A Local Characterization for Wyner Common Information

Shao-Lun Huang
DSIT Research Center
Tsinghua-Berkeley Shenzhen Institute
Shenzhen, China 518055
Email: shaolun.huang@sz.tsinghua.edu.cn

Xiangxiang Xu
Dept. of Electronic Engineering
Tsinghua University
Beijing, China 100084
Email: xuxx14@mails.tsinghua.edu.cn

Lizhong Zheng, Gregory W. Wornell
Dept. of Electrical & Computer Eng.
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
Email: {lizhong, gww}@mit.edu

***Abstract*—While the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation and the Wyner common information share similar information processing purposes of extracting common knowledge structures between random variables, the relationships between these approaches are generally unclear. In this paper, we demonstrate such relationships by considering the Wyner common information in the weakly dependent regime, called ϵ-common information. We show that the HGR maximal correlation functions coincide with the relative likelihood functions of estimating the auxiliary random variables in ϵ-common information, which establishes the fundamental connections these approaches. Moreover, we extend the ϵ-common information to multiple random variables, and derive a novel algorithm for extracting feature functions of data variables regarding their common information. Our approach is validated by the MNIST problem, and can potentially be useful in multi-modal data analyses.**

I. Introduction

For jointly distributed random variables X and Y, with ranges $\mathcal{X}$ and $\mathcal{Y}$ respectively, the (generalized) HGR maximal correlation [1]–[3], defined as

$$\rho_k(X;Y) \triangleq \sup_{\substack{\mathbf{f}\colon \mathcal{X}\mapsto\mathbb{R}^k,\ \mathbf{g}\colon \mathcal{Y}\mapsto\mathbb{R}^k \\ \mathbb{E}[\mathbf{f}(X)]=\mathbb{E}[\mathbf{g}(Y)]=\mathbf{0} \\ \mathbb{E}\left[\mathbf{f}(X)\mathbf{f}^{\mathrm{T}}(X)\right]=\mathbb{E}\left[\mathbf{g}(Y)\mathbf{g}^{\mathrm{T}}(Y)\right]=\mathbf{I}_k}} \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right], \tag{1}$$

attempts to extract k-dimensional features $f(X)$ and $g(Y)$ to optimize the Pearson correlation coefficient in the feature space, and then measure the dependence between X, Y by the resulted correlation coefficient. Conceptually, the formulation (1) intends to extract features from one variable that has high predictive power towards some aspects of the other variable, which can be interpreted as learning the common knowledge shared between these random variables. As such, the features extracted by (1) are shown useful in supervised learning, unsupervised learning, and multi-modal machine learning scenarios.

On the other hand, in traditional information theoretic discipline, the common knowledge between random variables is often described and extracted by auxiliary random variables, among which a widely adopted formulation is the Wyner common information, defined as

$$C(X,Y) = \min_{P_{W|XY}\colon\ X\leftrightarrow W\leftrightarrow Y} I(W;X,Y) \tag{2}$$

where W is the auxiliary random variable, and $X \leftrightarrow W \leftrightarrow Y$ forms the Markov chain. While the maximal correlation and Wyner common information share the same information processing purpose, the relationship between these two problems is generally unclear.

In this paper, we consider the Wyner common information for weakly dependent X, Y, from which the ϵ-common information problem is formulated. We show that in this formulation, the feature functions extracted from the HGR maximal correlation problem, with an appropriately chosen dimension k, coincide with the relative likelihood function of estimating the optimal W from X, Y in the ϵ-common information. This establishes a local-equivalence between HGR maximal correlation problem and Wyner common information. Moreover, our development indicates a (local) decomposition of the Wyner common information shared by X, Y, by the modal decomposition of their joint distribution P_{XY} corresponding to the singular value decomposition (SVD) of the canonical dependence matrix (CDM) [4].

Furthermore, motivated by [5], we generalize the ϵ-common information to multiple random variables with a class of joint distributions. Unlike two random variables, the resulted optimization problem is non-convex without analytical solutions, which is difficult to solve for continuous random variables. In this paper, we propose to solve a relaxed optimization problem that can be viewed as solving the original problem with quadratic regularizers. We show that the relaxed problem can be efficiently solved by deep neural networks, which also suggests a novel deep learning approach for extracting features from multi-modal data regarding to their common information. Finally, our experimental results demonstrate the effectiveness of the extracted features in real problems.

II. Preliminaries and Definitions

In this section, we briefly introduce some important results of HGR maximal correlation, as well as some definitions of local neighborhood of distributions that will be useful later on in our derivations.

A. HGR Maximal Correlation and SVD

For discrete random variables X, Y, the optimal $\mathbf{f}, \mathbf{g}$ of the HGR maximal correlation (1) is related to the SVD of the

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.

canonical dependence matrix (CDM).

