Common Information Dimension

Osama A. Hanna[†*], Xinlin Li[†*], Suhas Diggavi[†] and Christina Fragouli[†]
[†]University of California, Los Angeles
Email:{ohanna, xinlinli, suhasdiggavi, christina.fragouli}@ucla.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206887

***Abstract*—The exact common information between a set of random variables $X_1, ..., X_n$ is defined as the minimum entropy of a shared random variable that allows for the exact distributive simulation of $X_1, ..., X_n$. It has been established that, in certain instances, infinite entropy is required to achieve distributive simulation, suggesting that continuous random variables may be needed in such scenarios. However, to date, there is no established metric to characterize such cases. In this paper, we propose the concept of Common Information Dimension (CID) with respect to a given class of functions $\mathcal{F}$, defined as the minimum dimension of a random variable W required to distributively simulate a set of random variables $X_1, ..., X_n$, such that W can be expressed as a function of $X_1, \cdots, X_n$ using a member of $\mathcal{F}$. Our main contributions include the computation of the common information dimension for jointly Gaussian random vectors in a closed form, with $\mathcal{F}$ being the linear functions class.**

I. Introduction

The common randomness between dependent random variables is a fundamental problem in information theory [1]–[4] and has ubiquitous applications in a number of areas, such as key generation in cryptography [5]–[8], hypothesis testing in statistical inference [9]–[12], and multi-modal representation learning in machine learning [13]–[17]. Multiple information theoretical notions have been developed to measure the common randomness, for instance, Gács-Körner's common information [2], Wyner's common information [3], common entropy [18], and coordination capacity [19] (see also the monograph [12]); but as far as we know, all of them measure common information in terms of bits. In this paper, we introduce the notion of common information dimension, that uses dimensionality instead of bits to characterize common randomness for continuous random variables.

Gács-Körner's common information [2] and Wyner's common information [3] are perhaps the most classical notions of common information. Gács-Körner's common information is defined as the maximum number of bits per symbol of the information that can be individually extracted from two dependent discrete variables X, Y, namely

$$C_{\text{GK}}(X,Y) := \max_{f,g:f(X)=g(Y)} H(f(X)), \tag{1}$$

where f and g are deterministic functions. However, it is known from [2], [20] that $C_{\text{GK}}(X,Y)$ equals zero except for one special case where $X = (X', V)$, $Y = (Y', V)$ and X', Y', V are independent variables.

Wyner's common information was originally defined for a pair of discrete sources $(X,Y) \sim \pi_{XY}$ as

$$C_{\text{Wyner}}(X,Y) := \min_{P_W P_{X|W} P_{Y|W} : P_{XY} = \pi_{XY}} I(X\,Y; W). \tag{2}$$

Wyner [3] provided two operational interpretations. One is for the *source coding* problem: the minimum common rate for the lossless source coding problem over the Gray-Wyner network, subject to a sum rate constraint. The other is for the *distributed simulation* problem: the minimum amount of shared randomness to simulate a given joint distribution π_{XY}. Recently, the works in [21] and [19] generalized Wyner's common information to n discrete random variables settings; and the works [22], [23] and [24], [25] generalized its interpretations to continuous sources in lossy source coding and distributed simulation, respectively.

Wyner's distributed simulation assumes codes of large block length (i.e., multi-shot) and approximate generation: the relative entropy between the generated distribution and the target distribution goes to zero as the block length goes to infinity. Kumar, Li and El Gamal [18] extended Wyner's work to define the exact common information (also called common entropy) which requires a single-shot (i.e., block length 1) and exact generation of π_{XY}. The common entropy is defined as

$$G(X,Y) := \min_{P_W P_{X|W} P_{Y|W} : P_{XY} = \pi_{XY}} H(W). \tag{3}$$

To generalize this to the multi-shot (asymptotic) setting, they also defined the exact common information rate as

$$C_{\text{Exact}(X,Y)} := \lim_{n\to\infty} \frac{1}{n} G(X^n, Y^n). \tag{4}$$

The exact common information was extended to n continuous variables in [24], and was shown to provide an upper bound on Wyner's common information in [18].

Why do we need a new notion of common randomness? All of the above discussed notions of common randomness are expressed in terms of "bits", and thus they are only meaningful, for example in distributed simulation, when finite common randomness is sufficient. Nevertheless, in the general case of continuous variables, an infinite amount of randomness may be required. For instance, [23] calculated the Wyner's common information of bivariate Gaussian random variables X, Y as $C_{\text{Wyner}}(X,Y) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$, where ρ is the correlation coefficient, and $C_{\text{Wyner}}(X,Y)$ goes to infinity as ρ goes to 1.

∗ Equal contribution.

This work was supported in part by the NSF grants 2221871, 2139304 and 2007714, and the Army Research Laboratory grant under Cooperative Agreement W911NF-17-2-0196.

To the best of our knowledge, there is no proposed metric that distinguishes between different problem complexities when an infinite amount of randomness is needed.

