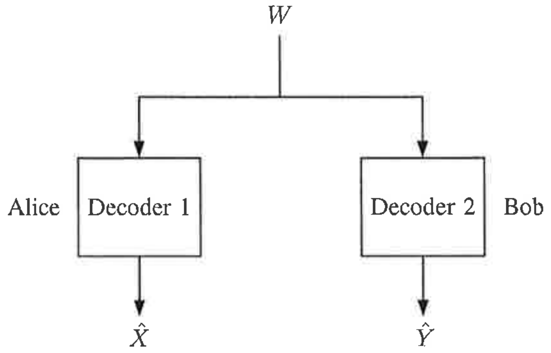

Fig. 1. Setting for distributed generation of correlated random variables. For exact generation, $(\hat{X}, \hat{Y}) \sim p_{X,Y}(x, y)$.

The random variable pair (X, Y) is said to be exactly generated by the simulation code (W, R) if $p_{\hat{X},\hat{Y}}(x, y) = p_{X,Y}(x, y)$. We wish to find the *exact common information* R^* between the sources X and Y, which is the infimum over all rates R such that the random variable pair (X, Y) can be exactly generated.

Remark: The exact common information R^* (and the exact common information rate defined in the next section) can be also defined through a "zero error" version of the Gray-Wyner system [4]. This approach, however, is neither operationally better motivated than the above setup nor yields better insights or results. Hence, we will not pursue this alternative setup any further.

Define the following quantity, which can be interpreted as the "common entropy" between X and Y,

$$G(X;Y) = \min_{W:\, X \to W \to Y} H(W). \qquad (1)$$

Remark: We can use min instead of inf in the definition of $G(X;Y)$ because the cardinality of W is bounded as we will see in Proposition 5, hence the optimization for computing $G(X;Y)$ is over a closed set.

Following the proof of Shannon's zero-error compression theorem, we can readily show the following.

Proposition 1.

$$G(X;Y) \leq R^* < G(X;Y) + 1.$$

Computing $G(X;Y)$ is in general quite difficult (see Section IV). In some special cases, we can find an explicit expression for it.

Example 1 The Symmetric Binary Erasure Source (X, Y) with parameter p (SBES(p)) is defined by

$$X \sim \mathrm{Bern}(1/2),$$
$$Y = \begin{cases} X & \text{w.p. } 1-p, \\ e & \text{w.p. } p, \end{cases}$$

where p is the *erasure probability* for the source. It can be shown that for the SBES(p),

$$G(X;Y) = \min\{1, H(p) + 1 - p\}.$$

Note that the Wyner common information for this source is [5]

$$J(X;Y) = \begin{cases} 1 & \text{if } p \leq 0.5, \\ H(p) & \text{if } p > 0.5. \end{cases}$$

In the following we present some basic properties of $G(X;Y)$.

A. Properties of $G(X;Y)$

1) $G(X;Y) \geq 0$ with equality if and only if X and Y are independent.
2) $G(X;Y) \geq J(X;Y)$.
3) *Data-processing Inequality:* If $U \to X \to Y$ forms a Markov chain, then $G(U;Y) \leq G(X;Y)$.
4) Define $G(X;Y|Z) = \sum_{z \in \mathcal{Z}} p_Z(z) G(X;Y|Z = z)$. Then $G(X;Y) \leq H(Z) + G(X;Y|Z)$.
5) If there exist functions $f(X)$ and $g(Y)$ such that $Z = f(X) = g(Y)$, then $G(X;Y) = H(Z) + G(X;Y|Z)$.
6) Let $T(X)$ be a sufficient statistic of X with respect to Y ([6], pg. 305). Then $G(X;Y) = G(T(X);Y)$. Further, if W achieves $G(X;Y)$, we have $H(W) \leq H(T(X))$. Thus a noisy description of X via W may potentially have a smaller entropy than the minimal sufficient statistic, which is a deterministic description.

III. Exact Common Information Rate

The distributed generation setup in Figure 1 can be readily extended to the n-letter setting in which Alice wishes to generate X^n from common randomness W_n and her local randomness and Bob wishes to generate Y^n from W_n and his local randomness such that $p_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) \sim \prod_{i=1}^n p_{X,Y}(x_i, y_i)$. We define a simulation code (W_n, R, n) for this setup in the same manner as for the one-shot case.