Definition 1 (CDM). *The canonical dependence matrix $\tilde{\mathbf{B}}$ is the $|\mathcal{Y}| \times |\mathcal{X}|$ matrix with the (y,x)th entry defined as*

$$\tilde{B}(y,x) \triangleq \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}} \tag{3}$$

We denote the singular values of $\tilde{\mathbf{B}}$ as $\sigma_1 \geq \cdots \geq \sigma_K$, where $K \triangleq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, and the corresponding right and left singular vectors as $\boldsymbol{\psi}_i^X$ and $\boldsymbol{\psi}_i^Y$, respectively, for all i. In addition, we define the functions $f_i^* : \mathcal{X} \mapsto \mathbb{R}$ and $g_i^* : \mathcal{Y} \mapsto \mathbb{R}$, for all i, as

$$f_i^*(x) = \frac{\psi_i^X(x)}{\sqrt{P_X(x)}}, \quad g_i^*(y) = \frac{\psi_i^Y(y)}{\sqrt{P_Y(y)}}, \tag{4}$$

where $\psi_i^X(x)$ and $\psi_i^Y(y)$ are the xth and yth entries of $\boldsymbol{\psi}_i^X$ and $\boldsymbol{\psi}_i^Y$, respectively. Then, the functions f_i^* and g_i^* admit a modal decomposition of the joint distribution P_{XY}:

$$P_{XY}(x,y) = P_X(x)P_Y(y)\left[1 + \sum_{i=1}^{K-1} \sigma_i f_i^*(x) g_i^*(y)\right] \tag{5}$$

where $\mathbb{E}\left[f_i^*(X)\right] = \mathbb{E}\left[g_i^*(Y)\right] = 0$, and[1] $\mathbb{E}\left[f_i^*(X) f_j^*(X)\right] = \mathbb{E}\left[g_i^*(Y) g_j^*(Y)\right] = \mathbb{1}_{i=j}$. Moreover, the following result holds for HGR maximal correlation (1).

Proposition 2. *For discrete X, Y and $k < K$, it follows that*

$$\rho_k(X;Y) = \sum_{i=1}^{k} \sigma_i,$$

which is achieved by choosing

$$\mathbf{f}(x) = f_*^k(x) \triangleq \left(f_1^*(x), \cdots, f_k^*(x)\right),$$
$$\mathbf{g}(y) = g_*^k(y) \triangleq \left(g_1^*(y), \cdots, g_k^*(y)\right).$$

Consequently, the SVD of CDM decomposes the correlation between X, Y into a sequence of dependent modes with the strength of the modes characterized by the singular values. This decomposition will be revisited when we develop the Wyner common information for weakly dependent variables in the ϵ-common information problem in the next section.

B. The Neighborhood of Distributions

Let $\mathcal{P}^{\mathcal{Z}}$ denote the space of distributions on some finite alphabet $\mathcal{Z}$, where $|\mathcal{Z}| < \infty$, and let $\mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$ denote the relative interior of $\mathcal{P}^{\mathcal{Z}}$, i.e., the subset of strictly positive distributions.

Definition 3 (ϵ-Neighborhood). *For a given $\epsilon > 0$, the ϵ-neighborhood of a reference distribution $P_0 \in \mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$ is the set of distributions in a (Neyman) χ^2-divergence [6] ball of radius ϵ^2 about P_0, i.e.,*

$$\mathcal{N}_\epsilon^{\mathcal{Z}}(P_0) \triangleq \left\{P' \in \mathcal{P}^{\mathcal{Z}} \colon D_{\chi^2}(P'\|P_0) \leq \epsilon^2\right\}, \tag{6a}$$

where for $P \in \mathcal{P}^{\mathcal{Z}}$ and $Q \in \mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$,

$$D_{\chi^2}(P\|Q) \triangleq \sum_{z \in \mathcal{Z}} \frac{\left(Q(z) - P(z)\right)^2}{Q(z)}. \tag{6b}$$

In the sequel, for a distribution $P \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$, we define the *information vector* ϕ and *feature function* associated to P

$$\phi(z) \triangleq \frac{P(z) - P_0(z)}{\epsilon\sqrt{P_0(z)}}, \quad f(x) = \frac{\phi(x)}{\sqrt{P(x)}} \tag{7}$$

which results in a one-to-one correspondence

$$P \leftrightarrow \phi \leftrightarrow f \tag{8}$$

In addition, given $\epsilon > 0$, we refer to random variables X and Y as *ϵ-dependent* if $P_{XY} \in \mathcal{N}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$, where P_X and P_Y are the marginal distributions associated with P_{XY}. Since $D_{\chi^2}(P_{XY}\|P_X P_Y) = \|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2$, where $\tilde{\mathbf{B}}$ is the CDM as defined in (3), it follows that X and Y are ϵ-dependent if and only if $\|\tilde{\mathbf{B}}\|_{\mathrm{F}} \leq \epsilon$.