Our observation is that, for the latter case, a real-valued random variable W may be needed to represent the common randomness, and its dimension could provide guidance for practical applications. This is akin to compressed sensing [26], [27], where we seek a low dimensional representation in a high dimensional space. Note that without imposing any structural assumptions, the minimum dimension does not exceed one. This is due to the existence of measurable bijections between $\mathbb{R}$ and $\mathbb{R}^d$ for any $d \geq 1$. However, these functions are not implementable and unstable, as noted in [28], and hence, are not useful for applications. Therefore, regularity constraints on the common variable need to be considered. In particular, in our definition, we allow for a restriction of the form $W = g(X_1, \cdots, X_n)$ for some $g \in \mathcal{F}$, where $\mathcal{F}$ is a given class of functions. For example, $\mathcal{F}$ may be chosen to be the set of linear, or smooth functions. If $\mathcal{F}$ is chosen to contain all possible functions, then the minimum dimension of W will be upper bounded by one, as previously explained.

In our work, we consider two notions of dimension: the number of elements of a vector and the information dimension (also called Rényi dimension). The Rényi dimension is an information measure for random vectors in Euclidean space that was proposed by Rényi in 1959 [29]. It characterizes the growth rate of the entropy of successively finer discretizations of random variables. The Rényi dimension of a random vector $W \in \mathbb{R}^{d_W}$ is defined as (when the limit exists)

$$d^R(W) = \lim_{m\to\infty} \frac{H(\langle W\rangle_m)}{\log m}, \tag{5}$$

where $\langle W\rangle_m$ is the element-wise discretization of W defined as $\langle W^{(i)}\rangle_m = \frac{\lfloor mW^{(i)}\rfloor}{m}$ and $H(V)$ is the entropy of V. Wu and Verdú [28] interpreted the Rényi dimension as the fundamental limit of almost lossless data compression for analog sources under regularity constraints that include linearity of the compressor and Lipschitz continuity of the decompressor.

Our contributions in this paper are as follows. We propose the concept of Common Information Dimension (CID), defined as the minimum dimension of a random variable, with respect to a class of functions, required to distributively simulate a set of random vector variables $X_1, ..., X_n$. We define the Rényi common information dimension (RCID) as the minimum Rényi dimension of a random variable, with respect to a class of functions, required to distributively simulate $X_1, ..., X_n$. We define the Gács-Körner's common information dimension (GKCID) as the maximum Rényi dimension of a common function that can be extracted from each random variable individually. We prove that for jointly Gaussian random vectors, CID and RCID coincide, GKCID is upper bounded by CID, and they can all be computed by examining ranks of covariance matrices. Moreover, we give a closed form solution for the CID, RCID and GKCID of jointly Gaussian vectors with $\mathcal{F}$ being the class of linear functions and an efficient method to construct W with minimum dimension that enables the distributed simulation of $X_1, \cdots, X_n$.

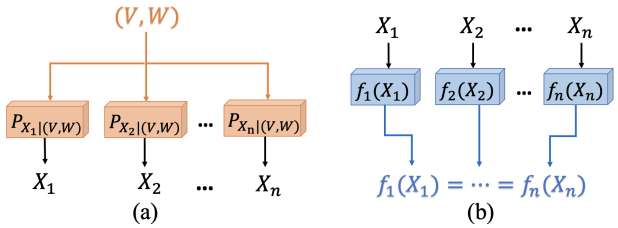

Fig. 1. The one-shot exact version of (a) Wyner's distributed simulation problem and (b) Gács-Körner's distributed randomness extraction problem.

II. Notation and Definitions

Notation. For a random (vector) variable X, we use d_X to denote the number of dimensions of X. We use $X = [X_1, \ldots, X_n]$ when X_i's are column vectors to refer to $X = [X_1^\top \ \ldots \ X_n^\top]^\top$. We use $[X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n|W]$ to abbreviate that $X_1, \cdots, X_n$ are conditionally independent given W. The entropy of a random variable V is denoted as $H(V)$. For a discrete random variable $H(V)$ is defined as $\sum_{v\in \text{Supp}(V)} -\mathbb{P}(V = v)\log_2 \mathbb{P}(V = v)$, where $\text{Supp}(V)$ is the support of V. For a matrix $\boldsymbol{M}$, we use $r(\boldsymbol{M})$ to denote the number of its rows.

Common Information Dimension (CID). We consider the one-shot exact version of the distributed simulation problem in Fig. 1 (a), where n distributed nodes leverage the common randomness (V, W), in addition to their own local randomness, to simulate random vectors $X_1, \cdots, X_n$ that follow a given joint distribution $\pi_{X_1,\cdots,X_n}$. We note that the distributed simulation is possible only if $X_1, \cdots, X_n$ are conditionally independent given (V, W). Specifically, each node i generates X_i according to a distribution $P_{X_i|(V,W)}$, and the joint distribution is required to satisfy $\pi_{X_1,\cdots,X_n} = \mathbb{E}_{V,W}\left[\prod_{i=1}^n P_{X_i|(V,W)}\right]$.

We assume that V is a (one-dimensional) random variable with finite entropy $H(V) < \infty$, and $W \in \mathbb{R}^{d_W}$ is a possibly continuous random vector of dimension d_W that can be expressed as $W = g(X_1, \ldots, X_n)$ for some function g in a given class of functions $\mathcal{F}$. Our goal is to determine the minimum dimension of W that is necessary to enable the distributed simulation. Note that we allow for the finite entropy random variable V to not follow the $\mathcal{F}$ restriction, and thus the common randomness needs to be expressed using a function in $\mathcal{F}$ only up to finite randomness. This allows the common information dimension to be zero when a finite amount of randomness is sufficient for the simulation, and avoids extra dimensions that may arise when this sufficient finite randomness cannot be expressed using $\mathcal{F}$.