We say that Alice and Bob can exactly generate the 2-DMS (X, Y) at rate R if for some $n \geq 1$, there exists a (W_n, R, n) simulation code that exactly generates (X^n, Y^n) (since we assume prefix-free codes for W_n, we can simulate for arbitrarily large lengths via concatenation of successive codewords). We wish to find the *exact common information rate* R^* between the sources X and Y, which is the infimum over all rates R such that the 2-DMS (X, Y) can be exactly generated.

Define the "joint common entropy"

$$G(X^n;Y^n) = \min_{W_n:\, X^n \to W_n \to Y^n} H(W_n). \qquad (2)$$

It can be readily shown that $\lim_{n\to\infty}(1/n)G(X^n;Y^n) = \inf_{n\in\mathbb{N}}(1/n)G(X^n;Y^n)$. Hence, we can define the limiting quantity

$$\overline{G}(X;Y) = \lim_{n\to\infty} \frac{1}{n} G(X^n;Y^n).$$

We are now are ready to establish the following multiletter characterization for the exact common information rate.

Proposition 2 (Multiletter Characterization of R^*). *The exact common information rate between the components X and Y of a 2-DMS (X, Y) is*

$$R^* = \overline{G}(X;Y).$$

As expected the exact common information rate is greater than or equal to the Wyner common information.

Proposition 3.

$$\overline{G}(X;Y) \geq J(X;Y).$$

In the following section, we show that they are equal for the SBES in Example 1. We do not know if this is the case in general, however.

A. Exact Common Information of the SBES

We will need the following result regarding computing the Wyner common information for the SBES.

Lemma 1. *To compute $J(X;Y)$ for the SBES(p), it suffices to consider*

$$W = \begin{cases} X & w.p.\ 1-p_1, \\ e & w.p.\ p_1, \end{cases}$$

and

$$Y = \begin{cases} W & w.p.\ 1-p_2, \\ e & w.p.\ p_2, \end{cases}$$

for some p_1 and p_2 satisfying $p_1 + p_2 - p_1 p_2 = p$.

The proof follows by [5], Appendix A.

We now present the main result on exact common information rate in this paper.

Theorem 1. *If (X,Y) is an SBES, then $\overline{G}(X;Y) = J(X;Y)$.*

Proof: In general $\overline{G}(X;Y) \geq J(X;Y)$. We will now provide an achievability scheme to show that for SBES, $\overline{G}(X;Y) \leq J(X;Y)$.

Choose a W as defined in Lemma 1 and define

$$\tilde{W} = \begin{cases} d & \text{if } W \in \{0,1\}, \\ e & \text{if } W = e, \end{cases}$$

$$\tilde{Y} = \begin{cases} d & \text{if } Y \in \{0,1\}, \\ e & \text{if } Y = e. \end{cases}$$

Note that $\tilde{Y}^n$, denoting the location of the erasures, is i.i.d. Bern(p) (with $1 \leftarrow e$, $0 \leftarrow d$) and independent of X^n. Furthermore, Y^n is a function of X^n and $\tilde{Y}^n$.

Codebook Generation: Generate a codebook $\mathcal{C}$ consisting of $2^{n(I(\tilde{Y};\tilde{W})+\epsilon)}$ sequences $\tilde{w}^n(m)$, $m \in [1:2^{n(I(\tilde{Y};\tilde{W})+\epsilon)}]$, that "covers" almost all the $\tilde{y}^n$ sequences except for a subset of small probability $\delta(\epsilon)$. By the covering lemma ([7], page 62), such a codebook exists for large enough n.

This lets us associate every covered sequence $\tilde{y}^n$ with a unique $\tilde{w}^n = \tilde{w}^n(\tilde{y}^n) \in \mathcal{C}$ such that $(\tilde{y}^n, \tilde{w}^n) \in \mathcal{T}_\epsilon^{(n)}$.

Define the random variable

$$\tilde{W}_n = \begin{cases} \tilde{w}^n(\tilde{y}^n) & \text{if } \tilde{y}^n \text{ is covered by } \mathcal{C}, \\ \tilde{y}^n & \text{if } \tilde{y}^n \text{ is not covered.} \end{cases} \tag{3}$$

Note that $\tilde{W}_n$ is a function of $\tilde{Y}^n$ and that the set of erasure coordinates in $\tilde{W}_n$ is a subset of those in $\tilde{Y}^n$.

