Sequential Dependence Decomposition and Feature Learning

Xiangxiang Xu[1] and Lizhong Zheng[1]

Abstract— **In this paper, we introduce an approach to decompose statistical dependence and learn informative features from sequential data. We first present a sequential decomposition of dependence, which extends the chain rule of mutual information. To learn this decomposition from data, we investigate the optimal feature representations associated with decomposed dependence modes and develop corresponding learning algorithms. Specifically, for stationary processes, we demonstrate applications of the learned features in computing optimal Markov approximation and testing the order of Markov processes.**

I. Introduction

Many learning applications involve data points collected presented in a meaningful order, such as natural language texts, DNA sequences, and time series. To effectively process such data, many algorithms have been proposed for exploiting the sequential structure in the data. One classical approach is based on probabilistic modeling of the data, as Markov model or hidden Markov models [1], [2], and then estimate model parameters. In contrast to the explicit modeling, another common practice is designing neural networks with new architectures to address the sequential structure, of which most notable examples include recurrent neural network [3], LSTM (long short-term memory) network [4], and the recent transformer model [5]. With the advance of deep learning techniques, these deep models have demonstrates the effectiveness in processing high-dimensional sequential data, particularly for natural languages. Despite of their empirical successes, the connection between learned models and the statistical dependence of data is unclear, making it difficult to analyze the learning processes and trained models.

On the other hand, theoretical analyses of sequential dependence focus on simple probability models with known forms, e.g., the discrete information source model [6] and the dependence decomposition framework with information geometry [7]. However, practical sequential data are typically structured and high-dimensional, with unknown probability distributions, limiting the application of such analyzes. Recent developments in information-theoretic and statistical analyzes have established a connection between statistical dependence and feature space, which bridges dependence structure and feature learning [8], [9]. Specifically, comprehensive discussions on statistical dependence for two variables have been presented in [8], [9], and were further extended to a trivariate setting for learning conditional dependence components [10].

In this paper, our goal is to analyze sequential dependence and designing algorithms to effectively learn such dependence from data. To this end, we first extend the feature learning framework introduced in [10] to the sequential setup, where we present a sequential decomposition of dependence. We show that this decomposition can be interpreted as an generalization of the chain rule of mutual information. We further introduce the informative features to represent decomposed dependence components, and design learning algorithms by applying the nesting technique proposed in [10]. In particular, the resulting learning objective is a nested H-score for sequential data, which is an extension of both the original H-score [11], [9], and the nested H-score for learning conditional dependence [10]. We further discuss the application of the learning algorithms, by considering feature learning from stationary processes. Specifically, we demonstrate that the learned features can be used in computing Markov approximations, and decide the order of Markov processes.

II. Notations and Preliminaries

A. Feature Geometry

1) Vector Space: 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 $\Pi\left(v; \mathcal{W}\right) \triangleq \operatorname*{arg\,min}_{w \in \mathcal{W}} \|v - w\|^2$. In addition, we use $\mathcal{V} \boxminus \mathcal{W}$ to denote the orthogonal complement of $\mathcal{W}$ in $\mathcal{V}$, viz., $\mathcal{V} \boxminus \mathcal{W} \triangleq \{v \in \mathcal{V} \colon \langle v, w \rangle = 0$ for all $w \in \mathcal{W}\}$.*

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 $\mathcal{V}_1 \perp \mathcal{V}_2$. Therefore, we have $\mathcal{V}_2 = \mathcal{V} \boxminus \mathcal{V}_1$ if $\mathcal{V} = \mathcal{V}_1 \boxplus \mathcal{V}_2$. In addition, if $\mathcal{W}$ is a subspace of $\mathcal{V}$, then $\mathcal{V} = \mathcal{W} \boxplus (\mathcal{V} \boxminus \mathcal{W})$.

Fact 1 (Orthogonality Principle)**.** *Given $v \in \mathcal{V}$ and a subspace $\mathcal{W}$ of $\mathcal{V}$, then $w = \Pi\left(v; \mathcal{W}\right)$ if and only if $w \in \mathcal{W}$ and $v - w \in \mathcal{V} \boxminus \mathcal{W}$.*

2) Feature 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 $\operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ 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.

For given $\mathcal{Z}$, we use $\mathcal{F}_{\mathcal{Z}} \triangleq \{\mathcal{Z} \to \mathbb{R}\}$ to denote the collection of features (functions) of $\mathcal{Z}$. For each $k \geq 1$, we

*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

use $\mathcal{F}_{\mathcal{Z}}^{k} \triangleq (\mathcal{F}_{\mathcal{Z}})^{k} = \{\mathcal{Z} \to \mathbb{R}^{k}\}$ to denote the collection of k-dimensional features. Specifically, we use $\mathcal{F}_{\varnothing}$ to represent the collection of constant features. Given $f_1, \dots, f_k \in \mathcal{F}_{\mathcal{Z}}$, for each $i = 1, \dots, k$, we denote the multidimensional feature $\left(z \mapsto (f_1(z), \dots, f_i(z))^{\mathrm{T}}\right) \in \mathcal{F}_{\mathcal{X}}^{i}$ 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 $R \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$, we define the inner product on $\mathcal{F}_{\mathcal{Z}}$ as $\langle f_1, f_2 \rangle \triangleq \mathbb{E}_R\left[f_1(Z)f_2(Z)\right]$, 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 $R_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ 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 *product functions*. Suppose $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ is a product function, there exist $f \in \mathcal{F}_{\mathcal{X}}, g \in \mathcal{F}_{\mathcal{Y}}$ and $\sigma \geq 0$ such that $\gamma = \sigma \cdot (f \otimes g)$, with $\sigma = \|\gamma\|$ and $\|f\| = \|g\| = 1$. We will refer to $\sigma \cdot (f \otimes g)$ as the standard form of γ. In addition, for given $f_1, \dots, f_k \in \mathcal{F}_{\mathcal{X}}$ and $g_1, \dots, g_k \in \mathcal{F}_{\mathcal{Y}}$, 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 denote the resulting orthogonal complement as $\mathcal{F}_{\mathcal{Y}|\mathcal{X}} \triangleq \mathcal{F}_{\mathcal{X}\times\mathcal{Y}} \boxminus \mathcal{F}_{\mathcal{X}}$.

We can establish a correspondence between distribution space and feature space by the density ratio functions.

Definition 2. *Given a metric distribution $R \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$, for each $P \in \mathcal{P}^{\mathcal{Z}}$, we define the (centered) density ratio function $\tilde{\ell}_{P;R} \in \mathcal{F}_{\mathcal{Z}}$ as*

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

When there is no ambiguity about metric distribution R, we will simply use $\tilde{\ell}_{P}$ to denote $\tilde{\ell}_{P;R}$. In particular, for given $(X, Y) \sim P_{X,Y}$, we will use $\mathfrak{i}_{X;Y} \triangleq \tilde{\ell}_{P_{X,Y};P_X P_Y}$ to denote the induced density ratio, and refer to it as the *canonical dependence kernel* (CDK) function. The CDK function $\mathfrak{i}_{X;Y}$ is also related to the information density $\tilde{\imath}_{X;Y}(x, y) \triangleq \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$, where we have $\tilde{\imath}_{X;Y} = \log(1 + \mathfrak{i}_{X;Y})$.