Definition 1: The **Common Information Dimension (CID)** of random variables $X_1, \cdots, X_n$ with respect to a class of functions $\mathcal{F}$, is defined as

$$d_{\mathcal{F}}(X_1, \cdots, X_n) = \min\{d_W | W \in \mathcal{W}_{\mathcal{F}}\}, \tag{6}$$

where

$$\mathcal{W}_{\mathcal{F}} = \{W|\ \exists V, g : \mathbb{R}^{\sum_i^n d_{X_i}} \to \mathbb{R}^{d_W} \in \mathcal{F},\ \text{such that}$$
$$X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n|(V, W),\ H(V) < \infty,\ W = g(X_1, \cdots, X_n)\}.$$

We next define the concept of the **Rényi Common Information Dimension (RCID)** by replacing the dimension of W with the Renyi dimension described in (5).

Definition 2: The **Rényi Common Information Dimension (RCID)** of random variables $X_1, \cdots, X_n$ with respect to a class of functions $\mathcal{F}$ is defined as

$$d^R_{\mathcal{F}}(X_1, \cdots, X_n) = \inf\{d^R(W) | W \in \mathcal{W}_{\mathcal{F}}\}. \tag{7}$$

Finally, we define the **Gács-Körner's Common Information Dimension (GKCID)** (illustrated in Fig. 1 (b)) for continuous random variables by replacing the entropy with the Rényi dimension in the Gács-Körner's common information definition. This measures the maximum dimension of a vector W that can be extracted from each random variable individually, using a potentially different function $f_i \in \mathcal{F}$.

Definition 3: The **Gács-Körner's Common Information Dimension (GKCID)** of random variables $X_1, \cdots, X_n$ with respect to a class of functions $\mathcal{F}$ is defined as

$$d^{GK}_{\mathcal{F}}(X_1, \cdots, X_n) = \sup_{(\forall i \in [n]) W = f_i(X_i), f_i \in \mathcal{F}} d^R(W), \tag{8}$$

where the optimization is over $W, \{f_i\}_{i=1}^n$.

III. CID and RCID for Jointly Gaussian Random Variables

We note that although CID, RCID and GKCID are well-defined, it is not clear whether and how they can be computed. Similarly, the calculation of Wyner's common information and its variants is challenging since it involves optimizing a concave function over a non-convex set. Therefore, closed-form solutions are available only in special cases [3], [23], [30], [31]. In this section, we characterize the CID, and RCID for jointly Gaussian random variables when $\mathcal{F}$ is the class of linear functions. Our results show that CID and RCID are equal and that CID, RCID can be computed simply from ranks of covariance matrices. A closed form for GKCID is provided in Section IV and complete proofs are in the Appendix in [32].

We consider a jointly Gaussian random vector $X = [X_1, \cdots, X_n]$, where $X_i \in \mathbb{R}^{d_{X_i}}$, with covariance matrix $\mathbf{\Sigma}_X$. We use $\mathbf{\Sigma}_{I|J}$ for $I, J \subseteq \{1, \cdots, n\}$ to denote the conditional covariance matrix of X_I conditioned on X_J, where X_I denotes the elements of X with indices in the set I. We also use $-I$ to denote the complement of I in $\{1, \cdots, n\}$. To simplify notation, we drop the parentheses when listing the elements of the sets I, J. Also, as we only consider $\mathcal{F}$ being the class of linear functions, we omit it in the subscripts.

We assume without loss of generality that the variables X_i are zero-mean. It is known from [23], [24] that $\mathbf{\Sigma}_X$ is non-singular if and only if a finite amount of randomness is sufficient to break the dependency between $X_1, \cdots, X_n$. Hence, when $\mathbf{\Sigma}_X$ is singular, infinite entropy is necessary.

A. CID of Jointly Gaussian Random Variables

Theorem 1 derives the CID for jointly Gaussian random variables in a closed form. The proof also provides a method to construct the pair (V, W), with $\mathcal{F}$ the class of linear functions.

TABLE I
Table of notation for Theorem 2

Notation	Definition	
X, Y	jointly Gaussian random vectors	
$\mathbf{\Sigma}_{X,Y}, \mathbf{\Sigma}_X, \mathbf{\Sigma}_Y$	covariance matrices of $[X, Y], X, Y$ respectively	
$\boldsymbol{N}$	basis of the null space of $\mathbf{\Sigma}_{X,Y}$	
$\boldsymbol{N}_X, \boldsymbol{N}_Y$	$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{N}_X & -\boldsymbol{N}_Y \end{bmatrix}$	(9)
$\boldsymbol{N}'_X, \boldsymbol{N}'_Y$	basis of the complementary space of $\boldsymbol{N}_X, \boldsymbol{N}_Y$	(10)
$\boldsymbol{M}_X, \boldsymbol{M}_Y$	$\boldsymbol{M}_X = \begin{bmatrix} \boldsymbol{N}_X \\ \boldsymbol{N}'_X \end{bmatrix}, \boldsymbol{M}_Y = \begin{bmatrix} \boldsymbol{N}_Y \\ \boldsymbol{N}'_Y \end{bmatrix}$	(11)

Theorem 1: Let $X = [X_1, \cdots, X_n]$ be a jointly Gaussian random vector. The common information dimension between $X_1, \cdots, X_n$ with respect to the class of linear functions is

$$d(X_1, \cdots, X_n) = \sum_{i=1}^{n} \text{rank}(\mathbf{\Sigma}_{-i}) - (n-1)\text{rank}(\mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the covariance matrix of X, and $\mathbf{\Sigma}_{-i}$ is the covariance matrix of $[X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n]$.