Channel Simulation Scheme:

1) The central node generates $\tilde{W}_n$ defined in (3) and sends it to both encoders.
2) Encoder 2 (Bob) generates $\tilde{Y}^n \sim p_{\tilde{Y}^n|\tilde{W}_n}(\tilde{y}^n|\tilde{w}^n)$
3) The central node generates and sends to both encoders a message M comprising i.i.d. Bern(1/2) bits for only those coordinates i of X^n where $\tilde{W}_n(i) = d$. Thus $H(M) \leq n(1-p_1+\delta(\epsilon))$.
4) Encoder 1 (Alice) generates the remaining bits of X^n not conveyed by M using local randomness. Then X^n is independent of $\tilde{W}_n, \tilde{Y}^n$ and is i.i.d. Bern(1/2).
5) Encoder 2 generates $Y^n = Y^n(\tilde{W}_n, X^n) = Y^n(\tilde{W}_n, M)$. He only needs the bits X_i such that $\tilde{Y}_i = d$, which are available via M.

To complete the proof, note that $X^n \to (\tilde{W}_n, M) \to Y^n$ forms a Markov chain. Therefore,

$$\begin{aligned} G(X^n;Y^n) &\leq H(\tilde{W}_n, M) + 1 \leq H(\tilde{W}_n) + H(M) + 1 \\ &\overset{(a)}{\leq} H(\delta(\epsilon)) + (1-\delta(\epsilon))H(\tilde{W}_n|\tilde{W}_n \in \mathcal{C}) \\ &\quad + \delta(\epsilon)H(\tilde{W}_n|\tilde{W}_n \notin \mathcal{C}) + n(1-p_1+\delta(\epsilon)) + 1 \\ &\overset{(b)}{\leq} H(\delta(\epsilon)) + (1-\delta(\epsilon))\log|\mathcal{C}| \\ &\qquad + \delta(\epsilon)\log|\tilde{\mathcal{Y}}^n| + n(1-p_1+\delta(\epsilon)) + 1 \\ &= n(I(\tilde{Y};\tilde{W}) + 1 - p_1 + \delta(\epsilon)) \\ &\overset{(c)}{=} n(I(W;X,Y) + \delta(\epsilon)), \end{aligned}$$

where (a) follows by the grouping lemma for entropy, since $\mathsf{P}\{\tilde{W}_n \notin \mathcal{C}\} = \mathsf{P}\{\tilde{Y}^n \text{ not covered}\} = \delta(\epsilon)$; (b) follows since entropy is upper bounded by log of the alphabet size; and (c) follows from the definition of mutual information and some algebraic manipulations.

If we let $n \to \infty$, we obtain $\overline{G}(X;Y) \leq I(W;X,Y) + \delta(\epsilon)$ for any $\epsilon > 0$. Minimizing $I(W;X,Y)$ over all W from Lemma 1 completes the proof. ■

Note that the single letter characterization of the Wyner common information for the 2-DMS $(X^k, Y^k) \sim \prod_{i=1}^k p_{X,Y}(x_i, y_i)$ is k times that of the 2-DMS $(X;Y)$, that is, $\min I(W;X^k,Y^k) = k \min I(W;X,Y)$. The same property holds for the Gács–Körner–Witsenhausen common information [8], and for mutual information. In the following we show that $G(X^k;Y^k)$ can be strictly smaller than $kG(X;Y)$. Hence, it is possible to realize gains in the "common entropy" when we increase the dimension.

By the fact that for the SBES(p), $\overline{G}(X;Y) = H(p)$ for $p > 1/2$ and $G(X;Y) = \min\{1, H(p) + 1 - p\}$, there exists a p such that $\overline{G}(X;Y) < G(X;Y)$. Hence, we can show by contradiction that there exists a 2-DMS (X,Y) such that $G(X^2;Y^2) < 2G(X;Y)$. We can also give an explicit example of a 2-DMS (X,Y) such that $G(X^2;Y^2) < 2G(X;Y)$. Let

$$p_{X,Y} = \begin{bmatrix} 1/3 & 1/3 \\ 1/3 & 0 \end{bmatrix}.$$

Then, by Proposition 8 in Section IV, we have $G(X;Y) = H(1/3)$, where $H(p)$, $0 \leq p \leq 1$, is the binary entropy function. Note that we can write

$$p_{X^2,Y^2} = \begin{bmatrix} 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 0 & 1/9 & 0 \\ 1/9 & 1/9 & 0 & 0 \\ 1/9 & 0 & 0 & 0 \end{bmatrix}$$

$$= \frac{4}{9}\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}^t + \frac{3}{9}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}^t$$

$$+ \frac{1}{9}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}^t + \frac{1}{9}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^t .$$

