Multivariate Feature Extraction

Xiangxiang Xu[1] and Lizhong Zheng[1]

***Abstract*— We propose a general framework to extract features for learning problems involving multiple variables. First, we decompose multivariate dependence into different components to obtain the piece relevant to the learning task. Then, we establish a modal decomposition approach to represent the component as informative features. We further demonstrate the algorithm design for extracting such features from real data, which can incorporate and utilize existing deep feature extractors. We also present an application of our framework in learning tasks with side information.**

I. Introduction

Extracting features from data is one of the most fundamental tasks in machine learning [1]. With the rapid development of deep learning [2], it is a common practice to extract features by training deep neural networks, where the learning objective is often set to a specific task, e.g., classification or prediction [3]. The effectiveness of such approaches has also been justified by information-theoretic and statistical analyses [4], [5]. In particular, it was shown in [5] that, the dominant dependence structures between X and Y are captured by deep neural networks that use X to predict Y.

On the other hand, machine learning problems involving more than two variables have also gained much attentions recently, e.g., multimodal learning [6] and distributed learning [7]. For such problems, it is usually difficult to construct a single learning task to extract useful features, due to the potentially complicated dependence among multiple variables. As a motivating example, let us consider a learning task with side information. Specifically, given data variable X, the goal is to infer Y based on feature extracted from X and some side information S. Such tasks can appear in the distributed learning scenarios, where a local node extracts the feature, and the cloud center conducts inference based on the extracted feature and some side information. To guarantee the performance in these tasks, the feature extracted from X shall capture information about Y that S cannot provide. Therefore, the feature design depends on some joint dependence among X, Y, and S, which cannot be converted to a simple bivariate learning task.

In this paper, our goal is to establish a general framework for solving such multivariate feature extraction problems. In particular, we develop a mathematical framework for multivariate dependence decomposition, based on the orthogonal decomposition of feature spaces. In addition, by decomposing joint function into orthogonal modals, we represent each dependence component as corresponding informative features. Moreover, we design a novel multivariate feature extraction algorithm, which can incorporate and make use of existing deep feature extractors to extract the informative features. Specifically, we propose a new training objective, called the *nested H-score*, which generalizes the H-score [8], [9] to multivariate scenarios and can also provide orthogonality guarantees for extracted features. Furthermore, we present an application of our framework in designing feature extraction and inference algorithms, for the learning task with side information.

*This work was supported in part by the National Science Foundation(NSF) under Award CNS-2002908 and the Office of Naval Research (ONR) under grant N00014-19-1-2621.

[1]Xiangxiang Xu and Lizhong Zheng are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 {xuxx, lizhong}@mit.edu

II. Notations

A. Vector Spaces

Given an inner product space with inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$, we can define the projection and orthogonal complement as follows.

Definition 1. *Give a subspace $\mathcal{W}$ of $\mathcal{V}$, we denote the projection of a vector $v \in \mathcal{V}$ onto $\mathcal{W}$ by*

$$\Pi(v; \mathcal{W}) \triangleq \arg\min_{w \in \mathcal{W}} \|v - w\|^2 . \tag{1}$$

In addition, we use $\mathcal{V} \boxminus \mathcal{W}$ to denote the orthogonal complement of $\mathcal{W}$ in $\mathcal{V}$, viz.,

$$\mathcal{V} \boxminus \mathcal{W} \triangleq \{v \in \mathcal{V} \colon \langle v, w \rangle = 0 \text{ for all } w \in \mathcal{W}\}. \tag{2}$$

We use "$\boxplus$" to denote the direct sum of orthogonal subspaces, i.e., $\mathcal{V} = \mathcal{V}_1 \boxplus \mathcal{V}_2$ indicates that $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2$ and $\mathcal{V}_1 \perp \mathcal{V}_2$. Then we have the following facts.

Fact 1. *If $\mathcal{V} = \mathcal{V}_1 \boxplus \mathcal{V}_2$, then $\mathcal{V}_2 = \mathcal{V} \boxminus \mathcal{V}_1$. In addition, if $\mathcal{W}$ is a subspace of $\mathcal{V}$, then $\mathcal{V} = \mathcal{W} \boxplus (\mathcal{V} \boxminus \mathcal{W})$.*

Fact 2 (Orthogonality Principle)**.** *Given $v \in \mathcal{V}$ and a subspace $\mathcal{W}$ of $\mathcal{V}$, then $w = \Pi(v; \mathcal{W})$ if and only if $w \in \mathcal{W}$ and $v - w \in \mathcal{V} \boxminus \mathcal{W}$. In addition, we have*

$$v = \Pi(v; \mathcal{W}) + \Pi(v; \mathcal{V} \boxminus \mathcal{W}) .$$

B. Distribution and Feature Spaces

1) Distribution Space: Given an alphabet $\mathcal{Z}$, we use $\mathcal{P}^{\mathcal{Z}}$ to denote the collection of probability distributions supported on $\mathcal{Z}$, and use $\mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$ to denote the relative interior of $\mathcal{P}^{\mathcal{Z}}$, i.e., the collection of distributions in $\mathcal{P}^{\mathcal{Z}}$ that have positive probability masses.

2) Feature Space: For given $\mathcal{Z}$, we use $\mathcal{F}_{\mathcal{Z}} \triangleq \{\mathcal{Z} \to \mathbb{R}\}$ to denote the collection of features (functions) of $\mathcal{Z}$. Specifically, we use $\mathcal{F}_{\varnothing}$ to represent the collection of constant features. Given $f_1, \ldots, f_k \in \mathcal{F}_{\mathcal{Z}}$, for each

$i = 1, \ldots, k$, we denote the multidimensional feature $\left(z \mapsto (f_1(z), \ldots, f_i(z))^{\mathrm{T}}\right)$ by $f_{[i]}$. We also denote $\Lambda_f \triangleq \mathbb{E}\left[f(Z)f^{\mathrm{T}}(Z)\right]$ for feature $f\colon \mathcal{Z} \to \mathbb{R}^k$.

Moreover, given a distribution $R \in \mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$, we define the inner product on $\mathcal{F}_{\mathcal{Z}}$ as $\langle f_1, f_2\rangle \triangleq \mathbb{E}_R[f_1(Z)f_2(Z)]$, which induces the norm $\|f\| \triangleq \sqrt{\langle f, f\rangle}$. We will refer to the distribution R as the *metric distribution*.

3) Joint Function: Given alphabets $\mathcal{X}, \mathcal{Y}$ and a metric distribution $R_{X,Y} \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ is composed of all joint functions of x and y. In particular, for given $f \in \mathcal{F}_{\mathcal{X}}$, $g \in \mathcal{F}_{\mathcal{Y}}$, we use $f \otimes g$ to denote their product $((x,y) \mapsto f(x)\cdot g(y)) \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, and refer to such features as *rank-one functions*. In addition, for given $f_1, \ldots, f_k \in \mathcal{F}_{\mathcal{X}}$ and $g_1, \ldots, g_k \in \mathcal{F}_{\mathcal{Y}}$, we denote $f_{[k]} \otimes g_{[k]} \triangleq \sum_{i=1}^k f_i \otimes g_i$.

Note that by extending each $f = (x \mapsto f(x)) \in \mathcal{F}_{\mathcal{X}}$ to $((x,y) \mapsto f(x)) \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, $\mathcal{F}_{\mathcal{X}}$ can be regarded as a subspace of $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, with the metric distribution being the marginal distribution R_X of $R_{X,Y}$. We then denote the orthogonal complement of $\mathcal{F}_{\mathcal{X}}$ in $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ as

$$\mathcal{F}_{\mathcal{Y}|\mathcal{X}} \triangleq \mathcal{F}_{\mathcal{X}\times\mathcal{Y}} \boxminus \mathcal{F}_{\mathcal{X}}. \tag{3}$$

Specifically, we use $\mathcal{F}_{\mathcal{X}|\varnothing} \triangleq \mathcal{F}_{\varnothing\times\mathcal{X}} \boxminus \mathcal{F}_{\varnothing} = \mathcal{F}_{\mathcal{X}} \boxminus \mathcal{F}_{\varnothing}$ to represent the collection of zero-mean functions on $\mathcal{X}$.

4) Likelihood Ratios: We establish a correspondence between the distribution space $\mathcal{P}^{\mathcal{Z}}$ and the feature space $\mathcal{F}_{\mathcal{Z}}$ by the likelihood ratio functions.