B. Modal Decomposition

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., $R_{X,Y} = R_X R_Y$.

We define the operator ζ on $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ as the optimal rank-1 approximation, i.e.,

$$\zeta(\gamma) \triangleq \operatorname*{arg\,min}_{\substack{\gamma'\colon \gamma' = f \otimes g \\ f \in \mathcal{F}_{\mathcal{X}}, g \in \mathcal{F}_{\mathcal{Y}}}} \|\gamma - \gamma'\|, \quad \text{for all } \gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}. \tag{1}$$

In addition, for all $k \geq 1$, we define the operator ζ_k as $\zeta_1 \triangleq \zeta$, and $\zeta_k(\gamma) \triangleq \zeta\left(\gamma - \sum_{i=1}^{k-1} \zeta_i(\gamma)\right)$, which we refer to as the k-th mode of γ. Then, we use $\zeta_{\leq k}(\gamma) \triangleq \sum_{i=1}^{k} \zeta_i(\gamma)$ and $r_k(\gamma) \triangleq \gamma - \zeta_{\leq k}(\gamma)$ to denote the superposition of the top k modes and the corresponding remainder, respectively.

For each $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$, we define the rank of γ as $\operatorname{rank}(\gamma) \triangleq \inf\{k \geq 0\colon \|r_k(\gamma)\| = 0\}$. Let $K \triangleq \operatorname{rank}(\gamma)$, and suppose $\zeta_i(\gamma) = \sigma_i\left(f_i^* \otimes g_i^*\right)$ denotes the standard form of $\zeta_i(\gamma)$ for $i = 1, \dots, K$. Then, we obtain

$$\gamma(x, y) = \sum_{i=1}^{K} \sigma_i \cdot f_i^*(x) g_i^*(y) \tag{2}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, where $\|f_i^*\| = \|g_i^*\| = 1$ and $\sigma_i = \|\zeta_i(\gamma)\|$. We refer to (2) as the modal decomposition of γ, which can be regarded as a special case of Schmidt decomposition [12], or singular value decomposition (SVD) in functional space. Specifically, we have the following property.

Property 1. *Let $K \triangleq \operatorname{rank}(\gamma)$, then $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K > 0$. In addition, for all $i, j \in [K]$, we have[1] $\langle f_i^*, f_j^* \rangle = \langle g_i^*, g_j^* \rangle = \delta_{ij}$.*

Moreover, we can express Eckart–Young–Mirsky theorem [13] as follows.

Property 2. *For all $\gamma \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ and $k \geq 1$, we have*

$$\zeta_{\leq k}(\gamma) = \operatorname*{arg\,min}_{\gamma'\colon \operatorname{rank}(\gamma') \leq k} \|\gamma - \gamma'\| = \operatorname*{arg\,min}_{\substack{\gamma'\colon \gamma' = f \otimes g, \\ f \in \mathcal{F}_{\mathcal{X}}^{k}, g \in \mathcal{F}_{\mathcal{Y}}^{k}}} \|\gamma - \gamma'\|.$$

Therefore, we will refer to $\zeta_{\leq k}(\gamma)$ as the rank-k approximation of γ, with the remainder $r_k(\gamma)$ representing the approximation error. Then, we have $\gamma = \zeta_{\leq k}(\gamma) + r_k(\gamma) = \sum_{i=1}^{k} \zeta_i(\gamma) + r_k(\gamma)$.

The following variational characterization will also be useful.

Property 3. *For all $i = 1, \dots, \operatorname{rank}(\gamma)$, we have $(f_i^*, g_i^*) = \operatorname*{arg\,max}_{f_i, g_i} \langle \gamma, f_i \otimes g_i \rangle$ 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 $j = 1, \dots, i - 1$.*

C. Decomposition of Statistical Dependence

1) Bivariate Case: For any joint distribution $P_{X,Y}$, with the metric distribution $R_{X,Y} = P_X P_Y$, the modal decomposition of $\mathfrak{i}_{X,Y}$ is also referred to as the modal decomposition of the joint distribution $P_{X,Y}$ (cf. [8, Proposition 2]). Let the decomposition be

$$\zeta_i(\mathfrak{i}_{X;Y}) = \sigma_i \cdot (f_i^* \otimes g_i^*), \quad i = 1, \dots, K, \tag{3}$$

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 [14], [15], [16]. Specifically, for given $f \in \mathcal{F}_{\mathcal{X}}, g \in \mathcal{F}_{\mathcal{Y}}$, let us denote their covariance as

$$\operatorname{cov}(f, g) \triangleq \mathbb{E}_{P_{X,Y}}\left[f(X)g(Y)\right] - \mathbb{E}_{P_X P_Y}\left[f(X)g(Y)\right]. \tag{4}$$

[1] We adopt the Kronecker delta notation

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

Note that since $\operatorname{cov}(f,g) = \langle \mathfrak{i}_{X;Y}, f \otimes g\rangle$, we have the following corollary of Property 3.

Corollary 1 (HGR Maximal Correlation Functions)**.** *For all $i = 1, \dots, K$, we have $\sigma_i = \operatorname{cov}(f_i^*, g_i^*) = \mathbb{E}_{P_{X,Y}}\left[f_i^*(X) g_i^*(Y)\right]$ and $(f_i^*, g_i^*) = \underset{f_i, g_i}{\arg\max} \operatorname{cov}(f_i, g_i)$, where the maximization is taken over all $f_i \in \mathcal{F}_{\mathcal{X}}$ and $g_i \in \mathcal{F}_{\mathcal{Y}}$ with $\|f_i\| = \|g_i\| = 1$ and $\langle f_i, f_j^* \rangle = \langle g_i, g_j^*\rangle = 0$ for $j = 1, \dots, i-1$.*

In addition, (3) can also be illustrated as a decomposition of mutual information.

Lemma 1 ([8, Lemma 16])**.** *If $\left\|\mathfrak{i}_{X,Y}\right\| \leq \epsilon$, then*

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

where $K = \operatorname{rank}(\mathfrak{i}_{X;Y})$, and $\sigma_i = \|\,\zeta_i(\mathfrak{i}_{X;Y})\|$.

The maximal correlation functions can be effectively extracted from data via maximizing the H-score [9], [11], and its nested variant [10].