Definition 4 (sub-ϵ dependence). *For random variables X and Y with joint distribution P_{XY}, X and Y are sub-ϵ dependent if the corresponding CDM $\tilde{\mathbf{B}}$ as defined in (3) satisfies*

$$\|\tilde{\mathbf{B}}\|_* \leq \epsilon, \tag{9}$$

where $\|\cdot\|_*$ denotes the nuclear norm. In addition, given marginal distributions P_X and P_Y, we use $\bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ to denote the set of P_{XY} satisfying (9).

Since $\|\tilde{\mathbf{B}}\|_{\mathrm{F}} \leq \|\tilde{\mathbf{B}}\|_*$, the sub-$\epsilon$ dependence implies ϵ-dependence, and hence $\bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y) \subset \mathcal{N}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$.

III. The ϵ-Common Information Problem

Under sub-ϵ dependence, we define the following restricted common information.

Definition 5 (ϵ-Common Information). *Given $P_{XY} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, the ϵ-common information is*

$$C_\epsilon(X,Y) = \min_{P_{W|XY} \in \mathcal{P}_\epsilon} I(W; X, Y), \tag{10}$$

where

$$\mathcal{P}_\epsilon \triangleq \Big\{ P_{W|XY},\ \textit{some } \mathcal{W}\colon P_{W|XY}(\cdot|x,y) \in \mathcal{P}^{\mathcal{W}} \textit{ for all } (x,y), X \leftrightarrow W \leftrightarrow Y,\ \textit{and } P_{X|W}(\cdot|w) \in \mathcal{N}_{\sqrt{\delta(\epsilon)}}^{\mathcal{X}}(P_X),\ P_{Y|W}(\cdot|w) \in \mathcal{N}_{\sqrt{\delta(\epsilon)}}^{\mathcal{Y}}(P_Y),\ \textit{for all } w \in \mathcal{W} \textit{ and some } \delta(\cdot) > 0 \textit{ such that } \lim_{\epsilon \to 0} \delta(\epsilon) \to 0. \Big\}. \tag{11}$$

In Definition 5, an auxiliary random variable W such that $P_{W|XY}(\cdot|x,y) \in \mathcal{P}_\epsilon$ is characterized by the *configuration* $\mathcal{C}_\epsilon^{\mathcal{X},\mathcal{Y}}(P_{XY})$ defined as:

$$\mathcal{C}_\epsilon^{\mathcal{X},\mathcal{Y}}(P_{XY}) \triangleq \Big\{ \mathcal{W},\ \{P_W(w),\ w \in \mathcal{W}\},\ \{P_{X|W}(\cdot|w),\ w \in \mathcal{W}\},\ \{P_{Y|W}(\cdot|w),\ w \in \mathcal{W}\} \Big\} \tag{12}$$

subject to the constraints

$$P_{X|W}(\cdot|w) \in \mathcal{N}_{\sqrt{\delta(\epsilon)}}^{\mathcal{X}}(P_X),\ w \in \mathcal{W}, \tag{13a}$$
$$P_{Y|W}(\cdot|w) \in \mathcal{N}_{\sqrt{\delta(\epsilon)}}^{\mathcal{Y}}(P_Y),\ w \in \mathcal{W}, \tag{13b}$$

[1] We use the Kronecker notation

$$\mathbb{1}_{\mathcal{A}} = \begin{cases} 1 & \mathcal{A} \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.

for some δ such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and

$$P_{X|W}(x|w)\, P_{Y|W}(y|w) = P_{XY|W}(x,y|w). \tag{14}$$

In turn, (14) implies the constraint

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{X|W}(x|w)\, P_{Y|W}(y|w) = P_{XY}(x,y), \tag{15}$$

which further implies the constraints

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{X|W}(x|w) = P_X(x) \tag{16a}$$

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{Y|W}(y|w) = P_Y(y). \tag{16b}$$