Definition 2. *Given a metric distribution $R \in \mathrm{relint}(\mathcal{P}^{\mathcal{Z}})$, for each $P \in \mathcal{P}^{\mathcal{Z}}$, we denote the likelihood ratio function by $\ell_{P;R} \in \mathcal{F}_{\mathcal{Z}}$ with $\ell_{P;R}(z) \triangleq \frac{P(z)}{R(z)}$. In addition, we define the* centered likelihood ratio *function $\tilde{\ell}_{P;R} \in \mathcal{F}_{\mathcal{Z}}$ as*

$$\tilde{\ell}_{P;R}(z) \triangleq \frac{P(z) - R(z)}{R(z)}, \quad \text{for all } z \in \mathcal{Z}.$$

When there is no ambiguity about metric distribution R, we will simply use $\ell_P, \tilde{\ell}_P$ to denote $\ell_{P;R}, \tilde{\ell}_{P;R}$, respectively.

III. Feature Transition and Decomposition

In this section, we further investigate the relation between joint functions in $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ and the features in $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Y}}$. For convenience, we will assume all metric distributions used in the section take the product form, i.e., $R_{X,Y} = R_X R_Y$.

A. Transition Operators

Definition 3 (Transition Operators)**.** *For given $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, its associated transition operator $\mathsf{T}(\gamma; \mathcal{X}, \mathcal{Y})$ is a mapping from $\mathcal{F}_{\mathcal{X}}$ to $\mathcal{F}_{\mathcal{Y}}$, such that for all $f \in \mathcal{F}_{\mathcal{X}}$, we have $[\tau(f)](y) = \mathbb{E}_{R_X}[\gamma(X,y)f(X)]$, where we have denoted $\tau \triangleq \mathsf{T}(\gamma; \mathcal{X}, \mathcal{Y})$.*

Given a transition operator $\tau = \mathsf{T}(\gamma; \mathcal{X}, \mathcal{Y})$, we use τ^* to denote the transition operator $\mathsf{T}(\gamma; \mathcal{Y}, \mathcal{X})$, referred to as the transpose (or the Hermitian adjoint) of τ.

Note that the transition operator is an extension of conditional expectation operators [10], which have been shown useful in analyzing bivariate dependence structure [4]. Indeed, for any given joint distribution $P_{X,Y}$, the conditional expectation operators $f \mapsto \mathbb{E}[f(X)|Y = \cdot\,]$ and $g \mapsto \mathbb{E}[g(Y)|X = \cdot\,]$ are the transition operators $\mathsf{T}\left(\ell_{P_{X,Y}}; \mathcal{X}, \mathcal{Y}\right)$ and and its adjoint $\mathsf{T}\left(\ell_{P_{X,Y}}; \mathcal{Y}, \mathcal{X}\right)$, respectively, where we have set the metric distribution $R_{X,Y} = P_X P_Y$. As we will see shortly in our development, this extension is useful for analyzing multivariate dependence structures, e.g., conditional dependence.

We also have the following fact of transition operators.

Fact 3. *The transition operator $\tau \triangleq \mathsf{T}(\gamma; \mathcal{X}, \mathcal{Y})$ is a linear operator, and is also linear to γ. In addition, for all $f \in \mathcal{F}_{\mathcal{X}}$ and $g \in \mathcal{F}_{\mathcal{Y}}$, we have $\langle \tau(f), g\rangle = \langle f, \tau^*(g)\rangle = \langle \gamma, f \otimes g\rangle$.*

B. Modal Decomposition

Given a joint function $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, γ is centered if we have $\mathbb{E}_{R_Y}[\gamma(x,Y)] + \mathbb{E}_{R_X}[\gamma(X,y)] \equiv 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, we have the following modal decomposition of joint functions in $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$. This decomposition can be regarded as the singular value decomposition of linear operators, also known as Schmidt decomposition [11], [12]. A proof of Proposition 1 is provided in Appendix A.

Proposition 1. *For each $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, there exist $K \geq 0$, $\sigma_1 \geq \cdots \geq \sigma_K > 0$, and features $f_1^*, \ldots, f_K^* \in \mathcal{F}_{\mathcal{X}|\varnothing}$, $g_1^*, \ldots, g_K^* \in \mathcal{F}_{\mathcal{Y}|\varnothing}$, such that*[1] *$\langle f_i^*, f_j^*\rangle = \langle g_i^*, g_j^*\rangle = \delta_{ij}$ for all $i, j = 1, \ldots, K$, and*

$$\gamma(x,y) = \sum_{i=1}^K \sigma_i f_i^*(x) g_i^*(y) - \mathbb{E}_{R_{X,Y}}[\gamma(X,Y)] + \mathbb{E}_{R_Y}[\gamma(x,Y)] + \mathbb{E}_{R_X}[\gamma(X,y)]. \tag{4}$$

In particular, if γ is centered, then we have

$$\gamma(x,y) = \sum_{i=1}^K \sigma_i f_i^*(x) g_i^*(y), \tag{5}$$

and we refer to the number K as the rank of γ.

C. Spectral Properties

Suppose $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ is centered and has modal decomposition (5). Let $\tau \triangleq \mathsf{T}(\gamma; \mathcal{X}, \mathcal{Y})$. Then, we can show that f_i^*'s, g_i^*'s are eigenfunctions of operators $\tau^* \circ \tau$ and $\tau \circ \tau^*$, respectively. In particular, for all $f \in \mathcal{F}_{\mathcal{X}}$ and $g \in \mathcal{F}_{\mathcal{Y}}$, we have

$$\tau(f) = \sum_{i=1}^K (\sigma_i \langle f, f_i^*\rangle) \cdot g_i^*, \tag{6}$$

$$\tau^*(g) = \sum_{i=1}^K (\sigma_i \langle g, g_i^*\rangle) \cdot f_i^*. \tag{7}$$

Therefore, for each $i = 1, \ldots, K$, we have $\tau(f_i^*) = \sigma_i \cdot g_i^*, \tau^*(g_i^*) = \sigma_i \cdot f_i^*$ and thus $[\tau^* \circ \tau](f_i^*) = \sigma_i^2 \cdot f_i^*, [\tau \circ \tau^*](g_i^*) = \sigma_i^2 \cdot g_i^*$.

In addition, we have the following property. A proof is provided in Appendix B.

Property 1. *For all $i = 1, \ldots, K$, we have*

$$f_i^* = \arg\max_{f_i} \|\tau(f_i)\|, \tag{8}$$

[1] We adopt the Kronecker delta notation

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

where the maximization is taken over all $f_i \in \mathcal{F}_\mathcal{X}$ with $\|f_i\| = 1$ and $\langle f_i, f_j^* \rangle = 0$ for $j = 1, \ldots, i-1$. Similarly,

$$(f_i^*, g_i^*) = \arg\max_{f_i, g_i} \langle \gamma, f_i \otimes g_i \rangle \tag{9}$$

where the maximization is taken over all $f_i \in \mathcal{F}_\mathcal{X}$ and $g_i \in \mathcal{F}_\mathcal{Y}$ with $\|f_i\| = \|g_i\| = 1$ and $\langle f_i, f_j^* \rangle = \langle g_i, g_j^* \rangle = 0$ for $j = 1, \ldots, i-1$.

Moreover, we have the following result on approximating a joint function by rank-one functions, which is an immediate corollary of Eckart–Young–Mirsky theorem [13].

Property 2. *For all $k \geq 1$, $f_1, \ldots, f_k \in \mathcal{F}_\mathcal{X}$, and $g_1, \ldots, g_k \in \mathcal{F}_\mathcal{Y}$, we have*

$$\left\| \gamma - f_{[k]} \otimes g_{[k]} \right\|^2 \geq \sum_{i=k+1}^{K} \sigma_i^2,$$

where the inequality holds with equality if and only if

$$\sum_{i=1}^{k} f_i \otimes g_i = \sum_{i=1}^{k} \sigma_i (f_i^* \otimes g_i^*).$$

D. Decomposition of Joint Distribution

For any joint distribution $P_{X,Y}$, with the metric distribution $R_{X,Y} = P_X P_Y$, the transition operator $\tau \triangleq \mathsf{T}\left(\tilde{\ell}_{P_{X,Y}}; \mathcal{X}, \mathcal{Y}\right)$ measures the amount of information about Y that is carried by features in $\mathcal{F}_\mathcal{X}$, as illustrated by the following result.