Definition 3. *Given $k \geq 1$ and $f \in \mathcal{F}_{\mathcal{X}}^k$, $g \in \mathcal{F}_{\mathcal{Y}}^k$, the H-score $\mathscr{H}(f,g)$ is defined as*

$$\begin{aligned}\mathscr{H}(f,g) &\triangleq \mathbb{E}\left[f^{\mathrm{T}}(X) g(Y)\right] - \left(\mathbb{E}\left[f(X)\right]\right)^{\mathrm{T}} \mathbb{E}\left[g(Y)\right] \\ &\quad - \frac{1}{2} \cdot \operatorname{tr}\left(\Lambda_f \Lambda_g\right),\end{aligned} \tag{5}$$

where $\Lambda_f \triangleq \mathbb{E}\left[f(X) f^{\mathrm{T}}(X)\right]$, $\Lambda_g \triangleq \mathbb{E}\left[g(Y) g^{\mathrm{T}}(Y)\right]$. The nested H-score[2] $\mathscr{H}^\star(f,g)$ is then defined as

$$\mathscr{H}^\star(f,g) \triangleq \sum_{i=1}^{k} \mathscr{H}(f_{[i]}, g_{[i]}). \tag{6}$$

From the relation

$$\mathscr{H}(f,g) = \frac{1}{2}\left(\left\|\mathfrak{i}_{X;Y}\right\|^2 - \left\|\mathfrak{i}_{X,Y} - f \otimes g\right\|^2\right), \tag{7}$$

we can obtain the following corollary of Property 2.

Proposition 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 $f \otimes g = \zeta_{\leq k}(\mathfrak{i}_{X;Y})$, and $\mathscr{H}^\star(f,g)$ is maximized if and only if $f_i \otimes g_i = \zeta_i(\mathfrak{i}_{X;Y})$, for all $i = 1, \dots, k$.*

2) Trivariate Case: We briefly summarize trivariate dependence decomposition results established in [10]. Given $(X, S, Y) \sim P_{X,S,Y}$, we set $R_{X,S,Y} = P_X R_{S,Y}$ as the metric distribution, and define the conditional CDK function $\mathfrak{i}_{X;Y|S} \in \mathcal{F}_{\mathcal{X} \times \mathcal{S} \times \mathcal{Y}}$ as, for each $x \in \mathcal{X}, s \in \mathcal{S}, y \in \mathcal{Y}$,

$$\mathfrak{i}_{X;Y|S}(x,s,y) \triangleq \mathfrak{i}_{X;S,Y}(x,s,y) - \mathfrak{i}_{X;S}(x,s), \tag{8}$$

which is analogous to[3] the conditional information density

$$\begin{aligned}\tilde{\imath}_{X;Y|S}(x,y|s) &\triangleq \log \frac{P_{X,Y|S}(x,y|s)}{P_{X|S}(x|s) P_{Y|S}(y|s)} \\ &= \tilde{\imath}_{X;S,Y}(x,s,y) - \tilde{\imath}_{X;S}(x,s).\end{aligned} \tag{9}$$

Then, we have the following proposition.

Proposition 2. *We have $\mathfrak{i}_{X;S} = \Pi\left(\mathfrak{i}_{X;S,Y}; \mathcal{F}_{\mathcal{X}\times\mathcal{S}}\right), \mathfrak{i}_{X;Y|S} \in \mathcal{F}_{\mathcal{Y}|\mathcal{X}\times\mathcal{S}}$ and $\|\mathfrak{i}_{X;S,Y}\|^2 = \|\mathfrak{i}_{X;Y|S}\|^2 + \|\mathfrak{i}_{X;S}\|^2$. Specifically, if $\|\mathfrak{i}_{X;S,Y}\| = O(\epsilon)$, then $I(X;Y|S) = \frac{1}{2} \cdot \|\mathfrak{i}_{X;Y|S}\|^2 + o(\epsilon^2)$.*

The following corollary is straightforward.

Corollary 2. *The following statements are equivalent:*

- *$\|\mathfrak{i}_{X;S,Y}\| = \|\mathfrak{i}_{X;S}\|$;*
- *X, S, Y follow the Markov relation $X - S - Y$;*
- *$\|\mathfrak{i}_{X;Y|S}\| = 0$.*

III. Sequential Dependence Decomposition

We consider the dependence between Y and sequential data $(X_1, \dots, X_n)$, where Y can be regarded as the target variable. Without loss of generality, we assume X_i takes values from $\mathcal{X}_i$, for each $i \in [n]$, and Y takes values from $\mathcal{Y}$, where we have defined $[i : j] \triangleq \{l \in \mathbb{Z} \colon i \leq l \leq j\}$ and $[n] \triangleq [1 : n]$. For a subset $\mathcal{I}$ of $[n]$, it will be convenient to denote the sequence $x_{\mathcal{I}} \triangleq (x_i, i \in \mathcal{I})$, random variable $X_{\mathcal{I}}$ with the associated alphabet $\mathcal{X}_{\mathcal{I}} \triangleq \bigtimes_{i \in \mathcal{I}} \mathcal{X}_i$. Specifically, we also denote $x^l \triangleq x_{[l]} = (x_1, \dots, x_l)$.

We then devlop the decomposition of dependence between Y and sequence X^n, characterized by their joint distribution P_{Y,X^n}. To this end, we consider the function space $\mathcal{F}_{\mathcal{Y} \times \mathcal{X}_{[n]}}$ with the metric distribution $R_{Y,X^n} = P_Y P_{X^n}$. The inner product and geometry on $\mathcal{F}_{\mathcal{Y} \times \mathcal{X}_{[n]}}$ are defined accordingly. We chracterize our dependence decomposition as follows. A proof is provided in Appendix A.

Proposition 3. *For each $l \in [n]$, we have the orthogonal decomposition*

$$\mathfrak{i}_{Y;X^l} = \sum_{i=1}^{l} \mathfrak{i}_{Y;X_i|X^{i-1}}, \tag{10}$$

where $\langle \mathfrak{i}_{Y;X_i|X^{i-1}}, \mathfrak{i}_{Y;X_j|X^{j-1}}\rangle = 0$ for all $1 \leq i < j \leq l$, and where we have defined $X^0 \triangleq \varnothing$.

Note that Proposition 3 is an extension of Proposition 2 for trivariate case, which corresponds to the special case of $n = 2$. Analogous to Corollary 2, we can readily conclude the equivalence of (1) $\|\mathfrak{i}_{Y;X^i}\| = \|\mathfrak{i}_{Y;X^{i-1}}\|$, (2) the Markov relation $X_i - X^{i-1} - Y$, and (3) $\|\mathfrak{i}_{Y;X_i|X^{i-1}}\| = 0$, for all $i \in [2 : l]$. Therefore, $\mathfrak{i}_{Y;X_i|X^{i-1}}$ corresponds to the conditional dependence between Y and X_i conditional on X^{i-1}.