Defining the information vectors

$$\phi_w^{X|W}(x) \triangleq \frac{P_{X|W}(x|w) - P_X(x)}{\sqrt{\delta(\epsilon)}\sqrt{P_X(x)}} \tag{17a}$$

$$\phi_w^{Y|W}(y) \triangleq \frac{P_{Y|W}(y|w) - P_Y(y)}{\sqrt{\delta(\epsilon)}\sqrt{P_Y(y)}}, \tag{17b}$$

we can equivalently express (12) in the form

$$\mathcal{C}_\epsilon^{\mathcal{X},\mathcal{Y}}(P_{XY}) \triangleq \Big\{ \mathcal{W},\ \{P_W(w),\ w\in\mathcal{W}\},\ \{\boldsymbol{\phi}_w^{X|W},\ w\in\mathcal{W}\}\ \{\boldsymbol{\phi}_w^{Y|W},\ w\in\mathcal{W}\}\Big\} \tag{18}$$

subject to the constraints

$$\|\boldsymbol{\phi}_w^{X|W}\| \le 1,\ \|\boldsymbol{\phi}_w^{Y|W}\| \le 1, \quad w\in\mathcal{W} \tag{19}$$

which correspond to (13), and

$$\sum_{w\in\mathcal{W}} P_W(w)\, \phi_w^{X|W}(x) = 0, \quad x\in\mathcal{X}, \tag{20a}$$

$$\sum_{w\in\mathcal{W}} P_W(w)\, \phi_w^{Y|W}(y) = 0, \quad y\in\mathcal{Y}, \tag{20b}$$

due to (16).

Moreover, it follows from (14) and (17a)–(17b)

$$P_{XY|W}(x,y|w) \tag{21}$$

$$= P_{X|W}(x|w)\, P_{Y|W}(y|w) \tag{22}$$

$$\begin{aligned} &= \big(P_X(x) + \sqrt{\delta(\epsilon)}\,\sqrt{P_X(x)}\,\phi_w^{X|W}(x)\big) \\ &\qquad \cdot \big(P_Y(y) + \sqrt{\delta(\epsilon)}\,\sqrt{P_Y(y)}\,\phi_w^{Y|W}(y)\big) \\ &= P_X(x)\,P_Y(y) + \sqrt{\delta(\epsilon)}\sqrt{P_X(x)\,P_Y(y)} \\ &\qquad \cdot \Big[\sqrt{P_Y(y)}\,\phi_w^{X|W}(x) + \sqrt{P_X(x)}\,\phi_w^{Y|W}(y) \\ &\qquad + \sqrt{\delta(\epsilon)}\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y)\Big]. \end{aligned} \tag{23}$$

Finally, it follows from the expectation of (23) with respect to P_W and (20) that

$$\begin{aligned} &P_{XY}(x,y) \\ &\quad = \sum_{w\in\mathcal{W}} P_W(w)\, P_{X|W}(x|w)\, P_{Y|W}(y|w) \\ &\quad = P_X(x)\,P_Y(y) \\ &\qquad + \sqrt{P_X(x)\,P_Y(y)} \\ &\qquad\quad \cdot \delta(\epsilon) \sum_{w\in\mathcal{W}} P_W(w)\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y), \end{aligned} \tag{24}$$

from which we obtain

$$\delta(\epsilon) \sum_{w\in\mathcal{W}} P_W(w)\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y) = \tilde{B}(y,x),\ x\in\mathcal{X},\ y\in\mathcal{Y}, \tag{25}$$

by recognizing $\tilde{B}(y,x)$ as defined in (3) as the final factor in (24).

The following variational characterization of the nuclear (i.e., trace) norm (see, e.g., [7]) is useful in our development.

Lemma 6. *Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$, we have*

$$\min_{\substack{\{k,\ \mathbf{M}_1\in\mathbb{R}^{k_1\times k},\ \mathbf{M}_2\in\mathbb{R}^{k\times k_2}:\\ \mathbf{M}_1\mathbf{M}_2=\mathbf{A}\}}} \left(\frac{1}{2}\|\mathbf{M}_1\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathbf{M}_2\|_{\mathrm{F}}^2\right) = \|\mathbf{A}\|_*. \tag{26}$$

In particular, we obtain that the ϵ-common information is given by the nuclear norm of $\tilde{\mathbf{B}}$.

Theorem 7. *Given $P_{XY} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, we have*[2]

$$C(X,Y) \le C_\epsilon(X,Y) = \|\tilde{\mathbf{B}}\|_* + o(\epsilon), \tag{27a}$$

where

$$\|\tilde{\mathbf{B}}\|_* = \sum_{i=1}^{K-1} \sigma_i, \tag{27b}$$

which is achieved by the configuration

$$\begin{aligned} &C_*^{\mathcal{X},\mathcal{Y}}(P_{XY}) \\ &\quad = \Big\{ \mathcal{W} = \{\pm 1, \ldots, \pm(K-1)\}, \\ &\qquad P_W(w) = \frac{\sigma_{|w|}}{2\|\tilde{\mathbf{B}}\|_*}, \\ &\qquad P_{X|W}(x|w) = P_X(x)\left(1 + \mathrm{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}\, f_{|w|}^*(x)\right), \\ &\qquad P_{Y|W}(y|w) = P_Y(y)\left(1 + \mathrm{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}\, g_{|w|}^*(y)\right)\Big\} \end{aligned} \tag{28}$$

and $\delta(\epsilon) = \epsilon$ in (11).