Let $W \sim p_W(w) = [4/9, 3/9, 1/9, 1/9]$, then

$$p_{X^2,Y^2}(x^2, y^2) = \sum_w p_W(w) p_{X^2|W}(x^2|w) p_{Y^2|W}(y^2|w),$$

that is, $X^2 \to W \to Y^2$ form a Markov chain. Thus,

$$\begin{aligned} G(X^2; Y^2) &\le H(W) \\ &< 2H(1/3) = 2G(X;Y). \end{aligned}$$

B. Approximate common information rate

Consider the approximate distributed generation setting in which Alice and Bob wish to generate 2-DMS (X, Y) with vanishing total variation

$$\lim_{n\to\infty} |p_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) - \prod_{i=1}^{n} p_{X,Y}(x_i, y_i)|_{\mathrm{TV}} = 0.$$

We define a (W_n, R, n)-simulation code for this setting in the same manner as for exact distributed generation. We define the *approximate common information rate* R^*_{TV} between the sources X and Y as the infimum over all rates R such that the 2-DMS (X, Y) can be approximately generated.

We can show that the approximate common information rate is equal to the Wyner common information.

Proposition 4.

$$R^*_{\mathrm{TV}} = J(X;Y).$$

Proof: Achievability: Achievability follows from Wyner's coding scheme [1]. Choose $W_n \sim \mathrm{Unif}[1 : 2^{nR}]$ and associate each $w_n \in \mathcal{W}_n$ with a codeword of fixed length $\ell(w_n) = \lceil nR \rceil$. Decoders 1 (Alice) and 2 (Bob) first decode W_n and then use Wyner's coding scheme to generate $\hat{X}^n, \hat{Y}^n$, respectively. Any rate $R > J(X;Y)$ is admissible and will guarantee the existence of a scheme such that $(\hat{X}^n, \hat{Y}^n)$ is close in total variation to (X^n, Y^n). Thus $R^*_{\mathrm{TV}} \le J(X;Y)$.
Converse: Suppose that for any $\epsilon > 0$, there exists a (W_n, R, n) simulation code that generates $(\hat{X}^n, \hat{Y}^n)$ whose pmf differs from that of (X^n, Y^n) by at most ϵ in total variation. Then we have

$$\begin{aligned} nR \ge H(W_n) &\ge I(\hat{X}^n, \hat{Y}^n; W_n) \\ &\quad - I(\hat{X}_q, \hat{Y}_q; \hat{X}^{q-1}, \hat{Y}^{q-1}) \\ &\overset{(a)}{\ge} \sum_{q=1}^{n} I(\hat{X}_q, \hat{Y}_q; W) - n\delta(\epsilon) \\ &= nI(\hat{X}_Q, \hat{Y}_Q; W, Q) - nI(\hat{X}_Q, \hat{Y}_Q; Q) - n\delta(\epsilon) \\ &\overset{(b)}{\ge} nI(\hat{X}_Q, \hat{Y}_Q; W, Q) - n\delta(\epsilon) \\ &\overset{(c)}{\ge} nJ(X;Y) - n\delta(\epsilon), \end{aligned}$$

where (a), (b) follow from Lemma 20 and Lemma 21 respectively in [9] since the pmf of $(\hat{X}^n, \hat{Y}^n)$ differs from that of (X^n, Y^n) by at most ϵ in total variation; and (c) follows from the continuity of $J(X;Y)$. ■

Remark: Note that if we replace the total variation constraint in Proposition 4 by the stronger condition

$$p_{X^n,Y^n}(x^n, y^n) = (1-\epsilon) p_{\hat{X}^n,\hat{Y}^n}(x^n, y^n) + \epsilon r(x^n, y^n) \quad (4)$$

for some pmf $r(x^n, y^n)$ over $\mathcal{X}^n \times \mathcal{Y}^n$, the required approximate common information rate R^*_{SD} becomes equal to the exact common information $\bar{G}(X;Y)$. To show this, note that $R^*_{SD} \le \bar{G}(X;Y)$ is trivial because the exact distributed generation constraint is stronger than (4).