In particular, the orthogonality of (10) implies the Pythagorean relation

$$\|\mathfrak{i}_{Y;X^n}\|^2 = \sum_{i=1}^{n} \|\mathfrak{i}_{Y;X_i|X^{i-1}}\|^2, \tag{11}$$

which can be regarded as the counterpart of the chain rule

$$I(Y;X^n) = \sum_{i=1}^{n} I(Y;X_i|X^{i-1}). \tag{12}$$

To see the connection between (11) and (12), we consider a local regime with $\|\mathfrak{i}_{Y;X^n}\| = O(\epsilon)$. Then, for each $i \in [n]$, we have [cf. Lemma 1 and Proposition 2]

$$I(Y;X_i|X^{i-1}) = \frac{1}{2} \cdot \|\mathfrak{i}_{Y;X_i|X^{i-1}}\|^2 + o(\epsilon^2). \tag{13}$$

Therefore, with the decomposition (10), we can interpret the chain rule of mutual information (11) as a Pythagorean

[2] In general, we can define a nested H-score by aggregating different H-scores, where with input features form a nested structure. See [17, Section 4.1] for detailed discussions.

[3] Unlike the obvious symmetry $\tilde{\imath}_{X;Y|S}(x,y|s) = \tilde{\imath}_{Y;X|S}(y,x|s)$, the analogous relation $\mathfrak{i}_{X;Y|S}(x,s,y) = \mathfrak{i}_{Y;X|S}(y,s,x)$ does not hold for conditional CDK in general.

relation in functional space, where the conditional mutual information corresponds to the length of functions. As we will soon demonstrate, this functional view allows us to develop effective algorithms to analyze the dependence and learn informative features from data samples.

To establish the connection between dependence components and features, we consider the modal decomposition. In particular, let us define

$$K_i \triangleq \operatorname{rank}(\mathfrak{i}_{Y;X_i|X^{i-1}}) \quad \text{for all } i \in [n], \tag{14}$$

and suppose the modal decompositions of $\mathfrak{i}_{Y;X_i|X^{i-1}}$ is

$$\sigma_j^{[i]}\left(g_j^{*[i]} \otimes f_j^{*[i]}\right) = \zeta_j(\mathfrak{i}_{Y;X_i|X^{i-1}}), \quad j \in [K_i] \tag{15}$$

for each $i \in [n]$, where we have $f_j^{*[i]} \in \mathcal{F}_{\mathcal{X}_{[i]}}, g_j^{*[i]} \in \mathcal{F}_{\mathcal{Y}}$. Then, we can show that $g_j^{*[i]}, f_j^{*[i]}$ are the solutions of a constrained extension of the maximal correlation problem (cf. Corollary 1). A proof is provided in Appendix B.

Proposition 4. *For each $i \in [n]$, we have $\sigma_j^{[i]} = \operatorname{cov}(g_j^{*[i]}, f_j^{*[i]}) = \mathbb{E}_{P_{Y,X^i}}\left[g_j^{*[i]}(Y) f_j^{*[i]}(X^i)\right]$, $(g_j^{*[i]}, f_j^{*[i]}) = \underset{g_j, f_j}{\arg\max} \operatorname{cov}(g_j, f_j)$, where* cov *denotes the covariance [cf. (4)], and where the maximization is taken over all $g_j \in \mathcal{F}_{\mathcal{Y}}$ and $f_j \in \mathcal{F}_{\mathcal{X}_i|\mathcal{X}_{[i-1]}}$ that satisfy $\|g_j\| = \|f_j\| = 1$ and*

$$\left\langle g_j, g_l^{*[i]} \right\rangle = \left\langle f_j, f_l^{*[i]} \right\rangle = 0 \text{ for } l \in [j-1]. \tag{16}$$

The following corollary is immediate.

Corollary 3. *For all $i, i' \in [n]$ and $j, j' \in [K_i]$, we have*

$$\left\langle f_j^{*[i]}, f_{j'}^{*[i']} \right\rangle = \delta_{ii'}\delta_{jj'}. \tag{17}$$

Therefore, the modal decomposition (15) gives a collection of orthogonal features $\{f_j^{*[i]} : i \in [n], j \in [K_i]\}$ of X^n, where for each $i \in [n]$, the value $f^{*[i]}(x^n)$ depends only on x^i.

IV. Learning Algorithms

We then develop the algorithms for learning the conditional dependence $\mathfrak{i}_{Y;X_i|X^{i-1}}$ from data samples. To this end, we apply the nesting technique introduced in [10], and extend the H-scores [cf. (5) and (6)] to the sequential data.

Specifically, given feature dimensions $k_1, \dots, k_n$ and n feature pairs $\left\{\left(f^{[i]}, g^{[i]}\right)\right\}_{i=1}^{n}$ with $f^{[i]} \in \mathcal{F}_{\mathcal{X}_{[i]}}^{k_i}, g^{[i]} \in \mathcal{F}_{\mathcal{Y}}^{k_i}$ for each $i \in [n]$, we define

$$\mathscr{H}_{\mathsf{seq}} \triangleq \sum_{i=1}^{n} \mathscr{H}\left(f^{[[i]]}, g^{[[i]]}\right), \tag{18}$$

$$\mathscr{H}^{\star}_{\mathsf{seq}} \triangleq \mathscr{H}^{\star}\left(f^{[[n]]}, g^{[[n]]}\right), \tag{19}$$

as the H-score and its nested variant for sequential data, respectively, where for each $i \in [n]$, we have defined

$$f^{[[i]]} \triangleq f^{[1]} +\!\!+ \cdots +\!\!+ f^{[i]}, \tag{20a}$$

$$g^{[[i]]} \triangleq g^{[1]} +\!\!+ \cdots +\!\!+ g^{[i]}, \tag{20b}$$

where we have used the "$+\!\!+$" symbol to indicate concatenation of two vectors, i.e., 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.

It is worth noting that if $n = 1$, the H-scores $\mathscr{H}_{\mathsf{seq}}$ and $\mathscr{H}^{\star}_{\mathsf{seq}}$ reduce to the original H-scores for two variables [cf. (5) and (6)]; if $n = 2$, $\mathscr{H}_{\mathsf{seq}}$ and $\mathscr{H}^{\star}_{\mathsf{seq}}$ reduce to the H-scores introduced in [10] for learning conditional dependence of three variables. The following result demonstrates that we can effectively learn the dependence modes via maximizing corresponding H-scores. A proof is provided in Appendix C.

Theorem 1. *Suppose $k_i \geq K_i$ for each $i \in [n]$, with K_i as defined in (14). Then for all $f^{[i]} \in \mathcal{F}_{\mathcal{X}_{[i]}}^{k_i}$, and $g^{[i]} \in \mathcal{F}_{\mathcal{Y}}^{k_i}$, $\mathscr{H}_{\mathsf{seq}}$ is maximized if and only if*

$$g^{[i]} \otimes f^{[i]} = \mathfrak{i}_{Y;X_i|X^{i-1}}, \quad \text{for all } i \in [n], \tag{21}$$

and $\mathscr{H}^{\star}_{\mathsf{seq}}$ is maximized if and only if

$$g_j^{[i]} \otimes f_j^{[i]} = \zeta_j(\mathfrak{i}_{Y;X_i|X^{i-1}}) \tag{22}$$

for all $i \in [n]$, $j \in [k_i]$.