Proof. Due to the space limitations, we present the main steps of the proof here, while referring the full proof to [4, Appendix V-I].

Notice that from (25) and (9), we have

$$\delta(\epsilon) = O(\epsilon). \tag{29}$$

In addition, it follows from (23) and (24) that $I(W;X,Y)$ can be approximated as

$$I(W;X,Y) = \sum_{w\in\mathcal{W}} P_W(w)\, D(P_{XY|W}(\cdot,\cdot|w)\|P_{XY})$$

[2]Since $I(W;X,Y) \ge \max\{I(W;X), I(W;Y)\}$ by the chain rule, it follows that our result does not change if we further include in (11) of Definition 5 all distributions $P_{X|W}(\cdot|w)$ and $P_{Y|W}(\cdot|w)$ that for all $w\in\mathcal{W}$ do not depend on ϵ, since they will give rise to nonvanishing $I(W;X)$ and $I(W;Y)$. In essence, the configurations our definition omits are those for which the P_W is increasingly severely imbalanced as $\epsilon \to 0$.

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.

$$= \frac{\delta(\epsilon)}{2} \sum_{w\in\mathcal{W}} P_W(w) \left(\|\boldsymbol{\phi}_w^{X|W}\|^2 + \|\boldsymbol{\phi}_w^{Y|W}\|^2 \right) + o(\delta(\epsilon)), \tag{30}$$

Then, by defining

$$\tilde{\boldsymbol{\Phi}}^{X|W} \triangleq \sqrt{\delta(\epsilon)}\, \boldsymbol{\Phi}^{X|W} \sqrt{\mathbf{P}_W} \tag{31a}$$

$$\tilde{\boldsymbol{\Phi}}^{Y|W} \triangleq \sqrt{\delta(\epsilon)}\, \boldsymbol{\Phi}^{Y|W} \sqrt{\mathbf{P}_W}, \tag{31b}$$

where $\boldsymbol{\Phi}^{X|W}$ is a $|\mathcal{X}| \times |\mathcal{W}|$ matrix whose wth column is $\boldsymbol{\phi}_w^{X|W}$, where $\boldsymbol{\Phi}^{Y|W}$ is a $|\mathcal{Y}| \times |\mathcal{W}|$ matrix whose wth column is $\boldsymbol{\phi}_w^{Y|W}$, and where $\mathbf{P}_W$ is a $|\mathcal{W}| \times |\mathcal{W}|$ diagonal matrix whose wth diagonal entry is $P_W(w)$, we can equivalently express the constraint (25) in the form

$$\tilde{\mathbf{B}} = \tilde{\boldsymbol{\Phi}}^{Y|W} \left(\tilde{\boldsymbol{\Phi}}^{X|W} \right)^{\mathrm{T}}, \tag{32}$$

and the objective function (30) as

$$I(W;X,Y) = \frac{1}{2} \left(\|\tilde{\boldsymbol{\Phi}}^{X|W}\|_{\mathrm{F}}^2 + \|\tilde{\boldsymbol{\Phi}}^{Y|W}\|_{\mathrm{F}}^2 \right) + o(\delta(\epsilon)) \tag{33}$$

$$\geq \|\tilde{\mathbf{B}}\|_* + o(\epsilon), \tag{34}$$

where to obtain the inequality we have used (29) and Lemma 6 with (32). The equality is achieved by choosing the configuration for W according to

$$\mathcal{W} = \{\pm 1, \ldots, \pm(K-1)\}$$

$$\boldsymbol{\phi}_i^{X|W} = -\boldsymbol{\phi}_{-i}^{X|W} = \sqrt{\frac{\sigma_i}{\tilde{\sigma}_i\, \delta(\epsilon)}}\, \boldsymbol{\psi}_i^X, \quad i = 1, \ldots, K-1$$

$$\boldsymbol{\phi}_i^{Y|W} = -\boldsymbol{\phi}_{-i}^{Y|W} = \sqrt{\frac{\sigma_i}{\tilde{\sigma}_i\, \delta(\epsilon)}}\, \boldsymbol{\psi}_i^Y, \quad i = 1, \ldots, K-1$$

$$P_W(i) = P_W(-i) = \frac{1}{2}\tilde{\sigma}_i, \quad i = 1, \ldots, K-1,$$

where

$$\tilde{\sigma}_i \triangleq \frac{\sigma_i}{\sum_{i'=1}^{K-1} \sigma_{i'}},$$

which results in (28) by the correspondence (8) and (4), and choosing $\delta(\epsilon) = \epsilon$. □

The ϵ-common information variable W can be related to the modal decomposition (5) of the correlation between X, Y. To develop this, let us equivalently express W as

$$W \triangleq W^{K-1} = (W_1, \ldots, W_{K-1}), \tag{35a}$$

where each W_i is a variable defined over alphabet

$$\mathcal{W}_\circ \triangleq \{-1, 0, +1\} \tag{35b}$$

according to

$$W_i \triangleq \begin{cases} +1 & W = i \\ -1 & W = -i \\ 0 & \text{otherwise.} \end{cases} \tag{35c}$$

In particular, we have the following result, where the proof is referred to [4, Appendix V-K].