Lemma 1 ([14])**.** *Suppose $(X, Y) \sim P_{X,Y}$, and $Z \sim \mathcal{N}(0,1)$ and is independent of X, Y. Then, for all $f \in \mathcal{F}_\mathcal{X}$ with $\|f\| = 1$, we have*

$$\lim_{\mathsf{snr} \to 0^+} \frac{1}{\mathsf{snr}} \cdot I\left(f(X) + \frac{Z}{\sqrt{\mathsf{snr}}}; Y\right) = \frac{1}{2} \cdot \|\tau(f)\|^2,$$

where snr represents signal-to-noise ratio, and $I(\cdot;\cdot)$ represents mutual information.

In particular, the modal decomposition of $\tilde{\ell}_{P_{X,Y}}$ is also referred to as the modal decomposition of the joint distribution $P_{X,Y}$ (cf. [4, Proposition 2]). Let the decomposition be

$$\tilde{\ell}_{P_{X,Y}} = \sum_{i=1}^{K} \sigma_i \cdot (f_i^* \otimes g_i^*), \tag{10}$$

then the functions f_i^*, g_i^* correspond to the maximally correlated functions in $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, known as Hirschfeld–Gebelein–Rényi (HGR) maximal correlation functions [15], [16], [17]. Then, an information-theoretic optimality of f_i^* can be obtained from Property 1 and Lemma 1. Moreover, it has been shown in [4] that, if X and Y are weakly dependent and satisfy

$$\left\| \tilde{\ell}_{P_{X,Y}} \right\| \leq \epsilon \tag{11}$$

for some small ϵ, then f_i^*, g_i^* are the most informative features for a series of inference tasks. In addition, (10) can also be illustrated as a decomposition of mutual information.

Lemma 2 ([4, Lemma 16])**.** *If $P_{X,Y}$ satisfies* (11)*, then*

$$I(X;Y) = \frac{1}{2}\left\| \tilde{\ell}_{P_{X,Y}} \right\|^2 + o(\epsilon^2) = \frac{1}{2} \sum_{i=1}^{K} \sigma_i^2 + o(\epsilon^2),$$

where $\sigma_1, \ldots, \sigma_K$ are defined by the modal decomposition (10)*.*

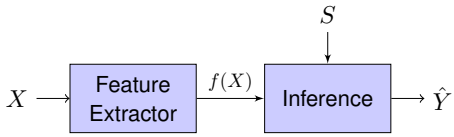

Fig. 1. Learning With Side Information S

The maximal correlation functions can be effectively extracted from data via maximizing the H-score [5], [8].

Definition 4. *Given $f\colon \mathcal{X} \to \mathbb{R}^k$, $g\colon \mathcal{X} \to \mathbb{R}^k$, the H-score $\mathscr{H}(f,g)$ is defined as*

$$\mathscr{H}(f,g) \triangleq \mathbb{E}\left[f^{\mathrm{T}}(X) g(Y)\right] - \left(\mathbb{E}\left[f(X)\right]\right)^{\mathrm{T}} \mathbb{E}\left[g(Y)\right] - \frac{1}{2} \cdot \mathrm{tr}\left(\Lambda_f \Lambda_g\right), \tag{12}$$

where we have defined $\Lambda_f \triangleq \mathbb{E}\left[f(X) f^{\mathrm{T}}(X)\right]$ and $\Lambda_g \triangleq \mathbb{E}\left[g(Y) g^{\mathrm{T}}(Y)\right]$, respectively.

Since we can equivalently express $\mathscr{H}(f,g)$ as

$$\mathscr{H}(f,g) = \frac{1}{2}\left(\left\| \tilde{\ell}_{P_{X,Y}} \right\|^2 - \left\| \tilde{\ell}_{P_{X,Y}} - f \otimes g \right\|^2 \right), \tag{13}$$

the following corollary of Property 2 is immediate.

Corollary 1. *Given $f\colon \mathcal{X} \to \mathbb{R}^k, g\colon \mathcal{Y} \to \mathbb{R}^k$, $\mathscr{H}(f,g)$ is maximized if and only if $f \otimes g = \sum_{i=1}^{k} \sigma_i f_i^* \otimes g_i^*$, where σ_i, f_i^*, g_i^* are given by the modal decomposition* (10)*.*

IV. Multivariate Dependence Decomposition

In this section, we introduce a multivariate dependence decomposition framework, by considering the learning problem with side information. As shown in Fig. 1, the goal is to infer the label Y based on extracted feature $f(X)$ and the side information S.

To begin, let $P_{X,S,Y}$ denote the joint distribution of data variable X, side information S, and label Y. Throughout our development, we set the product distribution $R_{X,S,Y} \triangleq P_X P_{S,Y}$ as the metric distributions, and consider the weak dependence regime with

$$\left\| \tilde{\ell}_{P_{X,S,Y}} \right\| \leq \epsilon \tag{14}$$

for some small $\epsilon > 0$. In addition, we define $P^{(\mathsf{M})}_{X,S,Y}$ as

$$P^{(\mathsf{M})}_{X,S,Y}(x,s,y) \triangleq P_{X|S}(x|s) P_S(s) P_{Y|S}(y|s), \tag{15}$$

which can be regarded as a coupling of $P_{X,S,Y}$ subject to the Markov relation $X - S - Y$.

Note that since the side information S has been revealed during the inference, the extracted feature $f(X)$ shall carry only the information not included in S. Therefore, the $X - S$ dependence will not be useful for the inference. To exclude this dependence component from the joint dependence, we consider the orthogonal decomposition [cf. Fact 1 and (3)]

$$\mathcal{F}_{\mathcal{X}\times\mathcal{S}\times\mathcal{Y}} = \mathcal{F}_{\mathcal{X}\times\mathcal{S}} \boxplus \mathcal{F}_{\mathcal{Y}|(\mathcal{X}\times\mathcal{S})}. \tag{16}$$

From Fact 2, we can further decompose $\tilde{\ell}_{P_{X,S,Y}}$ as

$$\tilde{\ell}_{P_{X,S,Y}} = \pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}}) + \pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}}), \tag{17}$$

where we have defined $\pi_{\mathsf{M}}(\gamma) \triangleq \Pi\left(\gamma; \mathcal{F}_{\mathcal{X}\times\mathcal{S}}\right)$ and $\pi_{\mathsf{C}}(\gamma) \triangleq \Pi\left(\gamma; \mathcal{F}_{\mathcal{Y}|\mathcal{X}\times\mathcal{S}}\right)$ for all $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{S}\times\mathcal{Y}}$. Then, we have the following result for the decomposed components, a proof of which is provided in Appendix C.

Proposition 2. *We have* $\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}}) = \tilde{\ell}_{P_{X,S}} = \tilde{\ell}_{P^{(\mathsf{M})}_{X,S,Y}}$ *and*

$$\left[\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})\right](x,s,y) = \frac{P_{X,S,Y}(x,s,y) - P^{(\mathsf{M})}_{X,S,Y}(x,s,y)}{R_{X,S,Y}(x,s,y)},$$

where $P^{(\mathsf{M})}_{X,S,Y}$ *is as defined in* (15).

Therefore, (17) decomposes the joint dependence between X and (Y,S), represented as $\tilde{\ell}_{P_{X,S,Y}}$, into two orthogonal components: the Markov component $\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}})$ that characterizes the dependence between X and S, and the Conditional dependence component $\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})$ that measures the conditional dependence between X and Y given S.