In practice, we can compute $\mathscr{H}_{\mathsf{seq}}$ using a nested neural network structure as shown in Fig. 1, where filled blocks are implemented by parameterized deep neural networks. Then, the corresponding features can be obtained by optimizing the parameters in networks to maximize the H-scores.

After learning the optimal features, we can readily obtain the magnitude of each dependence mode, by exploiting the following connection. A proof is provided in Appendix D.

Proposition 5. *For given $i \in [n]$ and k_i-dimensional features $g^{[i]} \in \mathcal{F}_{\mathcal{Y}}^{k_i}$, $f^{[i]} \in \mathcal{F}_{\mathcal{X}_{[i]}}^{k_i}$ with $g^{[i]} \otimes f^{[i]} = \mathfrak{i}_{Y;X_i|X^{i-1}}$, we have*

$$\left\|\mathfrak{i}_{Y;X_i|X^{i-1}}\right\|^2 = \mathbb{E}_{P_{Y,X^i}}\left[g^{[i]}(Y) \otimes f^{[i]}(X^i)\right]. \tag{23}$$

In addition, we can retrieve each $(X^i; Y)$ dependence from the optimal features (21), via

$$\mathfrak{i}_{Y;X^i} = g^{[[i]]} \otimes f^{[[i]]}, \quad \text{for all } i \in [n],$$

where $f^{[[i]]}, g^{[[i]]}$ are as defined in (20). Therefore, we can conveniently solve inference and estimation problem related to Y by using the set of optimal features. Formally, we have the following result, and the proof is omitted.

Proposition 6. *Suppose we have k_i-dimensional features $g^{[i]} \in \mathcal{F}_{\mathcal{Y}}^{k_i}$, $f^{[i]} \in \mathcal{F}_{\mathcal{X}_{[i]}}^{k_i}$ with $g^{[i]} \otimes f^{[i]} = \mathfrak{i}_{Y;X_i|X^{i-1}}$ for each $i \in [n]$. Then, for each $i \in [n]$, we have*

$$P_{Y|X^i}(y|x^i) = P_Y(y) \cdot \left[1 + g^{[[i]]}(y) \cdot f^{[[i]]}(x^n)\right], \tag{24}$$

and, for all $\psi \in \mathcal{F}_{\mathcal{Y}}^{d}$,

$$\mathbb{E}\left[\psi(Y)|X^i = x^i\right] = \mathbb{E}\left[\psi(Y)\right] + \Lambda_{\psi, g^{[[i]]}} \cdot f^{[[i]]}(x^i) \tag{25}$$

where $\Lambda_{\psi, g^{[[i]]}} = \mathbb{E}\left[\psi(Y)\left(g^{[[i]]}(Y)\right)^{\mathrm{T}}\right]$.

V. Stationary Processes

In this section, we consider the sequential dependence decomposition problem for stationary processes. In particular, we consider a process $\{X_n\}_{n\in\mathbb{Z}}$, i.e,

$$\cdots, X_{-2}, X_{-1}, X_0, X_1, X_2, \cdots$$

where each X_i take values from the same alphabet $\mathcal{X}_i = \mathcal{X}$. We assume the process is m-th order stationary, i.e.,

$$P_{X_{i_1}, X_{i_2}, \dots, X_{i_m}} = P_{X_{i_1+\tau}, X_{i_2+\tau}, \dots, X_{i_m+\tau}} \tag{26}$$

for all $i_1 < \cdots < i_m$ and $\tau \in \mathbb{Z}$.

It is of interest to characterize the dependence between variable X_t at time t and the history of length l, i.e., $(X_{t-1}, X_{t-2}, \dots, X_{t-l})$. Specifically, we assume $l \leq m-1$. Then, due to the m-th order stationarity, it is with out

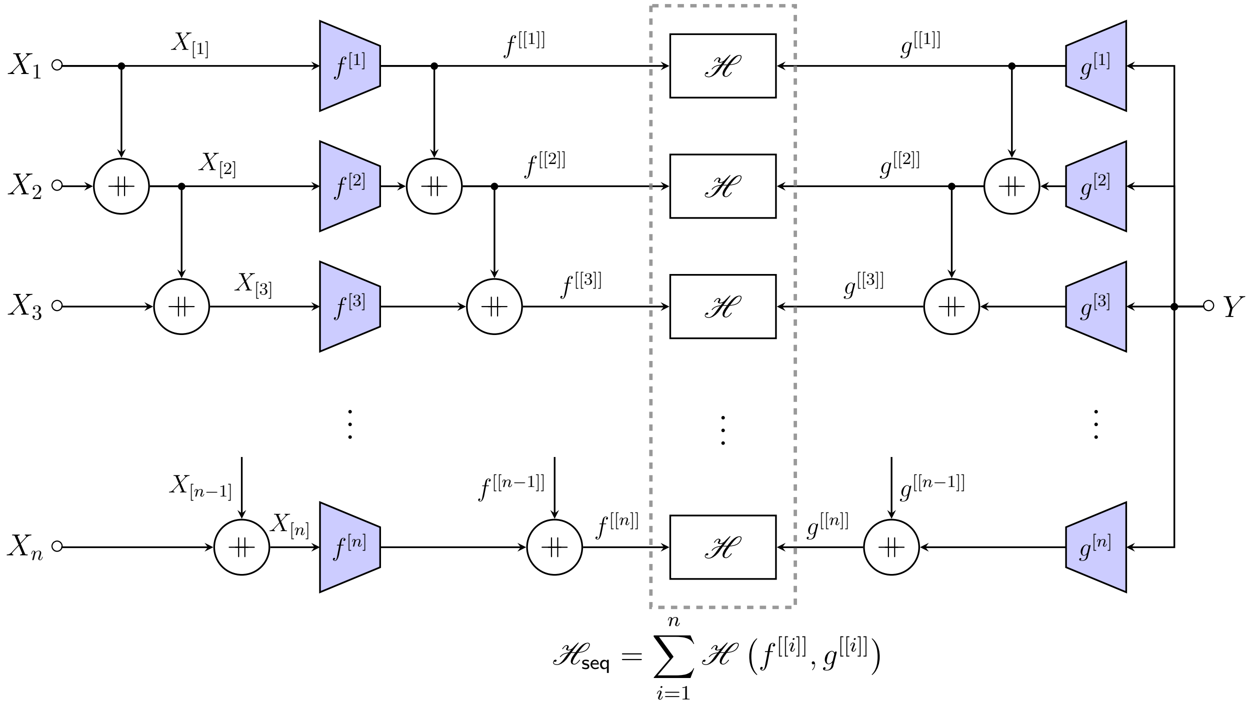

Fig. 1. Nested neural network structure for learning the sequential dependency decomposition, where "$+\!\!+$" denotes the concatenation operation of two vectors. Each filled block represents a deep feature extractor with trainable parameters.

loss of generality to set the current time as $t = 0$. For convenience, we introduce the notations $\overleftarrow{X}, \overleftarrow{x}$ to denote the reverse process and values, respectively: for all $i \in \mathbb{Z}$, we have $\overleftarrow{X}_i \triangleq X_{-i}$, $\overleftarrow{x}_i \triangleq x_{-i}$. Then, we can represent the past l observations $X_{[-l:-1]}$ as $\overleftarrow{X}_{[l]} = (\overleftarrow{X}_1, \dots, \overleftarrow{X}_l)$.