Theorem 8. *Given $P_{XY} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, and let W^{K-1} be the representation (35) of the optimizing ϵ-common information variable W in Proposition 7. Then*

$$C_\epsilon(X,Y) = I(W;X,Y) = \sum_{i=1}^{K-1} I(W_i;X,Y) + o(\epsilon), \tag{36}$$

where

$$I(W_i;X,Y) = \sigma_i + o(\epsilon), \quad i = 1, \ldots, K-1. \tag{37}$$

IV. The ϵ-Common Information for Multiple Random Variables

The Wyner common information was generalized by [5] to multiple variables, defined as

$$C(X^d) = \min I(W;X^d), \tag{38}$$

where $X^d \triangleq (X_1, \ldots, X_d)$ with the range $\mathcal{X}^d \triangleq \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, and the minimum is taken over all the joint distribution of W and X^d such that $X_1, \ldots, X_d$ are conditional independent given W.

In the sequel, we generalize the ϵ-common information to multiple random variables X^d with the joint distribution P_{X^d} belong to the class $\hat{\mathcal{P}}$, defined as

$$\hat{\mathcal{P}} \triangleq \Bigg\{ P_{X^d} \colon \mathbb{E}_{P_{X^d}} \left[\prod_{i\in\mathcal{I}} f_i(X_i) \right] = 0, \text{ for all } f_i : \mathcal{X}_i \mapsto \mathbb{R}, \text{ s.t. } \mathbb{E}\left[f_i(X_i)\right] = 0, \text{ and all } \mathcal{I} \subset \{1, \ldots, d\}, \text{ s.t. } |\mathcal{I}| \geq 3 \Bigg\}.$$

In particular, the distributions in $\hat{\mathcal{P}}$ only depend on the pairwise distributions of X^d, which is characterized by the following result with the proof referred to [8].

Proposition 9. *We have $P_{X^d} \in \hat{\mathcal{P}}$ if and only if, for all x^d,*

$$P_{X^d}(x^d) - \prod_{i=1}^d P_{X_i}(x_i) = \sum_{i<j} \left[P_{X_iX_j}(x_i,x_j) - P_{X_i}(x_i) P_{X_j}(x_j) \right] \prod_{l\neq i,j} P_{X_l}(x_l).$$

Then, under sub-ϵ dependence, we define the following ϵ-common information for random variables X^d with $P_{X^d} \in \hat{\mathcal{P}}$.

Definition 10. *Given $P_{X^d} \in \hat{\mathcal{P}} \cap \bar{\mathcal{N}}_\epsilon^{\mathcal{X}^d}(P_{X_1} \cdots P_{X_d})$ for $\epsilon > 0$, the multivariate ϵ-common information is defined as*

$$C_\epsilon(X^d) = \min_{P_{W|X^d} \in \mathcal{P}_\epsilon^{(d)}} I(W;X^d), \tag{39}$$

where

$$\mathcal{P}_\epsilon^{(d)} \triangleq \Bigg\{ P_{W|X^d},\ \text{some } \mathcal{W}\colon P_{W|X^d}(\cdot|x^d) \in \mathcal{P}^{\mathcal{W}} \text{ for all } x^d,\ P_{X^d|W} = \prod_{i=1}^d P_{X_i|W} \text{ and } P_{X_i|W}(\cdot|w) \in \mathcal{N}_{\sqrt{\delta(\epsilon)}}^{\mathcal{X}_i}(P_{X_i}),\ \text{for all } i = 1, \ldots, d, \text{ all } w \in \mathcal{W}, \text{ and some } \delta(\cdot) > 0 \text{ such that } \lim_{\epsilon\to 0} \delta(\epsilon) \to 0. \Bigg\}.$$

Similar to (31), we define, for $i = 1, \ldots, d$,

$$\tilde{\boldsymbol{\Phi}}^{X_i|W} \triangleq \sqrt{\delta(\epsilon)}\, \boldsymbol{\Phi}^{X_i|W} \sqrt{\mathbf{P}_W}$$

where $\boldsymbol{\Phi}^{X_i|W}$ is a $|\mathcal{X}_i| \times |\mathcal{W}|$ matrix whose wth column is the information vector $\boldsymbol{\phi}_w^{X_i|W}$ defined similar to (17). Then