Moreover, from (17) we can readily obtain

$$\left\|\tilde{\ell}_{P_{X,S,Y}}\right\|^2 = \left\|\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}})\right\|^2 + \|\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})\|^2, \quad (18)$$

which corresponds to a chain rule of mutual information. To see this, note that from Lemma 2 and (14) we have

$$\left\|\tilde{\ell}_{P_{X,S,Y}}\right\|^2 = 2\cdot I(X;S,Y) + o(\epsilon^2).$$

Similarly, since $\left\|\tilde{\ell}_{P_{X,S}}\right\| \le \left\|\tilde{\ell}_{P_{X,S,Y}}\right\| \le \epsilon$, we obtain

$$\left\|\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}})\right\|^2 = \left\|\tilde{\ell}_{P_{X,S}}\right\|^2 = 2\cdot I(X;S) + o(\epsilon^2),$$

Hence, from (18), we have the correspondence between the conditional dependence component $\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})$ and the conditional mutual information $I(X;Y|S)$:

$$\begin{aligned}\|\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})\|^2 &= \left\|\tilde{\ell}_{P_{X,S,Y}}\right\|^2 - \left\|\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}})\right\|^2 \\ &= 2\cdot I(X;Y|S) + o(\epsilon^2).\end{aligned}$$

Furthermore, let us define the corresponding transition operators $\tau_{\mathsf{J}}, \tau_{\mathsf{M}}, \tau_{\mathsf{C}} \colon \mathcal{F}_{\mathcal{X}} \to \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}$ as

$$\tau_{\mathsf{J}} \triangleq \mathsf{T}\left(\tilde{\ell}_{P_{X,S,Y}}; \mathcal{X}, \mathcal{S}\times\mathcal{Y}\right), \quad (19)$$

$$\tau_{\mathsf{M}} \triangleq \mathsf{T}\left(\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}}); \mathcal{X}, \mathcal{S}\times\mathcal{Y}\right), \quad (20)$$

$$\tau_{\mathsf{C}} \triangleq \mathsf{T}\left(\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}}); \mathcal{X}, \mathcal{S}\times\mathcal{Y}\right), \quad (21)$$

which correspond to the Joint dependence of X and (S,Y), the dependence contributed by the Markov distribution $P^{(\mathsf{M})}_{X,S,Y}$, and the Conditional dependence of X and Y given S, respectively. Note that since the transition operator $\mathsf{T}(\gamma;\cdot,\cdot)$ is linear to γ, from (17) we obtain the decomposition

$$\tau_{\mathsf{J}} = \tau_{\mathsf{C}} + \tau_{\mathsf{M}}. \quad (22)$$

In addition, the following theorem demonstrates that the images of τ_{C} and τ_{M} are also orthogonal. A proof is provided in Appendix D.

Theorem 1. *For all* $f \in \mathcal{F}_{\mathcal{X}}$, *we have* $\tau_{\mathsf{C}}(f) \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$, $\tau_{\mathsf{M}}(f) \in \mathcal{F}_{\mathcal{S}}$, *and*

$$\|\tau_{\mathsf{J}}(f)\|^2 = \|\tau_{\mathsf{C}}(f)\|^2 + \|\tau_{\mathsf{M}}(f)\|^2. \quad (23)$$

In addition, we have $\tau^*_{\mathsf{C}} \circ \tau_{\mathsf{M}} = 0, \tau^*_{\mathsf{M}} \circ \tau_{\mathsf{C}} = 0$.

Moreover, we can also interpret (23) as a chain rule of mutual information. To see this, let $Z \sim \mathcal{N}(0,1)$ be independent of X, Y, S. Then, it follows from Lemma 1 that for all $f \in \mathcal{F}_{\mathcal{X}}$ with $\|f\| = 1$, we have

$$\lim_{\mathsf{snr}\to 0^+} \frac{1}{\mathsf{snr}} \cdot I\left(f(X) + \frac{Z}{\sqrt{\mathsf{snr}}}; S, Y\right) = \frac{1}{2}\cdot\|\tau_{\mathsf{J}}(f)\|^2,$$

$$\lim_{\mathsf{snr}\to 0^+} \frac{1}{\mathsf{snr}} \cdot I\left(f(X) + \frac{Z}{\sqrt{\mathsf{snr}}}; S\right) = \frac{1}{2}\cdot\|\tau_{\mathsf{M}}(f)\|^2.$$

Therefore, from the chain rule of mutual information $I(U;S,Y) = I(U;S) + I(U;Y|S)$, where $U \triangleq f(X) + \frac{Z}{\sqrt{\mathsf{snr}}}$, we can readily obtain

$$\begin{aligned}\lim_{\mathsf{snr}\to 0^+} \frac{1}{\mathsf{snr}} \cdot I\left(f(X) + \frac{Z}{\sqrt{\mathsf{snr}}}; Y \middle| S\right) &= \frac{1}{2}\cdot\left(\|\tau_{\mathsf{J}}(f)\|^2 - \|\tau_{\mathsf{M}}(f)\|^2\right) \\ &= \frac{1}{2}\cdot\|\tau_{\mathsf{C}}(f)\|^2. \quad (24)\end{aligned}$$

Therefore, the transition operator τ_{C} measures the amount of additional information about Y carried by features in $\mathcal{F}_{\mathcal{X}}$, when the side information S has been revealed. Moreover, suppose the modal decomposition of $\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})$ is

$$\left[\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})\right](x,s,y) = \sum_{i=1}^{K} \sigma_i f_i^*(x) g_i^*(s,y). \quad (25)$$

Then, from Property 1, $f_1^*, \ldots, f_K^*$ are the most informative features in the task of inferring Y with the presence of side information.

Furthermore, we can interpret the features f_i^*, g_i^* as a generalization of HGR maximal correlation functions. In particular, let us define

$$C_{\mathrm{s}}(f,g) \triangleq \mathbb{E}_{P_{X,S,Y}}[f(X)g(S,Y)] - \mathbb{E}_{P^{(\mathsf{M})}_{X,S,Y}}[f(X)g(S,Y)]$$

for $f \in \mathcal{F}_{\mathcal{X}}, g \in \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}$, where $P^{(\mathsf{M})}_{X,S,Y}$ is as defined in (15). Then, note that since $C_{\mathrm{s}}(f,g) = \left\langle f\otimes g, \pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})\right\rangle$, we have the following corollary of Proposition 1.

Corollary 2. *For all* $i = 1, \ldots, K$, *we have* $\sigma_i = C_{\mathrm{s}}(f_i^*, g_i^*)$ *and* f_i^*, g_i^* *correspond to the optimal solution of the optimization problem*

$$(f_i^*, g_i^*) = \arg\max_{f_i, g_i} C_{\mathrm{s}}(f_i, g_i),$$

where the maximization is taken over all $f_i \in \mathcal{F}_{\mathcal{X}}$ *and* $g_i \in \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}$ *with* $\|f_i\| = \|g_i\| = 1$ *and* $\langle f_i, f_j^*\rangle = \langle g_i, g_j^*\rangle = 0$ *for* $j = 1, \ldots, i-1$.

Remark 1. *When* X *and* S *are independent, we have* $P^{(\mathsf{M})}_{X,Y,S} = P_{X|S}P_{S,Y} = P_X P_{S,Y}$. *Therefore,* f_i^*, g_i^* *correspond to the maximal correlation functions of* X *and* (Y,S). *Specifically, when there is no side information (i.e.,* S *is a constant),* f_i^*, g_i^* *will be the maximal correlation functions of* X *and* Y.

V. Learning Algorithms

In this section, we demonstrate the practical algorithm design based on our multivariate feature extraction framework, where we focus on the illustrating example of learning with side information.

A. Deep Feature Extraction

Suppose $\pi_{\mathsf{C}}(\tilde{\ell}_{P_{X,S,Y}})$ has the modal decomposition (25). Then, the features f_i^*, g_i^* characterize the conditional independence of X and Y given side information S, and thus are useful for the inference task. To extract such features from real data, we develop a novel training objective based on the multivariate dependence decomposition, called nested H-score. The nested H-score is a multivariate extension of the H-score [8], [9] and can be computed by a nested network structure.

In particular, for given $\lambda > 0$, features $\bar{f}\colon \mathcal{X} \to \mathbb{R}^{\bar{k}}, \bar{g}\colon \mathcal{S} \to \mathbb{R}^{\bar{k}}$ and $f\colon \mathcal{X} \to \mathbb{R}^k, g\colon \mathcal{Y} \times \mathcal{S} \to \mathbb{R}^k$, we define the nested H-score $\mathscr{H}_{\mathsf{C}}(\bar{f}, \bar{g}, f, g; \lambda)$ as

$$\mathscr{H}_{\mathsf{C}}(\bar{f}, \bar{g}, f, g; \lambda) \triangleq \mathscr{H}(\bar{f}, \bar{g}) + \lambda \cdot \mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right), \quad (26)$$

where $\mathscr{H}(\cdot, \cdot)$ is the H-score as defined in Definition 4.

Note that the nested H-score $\mathscr{H}_{\mathsf{C}}$ is the weighted sum of two H-scores, where the input features of the two H-scores form a nested structure. In particular, note that since $\bar{f}$ and $\bar{g}$ are features of X and S, respectively, we can extract $X - S$ dependence by maximizing the first H-score $\mathscr{H}(\bar{f}, \bar{g})$.