We then apply the decomposition established previously, by replacing Y and $X^l = X_{[l]}$ with $\overleftarrow{X}_0$ and $\overleftarrow{X}_{[l]}$, respectively. In particular, we establish the geometry on functional space $\mathcal{F}_{\mathcal{X}_{[0:l]}}$ by defining the metric distribution

$$R_{\overleftarrow{X}_{[0:l]}} \triangleq P_{\overleftarrow{X}_0} P_{\overleftarrow{X}_{[l]}} = P_X P_{\overleftarrow{X}_{[l]}}, \tag{27}$$

where we have denoted $P_X \triangleq P_{X_0}$.

Remark 1. *It is worth noting that while $\mathcal{X}_{\mathcal{I}} = \mathcal{X}_{\mathcal{I}'}$ for index sets $\mathcal{I}$ and $\mathcal{I}'$ with $|\mathcal{I}| = |\mathcal{I}'|$, the induced spaces $\mathcal{F}_{\mathcal{X}_{\mathcal{I}}}$ and $\mathcal{F}_{\mathcal{X}_{\mathcal{I}'}}$ generally have different geometries.*

Let us denote the dependence between the current $\overleftarrow{X}_0$ and the past l observations $\overleftarrow{X}_{[l]}$ by $\mathbb{i}_{\mathsf{seq}}^{[l]} \in \mathcal{F}_{\mathcal{X}_{[0:l]}}$, with $\mathbb{i}_{\mathsf{seq}}^{[0]} \triangleq 0$ and $\mathbb{i}_{\mathsf{seq}}^{[l]}(\overleftarrow{x}_{[0:l]}) \triangleq \mathbb{i}_{\overleftarrow{X}_0;\overleftarrow{X}_{[l]}}(\overleftarrow{x}_0, \overleftarrow{x}_{[l]})$ for $l \geq 1$. Then, we can obtain the conditional dependence at lag l by taking the first-order difference $\mathbb{i}_{\mathsf{seq}}^{\Delta[l]} \triangleq \mathbb{i}_{\mathsf{seq}}^{[l]} - \mathbb{i}_{\mathsf{seq}}^{[l-1]} = \mathbb{i}_{\overleftarrow{X}_0;\overleftarrow{X}_l|\overleftarrow{X}_{[l-1]}}$. To see this, note that

$$\begin{aligned}\mathbb{i}_{\mathsf{seq}}^{\Delta[l]}(\overleftarrow{x}_{[0:l]}) &\triangleq \mathbb{i}_{\mathsf{seq}}^{[l]}(\overleftarrow{x}_{[0:l]}) - \mathbb{i}_{\mathsf{seq}}^{[l-1]}(\overleftarrow{x}_{[0:l-1]})\\ &= \mathbb{i}_{\overleftarrow{X}_0;\overleftarrow{X}_{[l]}}(\overleftarrow{x}_0, \overleftarrow{x}_{[l]}) - \mathbb{i}_{\overleftarrow{X}_0;\overleftarrow{X}_{[l-1]}}(\overleftarrow{x}_0, \overleftarrow{x}_{[l-1]})\\ &= \mathbb{i}_{\overleftarrow{X}_0;\overleftarrow{X}_l|\overleftarrow{X}_{[l-1]}}(\overleftarrow{x}_0, \overleftarrow{x}_{[l-1]}, \overleftarrow{x}_l).\end{aligned}$$

This readily gives a decomposition of $\mathbb{i}_{\mathsf{seq}}^{[l]}$ into conditional dependencies of different lags, as

$$\mathbb{i}_{\mathsf{seq}}^{[l]} = \sum_{i=1}^{l} \mathbb{i}_{\mathsf{seq}}^{\Delta[i]}, \tag{28}$$

which is the counterpart of the orthogonal decomposition (10).

Then, we can represent the dependence components using learned features, by applying Theorem 1. Then we can use the learned features for inference, or estimating the strength of each dependence mode (cf. Proposition 5). In particular, for each $i \in [l]$, we can predict $\overleftarrow{X}_0$ by the previous i observations, via (cf. Proposition 6)

$$P_{\overleftarrow{X}_0|\overleftarrow{X}_{[i]}}(\overleftarrow{x}_0|\overleftarrow{x}_{[i]}) = P_X(\overleftarrow{x}_0)\left(1 + \mathbb{i}_{\mathsf{seq}}^{[i]}(\overleftarrow{x}_{[0:i]})\right). \tag{29}$$

A. Markov Approximation

Given process $\{X_n\}_{n\in\mathbb{Z}}$ with m-th order stationarity [cf. (26)], suppose we have learned dependence modes $\mathbb{i}_{\mathsf{seq}}^{[i]}$ for all $i \in [m-1]$ represented in features. Then, for each $i \in [m-1]$, we can define a Markov process of order i, of which the transition behaviors are specified by the conditional distribution $P_{\overleftarrow{X}_0|\overleftarrow{X}_{[i]}}$ [cf. (29)]. It is easy to verify that the resulting Markov process is also the optimal order-i approximation of the original process, in the sense of maximizing the likelihood function:

$$P_{\overleftarrow{X}_0|\overleftarrow{X}_{[i]}} = \underset{Q_{\overleftarrow{X}_0|\overleftarrow{X}_{[i]}}}{\arg\max}\, \mathbb{E}_{P_{\overleftarrow{X}_0,\overleftarrow{X}_{[l]}}}\left[\log Q_{\overleftarrow{X}_0|\overleftarrow{X}_{[i]}}(\overleftarrow{X}_0|\overleftarrow{X}_{[i]})\right].$$

Therefore, by learning the dependence decomposition (32) for $l = m-1$ from $\{X_n\}_{n\in\mathbb{Z}}$, we obtain optimal order-i Markov approximations of the original process, for all $i \in [m-1]$. One straightforward application is to test the order of Markov processes, which we illustrate as follows. A proof is provided in Appendix E.

Proposition 7. *Suppose $\{\mathcal{X}_n\}_{n\in\mathbb{Z}}$ is a Markov process of order M, and is stationary, i.e.,* (26) *holds for all $m \geq 1$. Then we have*

$$M = \max\left\{i \geq 1\colon \left\|\mathsf{i}_{\mathsf{seq}}^{\Delta[i]}\right\| > 0\right\}. \tag{30}$$

B. Joint Dependence and Decomposition

We then consider the joint dependence of variables in $\{X_n\}_{n\in\mathbb{Z}}$ of m-th order stationary. In particular, we characterize as the density ratio between joint distribution and marginal product. For given $X^n = (X_1, \dots, X_n)$, we denote this ratio by

$$\mathsf{i}_{X^n}(x^n) \triangleq \frac{P_{X^n}(x^n)}{\prod_{i=1}^n P_{X_i}(x_i)} - 1 \tag{31}$$

Then, we have the following result. A proof is provided in Appendix F.