To show $R^*_{\mathrm{SD}} \ge \bar{G}(X;Y)$, start with any (W_n, R, n) simulation code that generates $(\hat{X}^n, \hat{Y}^n)$ satisfying (4). Let

$$W'_n = \begin{cases} W_n & \text{w.p. } 1-\epsilon, \\ (\bar{X}^n, \bar{Y}^n) \sim r(x^n, y^n) & \text{w.p. } \epsilon. \end{cases}$$

We construct a (W'_n, R', n) code that generates (X^n, Y^n) exactly and satisfies $R' \le R + \delta(\epsilon)$. If the decoders receive $W'_n = W_n$, they follow the original achievability scheme to generate $(\hat{X}^n, \hat{Y}^n)$ satisfying (4). If $W'_n = (\bar{X}^n, \bar{Y}^n)$, then the decoders simply output $\bar{X}^n$ and $\bar{Y}^n$, respectively. Now,

$$\begin{aligned} H(W'_n) &\le H(\epsilon) + (1-\epsilon)H(W_n) + \epsilon \log |\mathcal{X}|^n |\mathcal{Y}|^n \\ &= H(W_n) + n\delta(\epsilon). \end{aligned}$$

Therefore, $R' \le (1/n)(H(W'_n) + 1) = R + \delta(\epsilon) + 1/n = R + \delta(\epsilon)$ for n large enough. Thus $R^*_{SD} \ge \bar{G}(X;Y)$.

IV. Computing $G(X;Y)$

The optimization problem for determining $G(X;Y)$ is in general quite difficult, involving the minimization of a concave function over a complex markovity constraint. In this section we provide some results on this optimization problem. We provide two bounds on the cardinality of W, establish two useful extremal lemmas, and use these results to analytically compute $G(X;Y)$ for binary alphabets. We then briefly discuss a connection to a problem in machine learning.

We first establish the following upper bound on cardinality.

Proposition 5. *To compute $G(X;Y)$, it suffices to consider W with cardinality $|\mathcal{W}| \le |\mathcal{X}||\mathcal{Y}|$.*

We now state an extremal lemma regarding the optimization problem for $G(X;Y)$ that will naturally lead to another cardinality bound.

Lemma 2. *Given $p_{X,Y}(x,y)$, let W attain $G(X;Y)$. Then for $w_1 \neq w_2$, the supports of $p_{Y|W}(\cdot|w_1)$ and $p_{Y|W}(\cdot|w_2)$ must be different.*

Lemma 2 yields the following cardinality bound.

Proposition 6. *To compute $G(X;Y)$ for a given pmf $p_{X,Y}(x,y)$, it suffices to consider W with cardinality $|\mathcal{W}| \leq 2^{\min(|\mathcal{X}|,|\mathcal{Y}|)} - 1$.*

The following shows that the bound in Proposition 6 is tight.

Example 2 Let (X,Y) be a SBES(0.1). Since $p_{X,Y}(0,1) = p_{X,Y}(1,0) = 0$, the markovity constraint $X \to W \to Y$ implies that the only W with $|\mathcal{W}| = 2$ is $W = X$; see [5], Appendix A. Hence, $G(X;Y) \leq H(X) = 1$. However, $H(Y) = H(0.1) + 0.1 < 1$. Thus, the optimal W^* that achieves $G(X;Y)$ requires $|\mathcal{W}^*| = 3$, making the bound in Proposition 6 tight.

The following is another extremal property of $G(X;Y)$.

Proposition 7. *Suppose W attains $G(X;Y)$. Consider a non-empty subset $\mathcal{W}' \subseteq \mathcal{W}$. Let (X',Y') be defined by the pmf*

$$p_{X',Y'}(x,y) = \sum_{w \in \mathcal{W}'} \frac{p_W(w)}{\sum_{w' \in \mathcal{W}'} p_W(w')} p_{X|W}(x|w) p_{Y|W}(y|w).$$

Then $H(X';Y') = H(W|W \in \mathcal{W}')$.