Moreover, by applying a two-phase training strategy, the second H-score can be used to extract the conditional dependence between X and Y given S. Specifically, in the first phase, we maximize $\mathscr{H}(\bar{f}, \bar{g})$ such that $\bar{f}, \bar{g}$ have captured all dependence between X and S. Next, in the second phase, we fix $\bar{f}, \bar{g}$, and maximize the second H-score $\mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right)$ over only f and g. Then, the optimal f and g shall extract the dependence structure between X and (S, Y) that is not captured by $\bar{f}$ and $\bar{g}$, which corresponds to the conditional dependence of X and Y given S.

The following theorem demonstrates that, directly optimizing the nested H-score $\mathscr{H}_{\mathsf{C}}$ provides the same solution as this two-phase strategy. A proof is provided in Appendix E.

Theorem 2. *Let $\bar{K}$ denote the rank of $\tilde{\ell}_{P_{X,S}}$. Given $k \geq 1$, $\bar{k} \geq \bar{K}$, $\bar{f}\colon \mathcal{X} \to \mathbb{R}^{\bar{k}}, \bar{g}\colon \mathcal{S} \to \mathbb{R}^{\bar{k}}$, and $f\colon \mathcal{X} \to \mathbb{R}^k, g\colon \mathcal{Y} \times \mathcal{S} \to \mathbb{R}^k$, for all $\lambda > 0$, the H-score $\mathscr{H}_{\mathsf{C}}(\bar{f}, \bar{g}, f, g; \lambda)$ is maximized if and only if*

$$\bar{f} \otimes \bar{g} = \tilde{\ell}_{P_{X,S}}, \quad (27a)$$

$$f \otimes g = \sum_{i=1}^{k} \sigma_i \left(f_i^* \otimes g_i^*\right). \quad (27b)$$

From Theorem 2, the learned feature $f = f_1, \ldots, f_k$ lie in the subspace spanned by $f_1^*, \ldots, f_k^*$. However, unlike $f_1^*, \ldots, f_k^*$ that are orthogonal and ordered by their informativeness (cf. Property 1 and Lemma 1), $f_1, \ldots, f_k$ can take any collection of k linearly independent features in the subspace. To guarantee that the extracted feature f preserve the order and orthogonality structure as f_i^*'s, we can again apply the nested trick. To this end, we define the orthogonalized nested H-score $\mathscr{H}_{\mathsf{C}}^{\perp}$, such that for $\bar{f}\colon \mathcal{X} \to \mathbb{R}^{\bar{k}}, \bar{g}\colon \mathcal{Y} \to \mathbb{R}^{\bar{k}}$ and $f = (f_1, \ldots, f_k)^{\mathrm{T}}\colon \mathcal{X} \to \mathbb{R}^k, g = (g_1, \ldots, g_k)^{\mathrm{T}}\colon \mathcal{Y} \times \mathcal{S} \to \mathbb{R}^k$,

$$\mathscr{H}_{\mathsf{C}}^{\perp}(\bar{f}, \bar{g}, f, g) \triangleq \sum_{i=0}^{k} \mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f_{[i]}\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g_{[i]}\end{bmatrix}\right),$$

where we have defined $\begin{bmatrix}\bar{f}\\ f_{[0]}\end{bmatrix} \triangleq \bar{f}, \begin{bmatrix}\bar{g}\\ g_{[0]}\end{bmatrix} \triangleq \bar{g}$.

Note that the orthogonalized nested H-score $\mathscr{H}_{\mathsf{C}}^{\perp}$ is related to the nested H-score $\mathscr{H}_{\mathsf{C}}$ via

$$\mathscr{H}_{\mathsf{C}}^{\perp}(\bar{f}, \bar{g}, f, g) = \frac{1}{k} \sum_{i=1}^{k} \mathscr{H}_{\mathsf{C}}(\bar{f}, \bar{g}, f_{[i]}, g_{[i]}; k). \quad (28)$$

Hence, by applying Theorem 2 to all $\mathscr{H}_{\mathsf{C}}(\bar{f}, \bar{g}, f_{[i]}, g_{[i]}; k)$, $i = 1, \ldots, k$, we can readily obtain the following result.

Corollary 3. *Given $\bar{f}\colon \mathcal{X} \to \mathbb{R}^{\bar{k}}, \bar{g}\colon \mathcal{S} \to \mathbb{R}^{\bar{k}}$ with $\bar{k} \geq \bar{K}$, and $f\colon \mathcal{X} \to \mathbb{R}^k, g\colon \mathcal{Y} \times \mathcal{S} \to \mathbb{R}^k$, the orthogonalized nested H-score $\mathscr{H}_{\mathsf{C}}^{\perp}(\bar{f}, \bar{g}, f, g)$ is maximized if and only if*

$$\bar{f} \otimes \bar{g} = \tilde{\ell}_{P_{X,S}}, \quad (29a)$$

$$f_i \otimes g_i = \sigma_i \left(f_i^* \otimes g_i^*\right), \quad i = 1, \ldots, k. \quad (29b)$$

Remark 2. *From Remark 1, when we set the side information S to a constant, we can learn the maximal correlation functions of X and Y by maximizing $\mathscr{H}_{\mathsf{C}}^{\perp}(0, 0, f, g)$.*

Remark 3. *By introducing monotonic mappings, we can construct functions of $\bar{f}, \bar{g}, f, g$ that satisfy the same property as demonstrated in Corollary 3, i.e., that can be used as the training objective. For example, for all $c_0, \ldots, c_k > 0$, it can be verified that the function*

$$\sum_{i=0}^{k} \exp\left(c_i \cdot \mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f_{[i]}\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g_{[i]}\end{bmatrix}\right)\right)$$

is maximized if and only if (29) *holds.*

To compute the orthogonalized nested H-score $\mathscr{H}_{\mathsf{C}}^{\perp}$, we design the nested network structure as illustrated in Fig. 2, where we have used the "$+\!\!+$" symbol to indicate concatenation of two vectors, we have $v_1 +\!\!+ v_2 \triangleq \left[\begin{smallmatrix}v_1\\ v_2\end{smallmatrix}\right]$ for two column vectors v_1, v_2. In addition, each gray block in the figure represents a deep feature extractor with trainable parameters. The architectures of these feature extractors can be chosen according to the data type. Then, the features $\bar{f}, \bar{g}, f, g$ are learned by training the parameters in these deep feature extractors to maximize $\mathscr{H}_{\mathsf{C}}^{\perp}$.

B. Inference

First, we have the following result on the posterior distribution $P_{Y|X,S}$. A proof is provided in Appendix F.

Theorem 3. *We have*

$$P_{Y|X,S}(y|x,s) = P_{Y|S}(y|s)\left(1 + \sum_{i=1}^{K} \sigma_i f_i^*(x) g_i^*(s,y)\right) + o(\epsilon), \quad (30)$$

where σ_i, f_i^*, g_i^* are as defined in (25)*.*

Note that the posterior distribution depends on x only through $f_i^*(x)$, $i = 1, \ldots, K$. Hence, the K-dimensional feature $f_{[K]}^* = (f_1^*, \cdots, f_K^*)^{\mathrm{T}}$ is a sufficient statistic of X for the side information inference problem. Then, the corresponding MAP (maximum a posteriori) estimation for Y can be expressed as

$$\hat{y}_{\mathrm{MAP}}(x, s) = \arg\max_{y \in \mathcal{Y}} P_{Y|X,S}(y|x,s). \quad (31)$$

In practical learning tasks, due to the computation or communication constraints, it may happen that the feature extracted from X is constrained to have only some $k < K$ dimensions. In such cases, we cannot directly use sufficient statistic $f_{[K]}^*$ and apply MAP estimation. Instead, we can approximate the posterior distribution $P_{Y|X,S}(y|x,s)$ by

$$P_{Y|X,S}^{(k)}(y|x,s) \triangleq P_{Y|S}(y|s)\left(1 + \sum_{i=1}^{k} \sigma_i f_i^*(x) g_i^*(s,y)\right),$$