Theorem 2 proves the result for the special case of two random variables X, Y; the proof derives properties that are later used to prove Theorem 1.

Theorem 2: Let $[X, Y]$ be a jointly Gaussian random vector. Then, the common information dimension between X, Y with respect to the class of linear functions equals

$$d(X, Y) = \text{rank}(\mathbf{\Sigma}_X) + \text{rank}(\mathbf{\Sigma}_Y) - \text{rank}(\mathbf{\Sigma}_{X,Y}).$$

Proof of Theorem 2: We start by stating the following lemma that enables to discover deterministic relations between X, Y by just examining the covariance matrix $\mathbf{\Sigma}_{X,Y}$. The proofs of all lemmas are in the Appendix in [32].

Lemma 1: Let $X = [X_1, X_2, ..., X_n]$ be a d_X-dimensional random vector, $d_X = \sum_{i=1}^n d_{X_i}$, with zero mean and covariance matrix $\mathbf{\Sigma}$. For any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{d_X}$, we have that

$$\boldsymbol{a}^\top X = \boldsymbol{b}^\top X \text{ almost surely, if and only if } \boldsymbol{a}^\top \mathbf{\Sigma} = \boldsymbol{b}^\top \mathbf{\Sigma}.$$

Corrollary 1: There is a subset $I \subseteq \{1, ..., d_X\}$ such that $|I| = \text{rank}(\mathbf{\Sigma}_X)$, and $X_I \perp\!\!\!\perp Y | (V, W)$ if and only if $X \perp\!\!\!\perp Y | (V, W)$.

Therefore, without loss of generality, we assume that $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are non-degenerate, which implies that $d_X = \text{rank}(\mathbf{\Sigma}_X)$, and $d_Y = \text{rank}(\mathbf{\Sigma}_Y)$.

Let $\boldsymbol{N} \in \mathbb{R}^{r(\boldsymbol{N}) \times (d_X + d_Y)}$ be a basis of the left null space of $\mathbf{\Sigma}_{X,Y}$. We next show some properties for the matrix $\boldsymbol{N}$ that will help us to prove the theorem. By definition of $\boldsymbol{N}$, we have the following facts

$$\textbf{Fact 1. } \boldsymbol{N}\mathbf{\Sigma}_{X,Y} = \mathbf{0}. \tag{12}$$

$$\begin{aligned} \textbf{Fact 2. } \text{rank}(\boldsymbol{N}) &= d_X + d_Y - \text{rank}(\mathbf{\Sigma}_{X,Y}) \\ &= \text{rank}(\mathbf{\Sigma}_X) + \text{rank}(\mathbf{\Sigma}_Y) - \text{rank}(\mathbf{\Sigma}_{X,Y}). \end{aligned} \tag{13}$$

Using Lemma 1 and (12), we have that $\boldsymbol{N}\begin{bmatrix} X^\top & Y^\top \end{bmatrix}^\top = \mathbf{0}$ almost surely. We partition $\boldsymbol{N}$ as $\begin{bmatrix} \boldsymbol{N}_X & -\boldsymbol{N}_Y \end{bmatrix}$, where $\boldsymbol{N}_X \in \mathbb{R}^{r(\boldsymbol{N}) \times d_X}$ and $\boldsymbol{N}_Y \in \mathbb{R}^{r(\boldsymbol{N}) \times d_Y}$. Then we have that

$$\boldsymbol{N}_X X = \boldsymbol{N}_Y Y. \tag{14}$$

$\boldsymbol{N}$ is full-row-rank by definition; in the following, we show that $\boldsymbol{N}_X$ (and similarly $\boldsymbol{N}_Y$) are also full-row-rank.

Lemma 2: Let $[X,Y]$ be a random vector with covariance matrix $\boldsymbol{\Sigma}_{X,Y}$ and $\boldsymbol{N} = \begin{bmatrix}\boldsymbol{N}_X & -\boldsymbol{N}_Y\end{bmatrix}$ be a basis for the null space of $\boldsymbol{\Sigma}_{X,Y}$, where $\boldsymbol{N} \in \mathbb{R}^{r(\boldsymbol{N})\times(d_X+d_Y)}$, $\boldsymbol{N}_X \in \mathbb{R}^{r(\boldsymbol{N})\times d_X}$, and $\boldsymbol{N}_Y \in \mathbb{R}^{r(\boldsymbol{N})\times d_Y}$. If $\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y$ are non-degenerate, we have that

$$\text{rank}(\boldsymbol{N}_X) = \text{rank}(\boldsymbol{N}_Y) = \text{rank}(\boldsymbol{N}). \tag{15}$$

Next, we define two square non-singular matrices $\boldsymbol{M}_X$ and $\boldsymbol{M}_Y$ as $\boldsymbol{M}_X = \begin{bmatrix}\boldsymbol{N}_X^\top & \boldsymbol{N}_X'^\top\end{bmatrix}^\top \in \mathbb{R}^{d_X\times d_X}$ and $\boldsymbol{M}_Y = \begin{bmatrix}\boldsymbol{N}_Y^\top & \boldsymbol{N}_Y'^\top\end{bmatrix}^\top \in \mathbb{R}^{d_Y\times d_Y}$, where $\boldsymbol{N}_X', \boldsymbol{N}_Y'$ are a basis for the complementary space of $\boldsymbol{N}_X$ and $\boldsymbol{N}_Y$, respectively. Lemma 3, and Lemma 4 show two properties of the quantities $\boldsymbol{M}_X, \boldsymbol{M}_Y, \boldsymbol{N}_X', \boldsymbol{N}_Y'$ which we prove in the Appendix A in [32].