We now use the above results to analytically compute $G(X;Y)$ for binary alphabets, i.e., when $|\mathcal{X}| = |\mathcal{Y}| = 2$.

Proposition 8. *Let $X \sim \text{Bern}(p)$ and*

$$p_{Y|X} = \begin{bmatrix} \alpha & \beta \\ \bar{\alpha} & \bar{\beta} \end{bmatrix}$$

for some $\alpha, \beta \in [0,1], \bar{\alpha} = 1 - \alpha, \bar{\beta} = 1 - \beta$. Let W achieve $G(X;Y)$. Then either

$$p_{Y|W} = \begin{bmatrix} \alpha & 1 \\ \bar{\alpha} & 0 \end{bmatrix}, \; p_{W|X} = \begin{bmatrix} 1 & \bar{\beta}/\bar{\alpha} \\ 0 & 1 - \bar{\beta}/\bar{\alpha} \end{bmatrix}, \; \textit{and}$$
$$W \sim \text{Bern}\left(\bar{p}\left(1 - \bar{\beta}/\bar{\alpha}\right)\right),$$

or

$$p_{Y|W} = \begin{bmatrix} 0 & \beta \\ 1 & \bar{\beta} \end{bmatrix}, \; p_{W|X} = \begin{bmatrix} 1 - \alpha/\beta & 0 \\ \alpha/\beta & 1 \end{bmatrix}, \; \textit{and}$$
$$W \sim \text{Bern}\left(p(1 - \alpha/\beta)\right).$$

The proof of this proposition uses Lemma 2 as well as the cardinality bound $|\mathcal{W}| \leq 3$ derived from Proposition 6. It considers all possible cases for W and finally concludes that $|\mathcal{W}| = 2$ suffices.

Remark (Relationship to machine learning): Computing $G(X;Y)$ is closely related to *positive matrix factorization*, which has applications in recommendation systems, e.g., [10]. In that problem, one wishes to factorize a matrix M with positive entries in the form $M = AB$, where A and B are both matrices with positive entries. Indeed, finding a Markov chain $X \to W \to Y$ for a fixed $p_{X,Y}$ is akin to factorizing $p_{Y|X} = p_{Y|W} p_{W|X}$ and numerical methods such as in [11] can be used. Rather than minimizing the number of factors $|\mathcal{W}|$ as is done in positive matrix factorization literature, it may be more meaningful for recommendation systems to minimize the entropy of the factors W. Computing $G(X;Y)$ for large alphabets appears to be very difficult, however.

V. Conclusion

We introduced the notion of exact common information for correlated random variables (X,Y) and bounded it by the common entropy quantity $G(X;Y)$. For the exact generation of a 2-DMS, we established a multiletter characterization of the exact common information rate. While this multiletter characterization is in general greater than or equal to the Wyner common information, we showed that they are equal for the SBES. The main open question is whether the exact common information rate has a single letter characterization in general. Is it always equal to the Wyner common information? Is there an example 2-DMS for which the exact common information rate is strictly larger than the Wyner common information? It would also be interesting to further explore the application to machine learning.

We also remark that our setting and results can be readily extended to the coordination via communication problem [12]. In the arXiv version of this paper, we show that for the SBES, the set of achievable rates for exact coordination coincides with that for coordination under the total variation constraint.

VI. Acknowledgments

The authors are indebted to Young-Han Kim for comments that greatly improved the presentation of this paper.

References

[1] A. D. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.

[2] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptogra-phy—I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1121–1132, 1993.

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

[4] R. Gray and A. Wyner, "Source coding for a simple network," *Bell Systems Tech. Journal*, vol. 53, pp. 1681–1721, Nov. 1974.

[5] P. Cuff, "Distributed channel synthesis," *arXiv:1208.4415v3*, 2013.

[6] A. A. Borokov, "Mathematical statistics," *Gordon and Breach Science Publishers*, 1998.

[7] A. E. Gamal and Y. H. Kim, "Network information theory," *Cambridge University Press*, 2011.

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

[9] P. Cuff, "Communication in networks for coordinating behavior," *PhD Dissertation submitted to the dept. of Electrical Engg. at Stanford University*, 2009.

[10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer Society*, 2009.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. MIT Press.*, pp. 556–562, 2001.

[12] P. Cuff, H. Permuter, and T. Cover, "Coordination capacity," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, p. 4181 4206, September 2010.