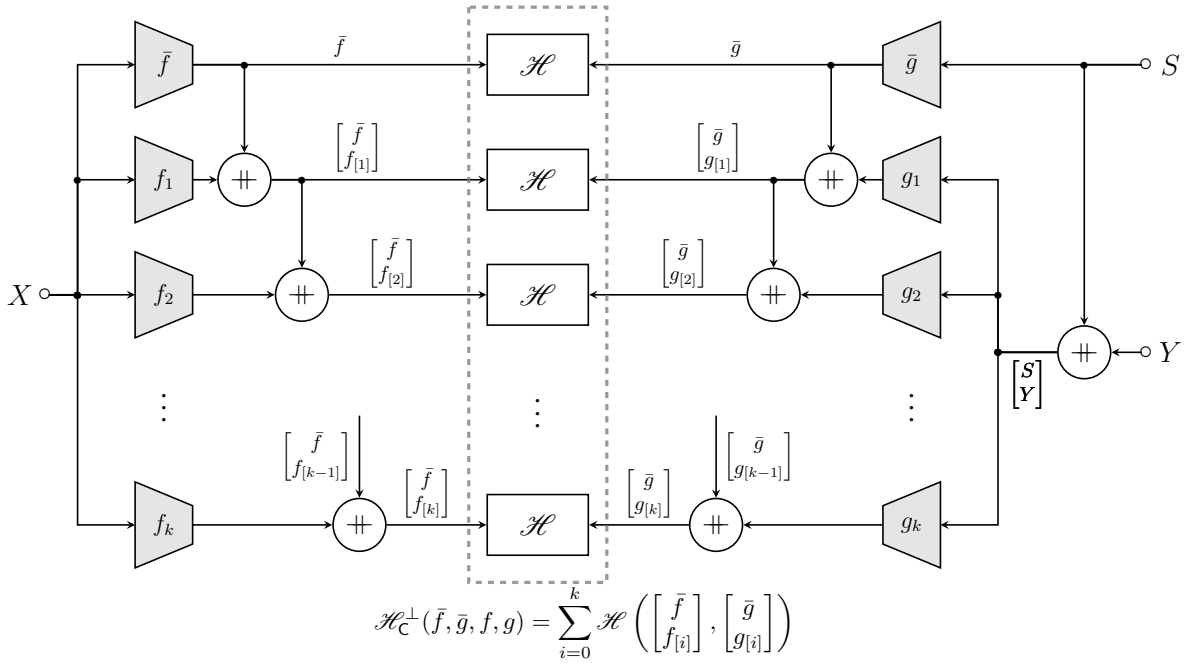

Fig. 2. Nested neural network structure for learning the features f_i^*, g_i^*, $i = 1, \ldots, k$ [cf. (25)], where "$+\!\!+$" denotes the concatenation operation of two vectors. Each gray block represents a deep feature extractor with trainable parameters.

which is composed of the top-k modals of the conditional dependence. Then, from Corollary 3, we have

$$P_{Y|X,S}^{(k)} = P_{Y|S}(y|s) \cdot (1 + f(x) \otimes g(s,y)), \tag{32}$$

where f, g are k-dimensional features of X and (S, Y) learned from maximizing the nested H-score $\mathscr{H}_{\mathsf{C}}$ (or $\mathscr{H}_{\mathsf{C}}^{\perp}$).

Therefore, we can use (32) as the implementation for inference, where $P_{Y|S}$ can be obtained by a discriminative model that predicts Y from side information S.

APPENDIX

A. Proof of Proposition 1

Let us define $\tilde{\gamma} \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ as

$$\tilde{\gamma}(x,y) \triangleq \gamma(x,y) + \mathbb{E}_{R_{X,Y}}[\gamma(X,Y)] - \mathbb{E}_{R_Y}[\gamma(x,Y)] - \mathbb{E}_{R_X}[\gamma(X,y)].$$

It can be verified that $\tilde{\gamma}$ is centered, i.e., for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we have $\mathbb{E}_{R_X}[\tilde{\gamma}(X,y)] + \mathbb{E}_{R_Y}[\tilde{\gamma}(x,Y)] = 0$, which is also equivalent to $\mathbb{E}_{R_X}[\tilde{\gamma}(X,y)] = \mathbb{E}_{R_Y}[\tilde{\gamma}(x,Y)] = 0$.

Then, from the Schmidt decomposition [11] of $\tilde{\gamma}$, there exist $K \geq 0$, $\sigma_1 \geq \cdots \geq \sigma_K > 0$, $f_1^*, \ldots, f_K^* \in \mathcal{F}_{\mathcal{X}}$, and $g_1^*, \ldots, g_K^* \in \mathcal{F}_{\mathcal{Y}}$, such that $\langle f_i^*, f_j^* \rangle = \langle g_i^*, g_j^* \rangle = \delta_{ij}$ for all $i, j = 1, \ldots, K$ and

$$\tilde{\gamma}(x,y) = \sum_{i=1}^{K} \sigma_i f_i^*(x) g_i^*(y). \tag{33}$$

It remains only to show that $f_i^* \in \mathcal{F}_{\mathcal{X}|\varnothing}, g_i^* \in \mathcal{F}_{\mathcal{Y}|\varnothing}$, $i = 1, \ldots, K$. To see this, note that from (33) we have, for each $i = 1, \ldots, K$,

$$\begin{aligned}\mathbb{E}_{R_X R_Y}[\tilde{\gamma}(X,Y) g_i^*(Y)] &= \sum_{j=1}^{K} \sigma_j \cdot \mathbb{E}_{R_X}\left[f_j^*(X)\right] \cdot \langle g_j^*, g_i^* \rangle \\ &= \sum_{j=1}^{K} \sigma_j \cdot \mathbb{E}_{R_X}\left[f_j^*(X)\right] \cdot \delta_{ij} \\ &= \sigma_i \cdot \mathbb{E}_{R_X}\left[f_i^*(X)\right].\end{aligned}$$

In addition, note that

$$\begin{aligned}\mathbb{E}_{R_X R_Y}[\tilde{\gamma}(X,Y) g_i^*(Y)] &= \sum_{y \in \mathcal{Y}} R_Y(y) g_i^*(y) \cdot \mathbb{E}_{R_X}[\tilde{\gamma}(X,y)] \\ &= 0.\end{aligned}$$

Therefore, we obtain $\mathbb{E}_{R_X}[f_i^*(X)] = 0$, i.e., $f_i^* \in \mathcal{F}_{\mathcal{X}|\varnothing}$. Similarly, we have $g_i^* \in \mathcal{F}_{\mathcal{Y}|\varnothing}$, for all $i = 1, \ldots, K$. □

B. Proof of Property 1

We first establish (8). For all $i = 1, \ldots, K$, since $\langle f_i, f_j^* \rangle = 0$ for $j = 1, \ldots, i-1$, from (6) we have

$$\tau(f_i) = \sum_{j=i}^{K} (\sigma_i \langle f_i, f_i^* \rangle) \cdot g_i^*.$$

Therefore, we have

$$\begin{aligned}\left\|\tau(f_i)\right\|^2 = \sum_{j=i}^{K} \left(\sigma_j \langle f_i, f_j^* \rangle\right)^2 &\leq \sigma_i^2 \sum_{j=i}^{K} \langle f_i, f_j^* \rangle^2 \\ &\leq \sigma_i^2 \|f_i\|^2 = \sigma_i^2,\end{aligned} \tag{34}$$

where we have used the fact that $\langle f_i^*, f_j^* \rangle = \langle g_i^*, g_j^* \rangle = \sigma_{ij}$, and where all inequalities hold with equality if $f_i = f_i^*$.

Similarly, to establish (9), note that for all $i = 1, \ldots, K$ we have

$$\begin{aligned}\langle \gamma, f_i \otimes g_i \rangle = \langle \tau(f_i), g_i \rangle &\leq \|\tau(f_i)\| \cdot \|g_i\| \\ &\leq \sigma_i \cdot \|f_i\| \cdot \|g_i\| \\ &= \sigma_i,\end{aligned}$$

where the first equality follows from Cauchy–Schwarz inequality, and the second inequality follows from (34). In addition, it can be verified that all inequalities hold with equality if $f_i = f_i^*$ and $g_i = g_i^*$. □

C. Proof of Proposition 2

The relation $\tilde{\ell}_{P_{X,S}} = \tilde{\ell}_{P^{(\mathsf{M})}_{X,S,Y}}$ can be directly verified from definition. To establish $\pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}}) = \Pi\left(\tilde{\ell}_{P_{X,S,Y}}; \mathcal{F}_{\mathcal{X}\times\mathcal{S}}\right) = \tilde{\ell}_{P_{X,S}}$, from Fact 2, it suffices to show that $(\tilde{\ell}_{P_{X,S,Y}} - \tilde{\ell}_{P_{X,S}}) \perp \mathcal{F}_{\mathcal{X}\times\mathcal{S}}$.