Lemma 3: Let $\boldsymbol{M}_X \in \mathbb{R}^{d_X\times d_X}, \boldsymbol{M}_Y \in \mathbb{R}^{d_Y\times d_Y}$ be full-rank matrices, and X, Y, V, W be random vectors of dimension d_X, d_Y, d_V, and d_W respectively. We have that

$$\boldsymbol{M}_X X \perp\!\!\!\perp \boldsymbol{M}_Y Y | (V,W) \text{ if and only if } X \perp\!\!\!\perp Y | (V,W).$$

Lemma 4: Let $\boldsymbol{N}_X, \boldsymbol{N}_Y, \boldsymbol{N}_X', \boldsymbol{N}_Y'$ be as defined in (9) and (10). Conditioned on $\boldsymbol{N}_X X$, we have that $\begin{bmatrix}(\boldsymbol{N}_X' X)^\top & (\boldsymbol{N}_Y' Y)^\top\end{bmatrix}^\top$ has full-rank covariance matrix.

We are now ready to prove Theorem 2. We first show that the common information dimension is upper bounded as $d(X,Y) \le \text{rank}(\boldsymbol{\Sigma}_X) + \text{rank}(\boldsymbol{\Sigma}_Y) - \text{rank}(\boldsymbol{\Sigma}_{X,Y})$. Consider $\boldsymbol{N}_X X$ as a possible W. Conditioned on $\boldsymbol{N}_X X$, both $\boldsymbol{N}_X X$ and $\boldsymbol{N}_Y Y$ are deterministic, since $\boldsymbol{N}_X X = \boldsymbol{N}_Y Y$ from (14). Hence, conditioned on $\boldsymbol{N}_X X$, breaking the dependency between $\boldsymbol{M}_X X$ and $\boldsymbol{M}_Y Y$ reduces to breaking the dependency between $\boldsymbol{N}_X' X$ and $\boldsymbol{N}_Y' Y$ conditioned on $\boldsymbol{N}_X X$.

From Lemma 4, we have that conditioned on $\boldsymbol{N}_X X$, $\begin{bmatrix}(\boldsymbol{N}_X' X)^\top & (\boldsymbol{N}_Y' Y)^\top\end{bmatrix}^\top$ is jointly Gaussian with full-rank covariance matrix. Hence, by the result in [24], there is a V_W with $H(V_W) < \infty$ such that $\boldsymbol{N}_X' X \perp\!\!\!\perp \boldsymbol{N}_Y' Y | (W, V_W)$, where $W = \boldsymbol{N}_X X$. Since the covariance matrix of $\begin{bmatrix}(\boldsymbol{N}_X' X)^\top & (\boldsymbol{N}_Y' Y)^\top\end{bmatrix}^\top$ conditioned on $\boldsymbol{N}_X X$ does not depend on the value of $\boldsymbol{N}_X X$ and is only a function of the covariance matrix of $\boldsymbol{N}_X X$, V_W can be the same for all W, and we can refer to V_W as V. This shows that $\boldsymbol{M}_X X \perp\!\!\!\perp \boldsymbol{M}_Y Y | (\boldsymbol{N}_X X, V)$. By Lemma 3, $X \perp\!\!\!\perp Y | (\boldsymbol{N}_X X, V)$. Thus,

$$\begin{aligned} d(X,Y) \le d_{\boldsymbol{N}_X X} &\overset{(i)}{=} \text{rank}(\boldsymbol{N}_X) \overset{(15)}{=} \text{rank}(\boldsymbol{N}) \\ &\overset{(12)}{=} \text{rank}(\boldsymbol{\Sigma}_X) + \text{rank}(\boldsymbol{\Sigma}_Y) - \text{rank}(\boldsymbol{\Sigma}_{X,Y}), \end{aligned} \tag{16}$$

where (i) follows since $\boldsymbol{N}_X$ is full rank by Lemma 2.

Next, we prove in Lemma 5 the other direction, that $d(X,Y) \ge \text{rank}(\boldsymbol{N})$. At the heart of the lemma, we prove that if there is a common function that can be extracted from both X, Y, namely, $f_X(X) = f_Y(Y)$ a.s. for some f_X, f_Y, then for (V,W) to break the X,Y dependency, $f_X(X)$ (and hence $f_Y(Y)$) is a deterministic function of (V,W). We also show that if f_X, f_Y are linear and $W = \boldsymbol{A}\begin{bmatrix}X^\top & Y^\top\end{bmatrix}^\top$ for some $\boldsymbol{A}$, then $d_W \ge d_{f_X(X)}$.