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.

similar to (33), to the objective function (39) can be equivalently expressed as

$$I(W;X^d) = \frac{1}{2}\sum_{i=1}^{d} \|\tilde{\mathbf{\Phi}}^{X_i|W}\|_{\mathrm{F}}^2 + o(\delta(\epsilon)). \tag{40}$$

Moreover, similar to what (25) was obtained, we have the constraints [cf. (32)]

$$\tilde{\mathbf{B}}_{ij} = \tilde{\mathbf{\Phi}}^{X_i|W}\left(\tilde{\mathbf{\Phi}}^{X_j|W}\right)^{\mathrm{T}}, \quad 1 \le i < j \le d, \tag{41}$$

where $\tilde{\mathbf{B}}_{ij}$ is the CDM of X_i and X_j as defined in (3). Then, it follows from (40) and (41) that (39) can be expressed as

$$C_{|\mathcal{W}|} = \min_{\tilde{\mathbf{\Phi}}^{X_i|W} \in \mathbb{R}^{|\mathcal{X}_i|\times|\mathcal{W}|}} \frac{1}{2}\sum_{i=1}^{d} \left\|\tilde{\mathbf{\Phi}}^{X_i|W}\right\|_{\mathrm{F}}^2 \tag{42a}$$

$$\text{s.t.} \quad \tilde{\mathbf{B}}_{ij} = \tilde{\mathbf{\Phi}}^{X_i|W}\left(\tilde{\mathbf{\Phi}}^{X_j|W}\right)^{\mathrm{T}}, \quad 1 \le i < j \le d. \tag{42b}$$

As such, we have the following result for multivariate ϵ-common information, where the proof is referred to [8].

Proposition 11. *The ϵ-common information (39) satisfies*

$$C(X^d) \le C_\epsilon(X^d) = C_{k_0} + o(\epsilon),$$

where $k_0 = \sum_{i=1}^{d} |\mathcal{X}_i|$.

From the correspondence (8), the optimal solution of (42) provides feature functions of the variables regarding to their common information. However, unlike the two random variables case, (42) is non-convex without structural solutions, and is difficult to solve in practice when k_0 is very large. As such, we relax the constraints in (42) from two aspects: (i) we allow $|\mathcal{W}| = k$ to be designed less than k_0, which can be interpreted as designing the top k informative features regarding to the common information; (ii) we adopt the quadratic penalty method [9], which leads to a relaxed unconstrained optimization problem that minimizes

$$\frac{1}{2}\sum_{i=1}^{d} \left\|\tilde{\mathbf{\Phi}}^{X_i|W}\right\|_{\mathrm{F}}^2 + \frac{1}{2}\sum_{i<j} \gamma_{ij} \left\|\tilde{\mathbf{B}}_{ij} - \tilde{\mathbf{\Phi}}^{X_i|W}\left(\tilde{\mathbf{\Phi}}^{X_j|W}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \tag{43}$$

where the γ's are the penalty parameters. It can be shown that when the penalty parameters grow to infinity, the optimal solutions of (43) converge to the optimal solutions of (42) [9]. Moreover from (8), (43) can be equivalently expressed as a functional optimization problem, where the proof is referred to [8].

Proposition 12. *For given matrices $\tilde{\mathbf{\Phi}}^{X_i|W} \in \mathbb{R}^{|\mathcal{X}_i|\times k}$, $i = 1,\ldots,d$, define the feature functions $\mathbf{f}_i\colon \mathcal{X}_i \mapsto \mathbb{R}^k$ such that $\mathbf{f}_i(x_i) = \left(\tilde{\mathbf{\Phi}}^{X_i|W}(x_i)\right)^{\mathrm{T}} / \sqrt{P_{X_i}(x_i)}$, for all $i = 1,\ldots,d$ and $x_i \in \mathcal{X}_i$, where $\tilde{\mathbf{\Phi}}^{X_i|W}(x_i)$ denotes the x_ith row of $\tilde{\mathbf{\Phi}}^{X_i|W}$. Then, minimizing (43) is equivalent to the optimization problem*

$$\min_{\mathbf{f}_i\colon \mathcal{X}_i \mapsto \mathbb{R}^k} \frac{1}{2}\sum_{i=1}^{d} \mathbb{E}\left[\|\mathbf{f}_i(X_i)\|^2\right] - \sum_{i<j} \gamma_{ij} H(\mathbf{f}_i(X_i), \mathbf{f}_j(X_j)), \tag{44}$$

where $H(\mathbf{f}_i(X_i), \mathbf{f}_j(X_j))$ denotes the H-score [10] of $\mathbf{f}_i$ and $\mathbf{f}_j$, defined as

$$\begin{aligned} &H(\mathbf{f}_i(X_i), \mathbf{f}_j(X_j)) \\ &\triangleq \mathbb{E}\left[\mathbf{f}_i^{\mathrm{T}}(X_i)\mathbf{f}_j(X_j)\right] - \left(\mathbb{E}\left[\mathbf{f}_i(X_i)\right]\right)^{\mathrm{T}} \mathbb{E}\left[\mathbf{f}_j(X_j)\right] \\ &\quad - \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[\mathbf{f}_i(X_i)\mathbf{f}_i^{\mathrm{T}}(X_i)\right]\mathbb{E}\left[\mathbf{f}_j(X_j)\mathbf{f}_j^{\mathrm{T}}(X_j)\right]\right\}. \end{aligned}$$