To this end, note that since

$$\begin{aligned}
&\tilde{\ell}_{P_{X,S,Y}}(x,s,y) - \tilde{\ell}_{P_{X,S}}(x,s) \\
&\quad = \frac{P_{X,S,Y}(x,s,y) - P^{(\mathsf{M})}_{X,S,Y}(x,s,y)}{R_{X,S,Y}(x,s,y)} \\
&\quad = \frac{P_{X,S,Y}(x,s,y) - P_{X|S}(x|s)P_S(s)P_{Y|S}(y|s)}{R_{X,S,Y}(x,s,y)},
\end{aligned}$$

for all $f \in \mathcal{F}_{\mathcal{X}\times\mathcal{S}}$, we have

$$\begin{aligned}
&\left\langle \tilde{\ell}_{P_{X,S,Y}} - \tilde{\ell}_{P_{X,S}}, f \right\rangle \\
&\quad = \sum_{x\in\mathcal{X}, s\in\mathcal{S}, y\in\mathcal{Y}} P_{X,S,Y}(x,s,y) f(x,s) \\
&\qquad - \sum_{x\in\mathcal{X}, s\in\mathcal{S}, y\in\mathcal{Y}} P_{X|S}(x|s)P_S(s)P_{Y|S}(y|s)\cdot f(x,s) \\
&\quad = \mathbb{E}_{P_{X,S}}[f(X,S)] - \mathbb{E}_{P_{X,S}}[f(X,S)] \\
&\quad = 0.
\end{aligned}$$

Finally, the expression of $\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right)$ can be obtained by using the fact $\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right) = \tilde{\ell}_{P_{X,S,Y}} - \pi_{\mathsf{M}}(\tilde{\ell}_{P_{X,S,Y}}) = \tilde{\ell}_{P_{X,S,Y}} - \tilde{\ell}_{P_{X,S}}$. □

D. Proof of Theorem 1

Suppose the modal decomposition of $\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right)$ is

$$\left[\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right)\right](x,s,y) = \sum_{i=1}^{K} \sigma_i f_i^*(x) g_i^*(s,y).$$

We then prove that $g_i^* \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}, i = 1,\ldots,K$. To this end, it suffices to show that for each $g \in \mathcal{F}_{\mathcal{S}}$, we have $\langle g_i^*, g\rangle = 0$ for each $i = 1,\ldots,K$. Indeed, since $\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right) \in \mathcal{F}_{\mathcal{Y}|(\mathcal{X}\times\mathcal{S})}$, we have

$$\begin{aligned}
0 = \left\langle f_i^* \otimes g, \pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right)\right\rangle &= \sum_{j=1}^{K} \sigma_j \cdot \langle f_i^*, f_j^*\rangle \cdot \langle g, g_j^*\rangle \\
&= \sigma_i \cdot \langle g, g_i^*\rangle.
\end{aligned}$$

Therefore, for all $f \in \mathcal{F}_{\mathcal{X}}$, we have

$$\tau_{\mathsf{C}}(f) = \sum_{i=1}^{K} \left(\sigma_i \langle f, f_i^*\rangle\right) \cdot g_i^* \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}.$$

Similarly, suppose the modal decomposition of $\tilde{\ell}_{P_{X,S}}$ is

$$\tilde{\ell}_{P_{X,S}} = \sum_{i=1}^{\bar{K}} \bar{\sigma}_i \left(\bar{f}_i^* \otimes \bar{g}_i^*\right),$$

where $\bar{f}_1^*,\ldots,\bar{f}_{\bar{K}}^* \in \mathcal{F}_{\mathcal{X}}$, $\bar{g}_1^*,\ldots,\bar{g}_{\bar{K}}^* \in \mathcal{F}_{\mathcal{S}}$. Then, for all $f \in \mathcal{F}_{\mathcal{X}}$, we have

$$\tau_{\mathsf{M}}(f) = \sum_{i=1}^{\bar{K}} \left(\bar{\sigma}_i \langle f, \bar{f}_i^*\rangle\right) \cdot \bar{g}_i^* \in \mathcal{F}_{\mathcal{S}}.$$

Hence, for each $f \in \mathcal{F}_{\mathcal{X}}$, we have $\tau_{\mathsf{C}}(f) \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}, \tau_{\mathsf{M}}(f) \in \mathcal{F}_{\mathcal{S}}$, which implies that

$$\langle \tau_{\mathsf{C}}(f), \tau_{\mathsf{M}}(f)\rangle = 0. \tag{35}$$

As a result, (23) follows immediately from (22) and (35).

Finally, note that for all $f \in \mathcal{F}_{\mathcal{X}}$, we have

$$\begin{aligned}
\|\tau_{\mathsf{C}}^* \circ \tau_{\mathsf{M}}(f)\|^2 &= \|\tau_{\mathsf{C}}^*(\tau_{\mathsf{M}}(f))\|^2 \\
&= \langle \tau_{\mathsf{C}}^*(\tau_{\mathsf{M}}(f)), \tau_{\mathsf{C}}^*(\tau_{\mathsf{M}}(f))\rangle \\
&= \langle \tau_{\mathsf{M}}(f), \tau_{\mathsf{C}}(\tau_{\mathsf{C}}^*(\tau_{\mathsf{M}}(f)))\rangle \\
&= 0,
\end{aligned}$$

where the last equality follow from the facts that $\tau_{\mathsf{M}}(f) \in \mathcal{F}_{\mathcal{S}}$ and $\tau_{\mathsf{C}}(\tau_{\mathsf{C}}^*(\tau_{\mathsf{M}}(f))) \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$.

Hence, we have $\tau_{\mathsf{C}}^* \circ \tau_{\mathsf{M}} = 0$. Similarly, we can show that $\tau_{\mathsf{M}}^* \circ \tau_{\mathsf{C}} = 0$. □

E. Proof of Theorem 2

We first introduce a useful lemma.

Lemma 3. *For all $\bar{k}, k \geq 1$ and $\bar{f}\colon \mathcal{X} \to \mathbb{R}^{\bar{k}}, \bar{g}\colon \mathcal{S} \to \mathbb{R}^{\bar{k}}$, $f\colon \mathcal{X} \to \mathbb{R}^k, g\colon \mathcal{Y}\times\mathcal{S} \to \mathbb{R}^k$, we have*

$$\begin{aligned}
\mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right) = \frac{1}{2}\cdot\Big(&\left\|\tilde{\ell}_{P_{X,S,Y}}\right\|^2 \\
&- \left\|\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h}\right\|^2 \\
&- \left\|\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right) - f\otimes h\right\|^2\Big),
\end{aligned}$$

where we have defined $\bar{h} \triangleq (\bar{h}_1,\ldots,\bar{h}_k)^{\mathrm{T}}$, $h \triangleq (h_1,\ldots,h_k)^{\mathrm{T}}$, and where for each $i = 1,\ldots,k$, $\bar{h}_i$ and h_i are defined as $\bar{h}_i \triangleq \Pi(g_i; \mathcal{F}_{\mathcal{S}})$ and $h_i \triangleq \Pi\left(g_i; \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\right)$, respectively.