Lemma 5: Let $[X,Y]$ be a jointly Gaussian random vector and V, W be random variables such that $W = \boldsymbol{A}\begin{bmatrix}X^\top & Y^\top\end{bmatrix}^\top$ for some matrix $\boldsymbol{A}$, $H(V) \le \infty$ and $X \perp\!\!\!\perp Y | (V,W)$. Let matrix $\boldsymbol{N}_X$ be such that $\boldsymbol{N}_X X$ has a non-degenerate covariance matrix. If there exists matrix $\boldsymbol{N}_Y$ such that $\boldsymbol{N}_X X = \boldsymbol{N}_Y Y$ a.s., then $\boldsymbol{N}_X X = \boldsymbol{A}' W$ a.s. for some matrix $\boldsymbol{A}'$ and $d_W \ge d_{\boldsymbol{N}_X X}$.

Proof of Lemma 5: First, we show that $\boldsymbol{N}_X X$ is a deterministic function of (V,W). Suppose towards a contradiction that there is a set $\mathcal{S} \subseteq \mathbb{R}^{r(\boldsymbol{N})}$ such that $0 < \mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] < 1$. Since $\boldsymbol{N}_X X = \boldsymbol{N}_Y Y$ a.s., we have that

$$\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S}, \boldsymbol{N}_Y Y \in \mathcal{S}^C | (V,W)] = 0, \tag{17}$$

where $\mathcal{S}^C$ is the complement of $\mathcal{S}$. However, from $0 < \mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] < 1$ we get that

$$\begin{aligned} &\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] \mathbb{P}[\boldsymbol{N}_Y Y \in \mathcal{S}^C | (V,W)] \\ &= \mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] \mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S}^C | (V,W)] \neq 0. \end{aligned} \tag{18}$$

This implies that

$$\begin{aligned} &\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S}, \boldsymbol{N}_Y Y \in \mathcal{S}^C | (V,W)] \\ &\neq \mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] \mathbb{P}[\boldsymbol{N}_Y Y \in \mathcal{S}^C | (V,W)]. \end{aligned} \tag{19}$$

However, as functions of independent random variables are independent, we have that $\boldsymbol{N}_X X, \boldsymbol{N}_Y Y$ are conditionally independent given (V,W); being projections of $\boldsymbol{M}_X X, \boldsymbol{M}_Y Y$. This contradicts (19).

This implies that for any $\mathcal{S} \subseteq \mathbb{R}^{r(\boldsymbol{N})}$ we either have $\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] = 1$ or $\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] = 0$. We show next that this implies that $\boldsymbol{N}_X X$ is a deterministic function of (V,W).

As the interval from $(-\infty, \infty)$ can be partitioned into countably many sets of the form $(0+m, 1+m]$, by countable additivity of measures we get that there is a cube of the form $\mathcal{S} = \prod_{i=1}^{r(\boldsymbol{N})} (0+m_i, 1+m_i]$ that has $\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{S} | (V,W)] = 1$. If we repeatedly halve one of the largest dimensions of the cube we get a sequence of hyper-rectangles $... \subseteq \mathcal{R}_2 \subseteq \mathcal{R}_1$ such that $\mathbb{P}[\boldsymbol{N}_X X \in \mathcal{R}_i | (V,W)] = 1, \forall i = 1, 2, ...$ and $\cap_{i\in\mathbb{N}} \mathcal{R}_i$ contains exactly one member. The last fact is proved in the following. We notice that $\cap_{i\in\mathbb{N}} \mathcal{R}_i$ contains at most one member because for any two points $x_1, x_2 \in \mathbb{R}^{r(\boldsymbol{N})}$, there is some i such that the largest dimension of $\mathcal{R}_i$ is less than $\|x_1 - x_2\|_2$ which implies that at most one point of x_1, x_2 can be in R_i. It is also not possible that $\cap_{i\in\mathbb{N}} \mathcal{R}_i$ is empty as by the continuity from above of finite measures, we have that

$$\mathbb{P}[\boldsymbol{N}_X X \in \cap_{i\in\mathbb{N}} \mathcal{R}_i | (V,W)] = 1. \tag{20}$$

Therefore, $\cap_{i\in\mathbb{N}} \mathcal{R}_i$ must contain a single member. Let us denote the unique point in $\cap_{i\in\mathbb{N}} \mathcal{R}_i$ by $g(V,W)$, where g is a deterministic function. Then, we have that $\mathbb{P}[\boldsymbol{N}_X X = g(V,W) | (V,W)] = 1$. Hence, we have that

$$\begin{aligned} H(\boldsymbol{N}_X X | W) &\le H(\boldsymbol{N}_X X, V | W) \\ &= H(\boldsymbol{N}_X X | (V,W)) + H(V|W) \\ &= H(V|W) \le H(V) < \infty. \end{aligned} \tag{21}$$