In practice, instead of the true joint distribution P_{X_d}, it is often observed a sequence of training data $(x_1^{(i)},\ldots,x_d^{(i)})$ sampled from P_{X^d}, for $i = 1,\ldots,n$. In such cases, we notice that all the expectations in (44) can be computed by empirical averages of the training data, e.g., $\mathbb{E}\left[\|\mathbf{f}_i(X_i)\|^2\right]$ can be computed by $\frac{1}{n}\sum_{j=1}^{n}\|\mathbf{f}_i(x_i^{(j)})\|^2$. As such, (44) can be treated as a loss function, and the optimization problem can be solved by implementing d neural networks, where each function $\mathbf{f}_i$ is modeled by a neural network. Then, the parameters of the neural networks are trained by the training data to minimize (44) as the loss.

In particular, an algorithm similar to (44) was proposed in [11, Eq. (20)] to also extract the common structure between X^d by feature functions $\mathbf{f}_i(X_i)$ via minimizing the loss:

$$\min_{\mathbf{f}_i\colon \mathcal{X}_i \mapsto \mathbb{R}^k} -\frac{1}{2}\sum_{i=1}^{d} H(\mathbf{f}_i(X_i), \mathbf{f}_i(X_i)) - \sum_{i<j} H(\mathbf{f}_i(X_i), \mathbf{f}_j(X_j)) \tag{45}$$

We note from (45) and (44) that both approaches have the same target of extracting pairwise correlations of the random variables X^d via maximizing the H-scores $H(\mathbf{f}_i(X_i), \mathbf{f}_j(X_j))$, for $i < j$, and the difference between these two approaches can be viewed as applying different regularization terms in the loss functions. In practice, by fine tuning the parameters γ_{ij}'s, the features extracted from optimizing the loss (44) can perform better than optimizing the loss (45). For example, on the MNIST dataset, following the same setup as in [11], the k-dimensional features are extracted from $d = 64$ sub-images of the MNIST image via optimizing (44), with the parameters γ_{ij} set to $\gamma_{ij} = \gamma$, for all $i < j$. The extracted features are then used for label prediction, and the prediction errors (in percentages) are compared with (45), as shown in the following table. From the results we can see some performance gain by tuning γ_{ij}.

k	4	8	12	16	20	24
Approach (45)	3.46	1.73	1.43	**1.17**	1.15	1.11
(44), $\gamma = 10$	3.52	**1.69**	1.38	1.21	**1.10**	**1.03**
(44), $\gamma = 500$	**3.33**	1.71	1.48	1.27	1.30	1.14
(44), $\gamma = 1000$	3.47	1.80	**1.34**	**1.17**	1.27	1.17

ACKNOWLEDGMENT

The work of Shao-Lun Huang was supported in part by the Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20170818094022586, and in part by the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3–4, pp. 441–451, 1959.

[2] H. O. Hirschfeld, "A connection between correlation and contingency," *Proc. Cambridge Phil. Soc.*, vol. 31, pp. 520–524, 1935.

[3] H. Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichungsrechnung," *Z. für angewandte Math., Mech.*, vol. 21, pp. 364–379, 1941.

[4] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[5] 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.

[6] J. Neyman, "Contributions to the theory of the χ^2 test. in proceedings of the first berkeley symposium on mathematical statistics and probability," 1949.

[7] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[8] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "A local characterization for Wyner common information," *preprint*, 2020. [Online]. Available: https://xiangxiangxu.com/media/documents/isit2020a.pdf

[9] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, "Chapter iii constrained nonlinear programming," *Handbooks in Operations Research and Management Science*, vol. 1, pp. 171–210, 1989.

[10] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," *arXiv preprint arXiv:1905.06600*, 2019.

[11] S.-L. Huang, X. Xu, and L. Zheng, "An information-theoretic approach to unsupervised feature selection for high-dimensional data," *arXiv preprint arXiv:1911.03196*, 2019.

Authorized licensed use limited to: MIT Libraries. Downloaded on November 06,2020 at 19:56:13 UTC from IEEE Xplore. Restrictions apply.