Proof of Lemma 3. From the definition of H-score, we have

$$\begin{aligned}
&\mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right) \\
&\quad = \frac{1}{2}\cdot\left(\left\|\tilde{\ell}_{P_{X,S,Y}}\right\|^2 - \left\|\tilde{\ell}_{P_{X,S,Y}} - \bar{f}\otimes\bar{g} - f\otimes g\right\|^2\right).
\end{aligned}$$

Therefore, it suffices to prove that

$$\begin{aligned}
&\left\|\tilde{\ell}_{P_{X,S,Y}} - \bar{f}\otimes\bar{g} - f\otimes g\right\|^2 \\
&\quad = \left\|\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h}\right\|^2 + \left\|\pi_{\mathsf{C}}\left(\tilde{\ell}_{P_{X,S,Y}}\right) - f\otimes h\right\|^2.
\end{aligned} \tag{36}$$

From $\mathcal{F}_{\mathcal{S}\times\mathcal{Y}} = \mathcal{F}_{\mathcal{S}} \boxplus \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$, for each $i = 1,\ldots,k$, we have $g_i = \bar{h}_i + h_i$. Therefore, we obtain $g = \bar{h} + h$ and $f\otimes g = f\otimes\bar{h} + f\otimes h$. We then verify that $f\otimes h \in \mathcal{F}_{\mathcal{Y}|(\mathcal{X}\times\mathcal{S})}$. To see this, for each $\hat{f} \in \mathcal{F}_{\mathcal{X}\times\mathcal{S}}$, note that

$$\begin{aligned}
&\langle f\otimes h, \hat{f}\rangle \\
&\quad = \sum_{i=1}^{k}\sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}, y\in\mathcal{Y}} P_X(x)P_{S,Y}(s,y)f_i(x)h_i(y,s)\hat{f}(x,s) \\
&\quad = \sum_{s\in\mathcal{S}, y\in\mathcal{Y}} P_{S,Y}(s,y)h_i(y,s)\cdot\left(\sum_{x\in\mathcal{X}} P_X(x)f_i(x)\hat{f}(x,s)\right) \\
&\quad = \sum_{s\in\mathcal{S}, y\in\mathcal{Y}} P_{S,Y}(s,y)h_i(y,s)\cdot\mathbb{E}_{P_X}\left[f_i(X)\hat{f}(X,s)\right] \\
&\quad = \sum_{s\in\mathcal{S}, y\in\mathcal{Y}} P_{S,Y}(s,y)h_i(y,s)t_i(s) \\
&\quad = \sum_{i=1}^{k}\langle h_i, t_i\rangle = 0,
\end{aligned}$$

where we have defined $t_i\colon s \mapsto \mathbb{E}_{P_X}\left[f_i(X)\hat{f}(X,s)\right]$, and where the last equality follows from the fact that $h_i \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$ and $t_i \in \mathcal{F}_{\mathcal{S}}$.

Finally, to establish (36), note that

$$\begin{aligned}
&\left\|\tilde{\ell}_{P_{X,S,Y}} - \bar{f}\otimes\bar{g} - f\otimes g\right\|^2 \\
&\quad= \Big\|\pi_{\mathsf{M}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) + \pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) \\
&\qquad - \bar{f}\otimes\bar{g} - f\otimes\bar{h} - f\otimes h\Big\|^2 \\
&\quad= \Big\|\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h} + \pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) - f\otimes h\Big\|^2 \\
&\quad= \Big\|\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h}\Big\|^2 + \Big\|\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) - f\otimes h\Big\|^2,
\end{aligned}$$

where to obtain the last equality we have used the fact that $\big\langle\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h}, \pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) - f\otimes h\big\rangle = 0$ since $\big(\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g} - f\otimes\bar{h}\big) \in \mathcal{F}_{\mathcal{X}\times\mathcal{S}}$ and $\big(\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big) - f\otimes h\big) \in \mathcal{F}_{\mathcal{Y}|(\mathcal{X}\times\mathcal{S})}$. □

Proceeding to our proof of Theorem 2, note that from (13) we have

$$\mathscr{H}(\bar{f},\bar{g}) = \frac{1}{2}\left(\left\|\tilde{\ell}_{P_{X,S}}\right\|^2 - \left\|\tilde{\ell}_{P_{X,S}} - \bar{f}\otimes\bar{g}\right\|^2\right). \tag{37}$$

Therefore, from (2), $\mathscr{H}(\bar{f},\bar{g})$ is maximized if and only if

$$\bar{f}\otimes\bar{g} = \tilde{\ell}_{P_{X,S}}. \tag{38}$$

Similarly, from Lemma 3, $\mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right)$ is maximized if and only if

$$\bar{f}\otimes\bar{g} + f\otimes\bar{h} = \tilde{\ell}_{P_{X,S}} \tag{39}$$

$$f\otimes h = \sum_{i=1}^{k}\sigma_i(f_i^*\otimes g_i^*), \tag{40}$$

where $\bar{h} \triangleq (\bar{h}_1,\ldots,\bar{h}_k)^{\mathrm{T}}$, $h \triangleq (h_1,\ldots,h_k)^{\mathrm{T}}$, and where for each $i = 1,\ldots,k$, $\bar{h}_i$ and h_i are defined as $\bar{h}_i \triangleq \Pi\left(g_i;\mathcal{F}_{\mathcal{S}}\right)$ and $h_i \triangleq \Pi\left(g_i;\mathcal{F}_{\mathcal{Y}|\mathcal{S}}\right)$, respectively.

It is easy to verify that there exist $\bar{f},\bar{g},f,g$ such that (38)–(40) hold simultaneously. As a result, the two H-scores $\mathscr{H}(\bar{f},\bar{g})$ and $\mathscr{H}\left(\begin{bmatrix}\bar{f}\\ f\end{bmatrix}, \begin{bmatrix}\bar{g}\\ g\end{bmatrix}\right)$ are both maximized if and only if we have (38)–(40).

We then show that (38)–(40) are equivalent to (27). To see this, first note that (38) is equivalent to (27a). Moreover, from (38) and (39) we have $f\otimes\bar{h} = 0$, and thus from (40) we obtain

$$f\otimes g = f\otimes h + f\otimes\bar{h} = f\otimes h = \sum_{i=1}^{k}\sigma_i(f_i^*\otimes g_i^*),$$

which is (27b). In addition, from (27b) we have $g_i \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$ and thus $\bar{h} = 0$.

Finally, since $\mathscr{H}_{\mathsf{C}}$ is a weighted sum of these two H-scores, $\mathscr{H}_{\mathsf{C}}$ is maximized if and only if we have (27). □

F. Proof of Theorem 3

Note that from Proposition 2, we have

$$\begin{aligned}
&P_{X,S,Y}(x,s,y) - P_{X|S}(x|s)P_S(s)P_{Y|S}(y|s) \\
&\qquad= P_X(x)P_{S,Y}(s,y)\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y),
\end{aligned}$$

which implies that

$$\begin{aligned}
&P_{Y|X,S}(y|x,s) \\
&= P_{Y|S}(y|s)\cdot\left(1 + \frac{P_X(x)P_S(s)}{P_{X,S}(x,s)}\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y)\right) \\
&= P_{Y|S}(y|s)\cdot\left(1 + \frac{1}{\ell_{P_{X,S}}(x,s)}\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y)\right) \\
&= P_{Y|S}(y|s)\cdot\left(1 + \left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y)\right) + o(\epsilon).
\end{aligned}$$

To obtain the last equality we have used the fact that

$$\begin{aligned}
&\frac{1}{\ell_{P_{X,S}}(x,s)}\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y) \\
&\quad= \frac{1}{1+\tilde{\ell}_{P_{X,S}}(x,s)}\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y) \\
&\quad= \Big(1-\tilde{\ell}_{P_{X,S}}(x,s)\Big)\cdot\left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y) + o(\epsilon), \\
&\quad= \left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y) + o(\epsilon),
\end{aligned}$$

since [cf. (14)]

$$\tilde{\ell}_{P_{X,S}}(x,s) = O(\epsilon), \quad \left[\pi_{\mathsf{C}}\big(\tilde{\ell}_{P_{X,S,Y}}\big)\right](x,s,y) = O(\epsilon).$$

Finally, (30) is obtained by using the decomposition (25). □

REFERENCES

[1] D. Storcheus, A. Rostamizadeh, and S. Kumar, "A survey of modern questions and challenges in feature extraction," in *Feature Extraction: Modern Questions and Challenges*. PMLR, 2015, pp. 1–18.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[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] X. Xu, S.-L. Huang, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," *Entropy*, vol. 24, no. 1, p. 135, 2022.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[7] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–33, 2020.

[8] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5281–5288.

[9] X. Xu and S.-L. Huang, "Maximal correlation regression," *IEEE Access*, vol. 8, pp. 26 591–26 601, 2020.

[10] A. Makur and L. Zheng, "Polynomial spectral decomposition of conditional expectation operators," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 633–640.

[11] E. Schmidt, "Zur theorie der linearen und nichtlinearen integralgleichungen. i. teil: Entwicklung willkürlicher funktionen nach systemen vorgeschriebener," *Mathematische Annalen*, vol. 63, pp. 433–476, 1907.

[12] A. Ekert and P. L. Knight, "Entangled quantum systems and the schmidt decomposition," *American Journal of Physics*, vol. 63, no. 5, pp. 415–423, 1995.

[13] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[14] V. V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1567–1580, 2004.

[15] H. O. Hirschfeld, "A connection between correlation and contingency," in *Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4, 1935, pp. 520–524.

[16] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

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