Since $W = \boldsymbol{A}\begin{bmatrix} X^\top & Y^\top \end{bmatrix}^\top$, we have that $(\boldsymbol{N}_X X, W)$ follows a jointly Gaussian distribution. As a result, conditioned on W, we have that $\boldsymbol{N}_X X$ is also jointly Gaussian, whose entropy is either 0 (for zero variance) or ∞. Based on (21), it must be that $H(\boldsymbol{N}_X X|W) = 0$. Hence, we have that $\boldsymbol{N}_X X = \boldsymbol{B}W$, for some $\boldsymbol{B} \in \mathbb{R}^{\text{rank}(\boldsymbol{N}) \times d_W}$. And as a result, $\boldsymbol{N}_X \boldsymbol{\Sigma}_X = \boldsymbol{B}\mathbb{E}(WX^\top)$. Then we have that

$$\text{rank}(\boldsymbol{N}_X \boldsymbol{\Sigma}_X) \overset{(i)}{=} \text{rank}(\boldsymbol{N}_X) \overset{(15)}{=} \text{rank}(\boldsymbol{N}) \leq \text{rank}(\boldsymbol{B}) \leq d_W, \tag{22}$$

where (i) follows from the fact that $\boldsymbol{N}_X X$ has full rank covariance matrix. □

Combining (16) and (22), we conclude that

$$\begin{aligned} d(X,Y) &= \text{rank}(\boldsymbol{N}) \\ &= \text{rank}(\boldsymbol{\Sigma}_X) + \text{rank}(\boldsymbol{\Sigma}_Y) - \text{rank}(\boldsymbol{\Sigma}_{X,Y}). \end{aligned}$$ □

We next prove our main result stated in Theorem 1.

Proof Outline of Theorem 1: We here give a proof outline, then formally prove our result in App. B [32]. The main part of the proof is illustrated in Algorithm 1 which constructs variables $Z = [Z_1, \cdots, Z_n]$ that satisfy: (i) (Lemma 8 [32]) conditioned on Z, the dependency between $X_1, \cdots, X_n$ can be broken using finite randomness. This is proved by showing that after eliminating from X the parts that can be almost surely determined by Z, the remaining part is jointly Gaussian with non-degenerate covariance matrix. This shows that CID is upper bounded by the total number of dimensions of Z. (ii) (Lemma 9 [32]) for any V, W that break the dependency between $X_1, \cdots, X_n$, we have that Z is a linear function of W. By showing that Z is jointly Gaussian with non-degenerate covariance matrix, we prove that the dimension of W is lower bounded by the dimension of Z, hence, CID is lower bounded by the number of dimensions of Z.

Algorithm 1 Algorithm to find (V, W)

1: **for** $i = 1, \cdots, n$ **do**
2: Find $\boldsymbol{A}_i$, a basis of the row space of $\boldsymbol{\Sigma}_{i|1:i-1}$.
3: Define $U_i = \boldsymbol{A}_i X_i$ (remove parts from X_i that can be obtained from previous $X_1, \cdots, X_{i-1}$).
4: Find $\boldsymbol{B}_i$, a basis of the row space of $\boldsymbol{\Sigma}_{i+1:n|1:i-1}$.
5: Define $Y_i = \boldsymbol{B}_i[X_{i+1}, \cdots, X_n]$ (remove parts from $[X_{i+1}, \cdots, X_n]$ that can be obtained from previous $X_1, \cdots, X_{i-1}$).
6: Find $\bar{\boldsymbol{N}}_i = [\boldsymbol{N}_i \ \bar{\boldsymbol{N}}_i]$, the null space of $\boldsymbol{\Sigma}_{U_i Y_i}$.
7: Let $Z_i = \boldsymbol{N}_i U_i$ (the parts of $X_{i+1}, \cdots, X_n$ that can be obtained from X_i but cannot be obtained from $X_1, \cdots, X_{i-1}$).
8: **end for**
9: Let $W = [Z_1, \cdots, Z_n]$.
10: Find $\boldsymbol{C}_i$: basis for the covariance matrix of X_i conditioned on W (the parts of X_i that cannot be obtained from W). Let $T_i = \boldsymbol{C}_i X_i$.
11: Use [24] to get V that breaks the dependency of $T_1, \cdots, T_n$ conditioned on W.

We build Z gradually as follows. Z_1 represents the information that X_1 contains about $[X_2, \cdots, X_n]$; namely, the linearly independent dimensions of $[X_2, \cdots, X_n]$ that can be determined from X_1. Then, Z_2 contains the amount of information that X_2 contains about $X_3, \cdots, X_n$ that X_1 does not contain. Generally, Z_i contains the information that X_i contains about $X_{i+1}, \cdots, X_n$ which is not contained in any of the previous $X_1, \cdots, X_{i-1}$. □

B. RCID of Jointly Gaussian Random Variables

Our next result shows that for jointly Gaussian random variables RCID and CID are the same.

Lemma 6: Let $[X_1, \cdots, X_n]$ be a jointly Gaussian random vector. Then, the Rényi common information dimension between $X_1, \cdots, X_n$ with respect to the class of $\mathcal{F}$ of linear functions is given by

$$d^R(X_1, \cdots, X_n) = d(X_1, \cdots, X_n).$$

IV. GKCID of Jointly Gaussian Random Variables

In this section, we provide a closed form solution for GKCID of jointly Gaussian random variables and a class of linear functions, and show that GKCID is not larger than CID. The result is stated in Theorem 3. The proof of this theorem also gives a method to construct W, given in (69) in [32], with the maximum information dimension. The proof is provided in the Appendix C in [32].