Theorem 2. *For all $l \in [m-1]$, we have*

$$\mathsf{i}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]}) = \prod_{i=1}^{l}\left(1 + \mathsf{i}_{\mathsf{seq}}^{[i]}(\overleftarrow{x}_{[l-i:l]})\right) - 1 \tag{32}$$

Specifically, if $\left\|\mathsf{i}_{\overleftarrow{X}_{[0:l]}}\right\| = O(\epsilon)$, then we have

$$\begin{aligned}\mathsf{i}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]}) &= \sum_{i=1}^{l} \mathsf{i}_{\mathsf{seq}}^{[i]}(\overleftarrow{x}_{[l-i:l]}) + o(\epsilon) &(33)\\ &= \sum_{i=1}^{l}\sum_{j=0}^{l-i} \mathsf{i}_{\mathsf{seq}}^{\Delta[i]}(\overleftarrow{x}_{[j:j+i]}) + o(\epsilon) &(34)\end{aligned}$$

Note that when $\left\|\mathsf{i}_{\overleftarrow{X}_{[0:l]}}\right\| = O(\epsilon)$, it will be more convenient to use the product distribution $R'_{\overleftarrow{X}_{[0:m]}} \triangleq \prod_{i\in[0:m]} P_{\overleftarrow{X}_i}$ as the metric distribution. In fact, it can be verified that such replacement only induces an error of higher-order terms, i.e., $o(\epsilon^2)$. Then, the following result illustrates that the decompositions (33) and (33) are orthogonal. A proof is provided in Appendix G.

Lemma 2. *We have*

$$\left\langle \mathsf{i}_{\mathsf{seq}}^{[i]}(\overleftarrow{X}_{[l-i:l]}), \mathsf{i}_{\mathsf{seq}}^{[i']}(\overleftarrow{X}_{[l-i':l]})\right\rangle_{R'} = 0 \quad \textit{if } i \neq i',$$

and

$$\left\langle \mathsf{i}_{\mathsf{seq}}^{\Delta[i]}(\overleftarrow{X}_{[j:j+i]}), \mathsf{i}_{\mathsf{seq}}^{\Delta[i']}(\overleftarrow{X}_{[j':j'+i']})\right\rangle_{R'} = 0$$

if $(i', j') \neq (i, j)$.

The following corollary of Theorem 2 is then immediate.

Corollary 4. *If $\|\mathsf{i}_{\overleftarrow{X}_{[0:l]}}\| = O(\epsilon)$, we have*

$$\begin{aligned}\left\|\mathsf{i}_{\overleftarrow{X}_{[0:l]}}\right\|^2 &= \sum_{i=1}^{l}\left\|\mathsf{i}_{\mathsf{seq}}^{[i]}\right\|^2 + o(\epsilon^2) &(35)\\ &= \sum_{i=1}^{l}(l+1-i)\cdot\left\|\mathsf{i}_{\mathsf{seq}}^{\Delta[i]}\right\|^2 + o(\epsilon^2) &(36)\end{aligned}$$

As a final remark, the above decompositions are also well-connected to corresponding decompositions of information density and information measures. Specifically, for given X^n, let us define $\tilde{\imath}_{X^n}(x^n) \triangleq \log\frac{P_{X^n}(x^n)}{\prod_{i=1}^n P_{X_i}(x_i)}$ as the information density for X^n, then Watanabe's total correlation [18] of X^n is defined as $C(X^n) \triangleq \mathbb{E}_{\hat{X}^n\sim P_{X^n}}\left[\tilde{\imath}_{X^n}(\hat{X}^n)\right]$. The following result gives the counterparts of Theorem 2 and Corollary 4. We provide its proof in Appendix H for completeness.

Proposition 8. *For each $l \in [m-1]$, we have*

$$\begin{aligned}&\tilde{\imath}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]})\\ &\quad= \sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{[i]}}(\overleftarrow{x}_{[l-i:l]}) &(37)\\ &\quad= \sum_{i=1}^{l}\sum_{j=0}^{l-i}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}}(\overleftarrow{x}_j, \overleftarrow{x}_{i+j}|\overleftarrow{x}_{[j+1:i+j-1]}), &(38)\end{aligned}$$

where the conditional information density is as defined in (9). *In addition, we have*

$$\begin{aligned}C(\overleftarrow{X}_{[0:l]}) &= \sum_{i=1}^{l} I(\overleftarrow{X}_0;\overleftarrow{X}_{[i]}) &(39)\\ &= \sum_{i=1}^{l}(l-i+1)\cdot I\left(\overleftarrow{X}_0;\overleftarrow{X}_i\middle|\overleftarrow{X}_{[i-1]}\right). &(40)\end{aligned}$$

APPENDIX

A. Proof of Proposition 3

From the definition of conditional CDK function [cf. (8)], we have

$$\mathsf{i}_{Y;X_i|X^{i-1}} = \mathsf{i}_{Y;X^i} - \mathsf{i}_{Y;X^{i-i}} \quad \text{for all } i \in [l], \tag{41}$$

where $\mathsf{i}_{Y;X^0} = \mathsf{i}_{Y;\varnothing} = 0$. Taking the sum on both sides of over $i \in [l]$, we obtain (10).

To establish the orthogonality, note that for $i < j$, we have $i \leq j-1$ and $\mathsf{i}_{Y;X_i|X^{i-1}} \in \mathcal{F}_{\mathcal{Y}\times\mathcal{X}_{[i]}} \subset \mathcal{F}_{\mathcal{Y}\times\mathcal{X}_{[j-1]}}$. In addition, from Proposition 2, we have $\mathsf{i}_{Y;X_j|X^{j-1}} \in \mathcal{F}_{\mathcal{X}_j|(\mathcal{Y}\times\mathcal{X}_{[j-1]})}$, which implies that $\mathsf{i}_{Y;X_j|X^{j-1}} \perp \mathcal{F}_{\mathcal{Y}\times\mathcal{X}_{[j-1]}} \ni \mathsf{i}_{Y;X_i|X^{i-1}}$. □

B. Proof of Proposition 4

For all $i \in [n]$, $g \in \mathcal{F}_{\mathcal{Y}}$ and $f \in \mathcal{F}_{\mathcal{X}_i|\mathcal{X}_{[i-1]}}$, from $\mathsf{i}_{Y;X^i} = \mathsf{i}_{Y;X^{i-1}} + \mathsf{i}_{Y;X_i|X^{i-1}}$, we have

$$\begin{aligned}\operatorname{cov}(g, f) &= \langle \mathsf{i}_{Y;X^i}, g\otimes f\rangle\\ &= \langle \mathsf{i}_{Y;X^{i-1}}, g\otimes f\rangle + \langle \mathsf{i}_{Y;X_i|X^{i-1}}, g\otimes f\rangle\\ &= \langle \mathsf{i}_{Y;X^{i-1}}, g\otimes f\rangle, &(42)\end{aligned}$$

where the last equality follows from the fact that $\mathsf{i}_{Y;X_i|X^{i-1}} \perp \mathcal{F}_{\mathcal{Y}\times\mathcal{X}_{[i-1]}} \ni (g\otimes f)$. Therefore, for each $j \geq 1$, the maximum value of $\operatorname{cov}(g_j, f_j)$ cannot exceed the corresponding maximum value of $\langle \mathsf{i}_{Y;X^{i-1}}, g_j\otimes f_j\rangle$, under the same sets of constraints.