Theorem 3: Let $X = [X_1, \cdots, X_n]$ be a jointly Gaussian random vector. The GKCID between $X_1, \cdots, X_n$ with respect to the class of linear functions is given by

$$d^{GK}(X_1, \cdots, X_n) = r(\tilde{\boldsymbol{\Sigma}}) - \text{rank}(\tilde{\boldsymbol{\Sigma}}), \tag{23}$$

where $r(\tilde{\boldsymbol{\Sigma}})$ is the number of rows of $\tilde{\boldsymbol{\Sigma}}$, with

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}_{X'_1 X'_2} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{X'_2 X'_3} & \cdots & 0 \\ & \cdots & & \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}_{X'_{n-1} X'_n} \\ \mathbf{0}_1 \ \mathbf{1}_2 & \mathbf{1}_2 \ \mathbf{0}_3 & \cdots & \mathbf{0}_{n-1} \ \mathbf{0}_n \\ \mathbf{0}_1 \ \mathbf{0}_2 & \mathbf{0}_2 \ \mathbf{1}_3 & \cdots & \mathbf{0}_{n-1} \ \mathbf{0}_n \\ \mathbf{0}_1 \ \mathbf{0}_2 & \mathbf{0}_2 \ \mathbf{0}_3 & \cdots & \mathbf{1}_{n-1} \ \mathbf{0}_n \end{bmatrix}, \tag{24}$$

$X'_i = \boldsymbol{F}_i X_i, \forall i \in [n]$, $\boldsymbol{F}_i$ is a basis of the row space of $\boldsymbol{\Sigma}_{X_i}$, $\mathbf{0}_i \in \mathbb{R}^{1 \times d_{X'_i}}$ and $\mathbf{1}_i \in \mathbb{R}^{1 \times d_{X'_i}}$ are all zeros (and ones respectively) row vectors with the same dimension as X'_i.

The following corollary follows from Theorems 1 and 3.

Corrollary 2: For two jointly Gaussian random variables X_1, X_2 we have that $d(X_1, X_2) = d^{GK}(X_1, X_2)$.

This result does not extend to more than two variables: as the following example shows, GKCID can be strictly less than CID. We consider three random vectors X_1, X_2, X_3 with non-zero variance, $X_1 = X_2$ a.s., and $X_3, [X_1, X_2]$ independent. A W of dimension equal to d_{X_1} is required to break the dependency, hence, $d(X_1, X_2, X_3) = d_{X_1}$. However, as X_3 is independent of X_1, all functions with $f_1(X_1) = f_3(X_3)$ have zero entropy [2], [20], and zero information dimension. Corollary 3 follows from the proof of Theorem 3 and Lemma 5.

Corrollary 3: Let $X = [X_1, \cdots, X_n]$ be jointly Gaussian vectors, then $d^{GK}(X_1, \cdots, X_n) \leq d(X_1, \cdots, X_n)$.

REFERENCES

[1] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

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

[3] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.

[4] A. J. Bell, "The co-information lattice," in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, vol. 2003, 2003.

[5] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography. i. secret sharing," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1121–1132, 1993.

[6] ——, "Common randomness in information theory and cryptography. ii. cr capacity," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 225–240, 1998.

[7] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE transactions on information theory*, vol. 39, no. 3, pp. 733–742, 1993.

[8] I. Csiszár and P. Narayan, "Common randomness and secret key generation with a helper," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 344–366, 2000.

[9] H. Hotelling, "Relations between two sets of variates," *Breakthroughs in statistics: methodology and distribution*, pp. 162–190, 1992.

[10] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[11] O. Shayevitz, "On rényi measures and hypothesis testing," in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 894–898.

[12] L. Yu, V. Y. Tan *et al.*, "Common information, noise stability, and their extensions," *Foundations and Trends® in Communications and Information Theory*, vol. 19, no. 2, pp. 107–389, 2022.

[13] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[14] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[15] Y. Shi, B. Paige, P. Torr *et al.*, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] T. M. Sutter, I. Daunhawer, and J. E. Vogt, "Generalized multimodal ELBO," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=5Y21V0RDBV

[17] C. Zhang, Z. Han, H. Fu, J. T. Zhou, Q. Hu *et al.*, "Cpm-nets: Cross partial multi-view networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] G. R. Kumar, C. T. Li, and A. El Gamal, "Exact common information," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 161–165.

[19] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.

[20] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 100–113, 1975.

[21] W. Liu, G. Xu, and B. Chen, "The common information of n dependent random variables," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2010, pp. 836–843.

[22] K. B. Viswanatha, E. Akyol, and K. Rose, "The lossy common information of correlated sources," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3238–3253, 2014.

[23] G. Xu, W. Liu, and B. Chen, "A lossy source coding interpretation of wyner's common information," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 754–768, 2015.

[24] C. T. Li and A. El Gamal, "Distributed simulation of continuous random variables," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6329–6343, 2017.

[25] L. Yu and V. Y. Tan, "On exact and ∞-rényi common informations," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3366–3406, 2020.

[26] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[27] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

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

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

[30] H. S. Witsenhausen, "Values and bounds for the common information of two discrete random variables," *SIAM Journal on Applied Mathematics*, vol. 31, no. 2, pp. 313–333, 1976.

[31] S. Satpathy and P. Cuff, "Gaussian secure source coding and wyner's common information," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 116–120.

[32] O. Hanna, X. Li, S. Diggavi, and C. Fragouli, "Common information dimension," 2023, https://arxiv.org/abs/2305.06469.