However, from Property 3, the optimal functions that maximize $\langle \mathsf{i}_{Y;X^{i-1}}, g_j\otimes f_j\rangle$ are given by $g_j = g_j^{*[i]}, f_j = f_j^{*[i]}$ for all $j \geq 1$, which also satisfy the unit-norm and orthogonality constraints [cf. (16)]. Hence, we conclude that $g_j^{*[i]}, f_j^{*[i]}$ are the optimal solutions. Finally, the relation $\sigma_j^{[i]} = \operatorname{cov}(g_j^{*[i]}, f_j^{*[i]})$ can be readily obtained from (42) and (15). □

C. Proof of Theorem 1

From Proposition 1, for each $i \in [n]$, $\mathscr{H}\left(f^{[[i]]}, g^{[[i]]}\right)$ is maximized if and only if

$$g^{[[i]]} \otimes f^{[[i]]} = \mathfrak{i}_{Y;X^i}. \tag{43}$$

Therefore, $\mathscr{H}_{\mathsf{seq}}$ is maximized if (43) holds for all $i \in [n]$. This is further equivalent to, for all $i \in [n]$,

$$\begin{aligned} g^{[i]} \otimes f^{[i]} &= g^{[[i]]} \otimes f^{[[i]]} - g^{[[i-1]]} \otimes f^{[[i-1]]} \\ &= \mathfrak{i}_{Y;X^i} - \mathfrak{i}_{Y;X^{i-1}} = \mathfrak{i}_{Y;X_i|X^{i-1}}, \end{aligned}$$

where we have defined $g^{[[0]]} = f^{[[0]]} = 0$. This establishes (21).

Similarly, (22) is the only common solution such that each single H-score in $\mathscr{H}^{\star}_{\mathsf{seq}}$ gets maximized, which completes the proof. □

D. Proof of Proposition 5

Note that we have

$$\begin{aligned} &\mathbb{E}_{P_{Y,X^i}}\left[g^{[i]}(Y) \otimes f^{[i]}(X_{[i]})\right] \\ &\quad = \left\langle 1 + \mathfrak{i}_{Y;X^i}, g^{[i]}(Y) \otimes f^{[i]}(X_{[i]})\right\rangle \\ &\quad = \langle 1 + \mathfrak{i}_{Y;X^i}, \mathfrak{i}_{Y;X_i|X^{i-1}}\rangle \\ &\quad = \|\mathfrak{i}_{Y;X_i|X^{i-1}}\|^2, \end{aligned}$$

where to obtain the last equality we have used the facts that $\langle 1, \mathfrak{i}_{Y;X_i|X^{i-1}}\rangle = 0$ and $\mathfrak{i}_{Y;X_i|X^{i-1}} = \Pi\left(\mathfrak{i}_{Y;X^i}; \mathcal{F}_{\mathcal{Y}\times\mathcal{F}_{\mathcal{X}_{[i]}}}\right)$. □

E. Proof of Proposition 7

By definition, for all $\tau \geq 1$, we have

$$P_{\overleftarrow{X}_0|\overleftarrow{X}_{[M+\tau]}} = P_{\overleftarrow{X}_0|\overleftarrow{X}_{[M]}}. \tag{44}$$

Therefore, we obtain the Markov relation $\overleftarrow{X}_0 - \overleftarrow{X}_{[n-1]} - \overleftarrow{X}_n$ for all $n > M$. As a consequence, from Corollary 2 we have

$$\left\|\mathfrak{i}^{\Delta[i]}_{\mathsf{seq}}\right\| = \left\|\mathfrak{i}_{\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}}\right\| = 0 \quad \text{for all } i > M.$$

It suffices to establish $\left\|\mathfrak{i}^{\Delta[M]}_{\mathsf{seq}}\right\| > 0$. In fact, if $\left\|\mathfrak{i}^{\Delta[M]}_{\mathsf{seq}}\right\| = 0$, we obtain the Markov relation $\overleftarrow{X}_0 - \overleftarrow{X}_{[M-1]} - \overleftarrow{X}_M$, and the order of $\{\mathcal{X}_n\}_{n\in\mathcal{Z}}$ will be at most $M - 1$, leading to a contradiction. Hence, we obtain (30) as desired. □

F. Proof of Theorem 2

To obtain (32), it suffices to note that

$$\begin{aligned} &P_{X_{[-l:0]}}(x_{[-l:0]}) \\ &\quad = P_{X_{-l}}(x_{-l}) \prod_{i=-l}^{-1} P_{X_{i+1}|X_{[-l:i]}}(x_{i+1}|x_{[-l:i]}) \\ &\quad = P_X(x_{-l}) \prod_{i=-l}^{-1} P_{X_0|X_{[-l-i-1:-1]}}(x_{i+1}|x_{[-l:i]}) \\ &\quad = P_X(x_{-l}) \prod_{i=-l}^{-1} P_{X_0|\overleftarrow{X}_{[l+i+1]}}(\overleftarrow{x}_{-i-1}|\overleftarrow{x}_{[-i:l]}) \\ &\quad = P_X(x_{-l}) \prod_{i=1}^{l} P_{X_0|\overleftarrow{X}_{[i]}}(\overleftarrow{x}_{l-i}|\overleftarrow{x}_{[l-i+1:l]}) \\ &\quad = \left[\prod_{i=0}^{l} P_X(\overleftarrow{x}_i)\right] \cdot \prod_{i=1}^{l}\left[1 + \mathfrak{i}^{[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l]})\right], \end{aligned}$$

where to obtain the second equality we have used the stationarity, and where to obtain the last equality we have used (29).

Specifically, when $\|\mathfrak{i}_{X_{-l};\cdots;X_0}\| = O(\epsilon)$, we have $\mathfrak{i}^{[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l]}) = O(\epsilon)$ for all $i \in [l]$, and thus

$$\begin{aligned} &\prod_{i=1}^{l}\left[1 + \mathfrak{i}^{[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l]})\right] - 1 \\ &\quad = \sum_{i=1}^{l} \mathfrak{i}^{[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l]}) + o(\epsilon) &(45)\\ &\quad = \sum_{i=1}^{l}\sum_{j=1}^{i} \mathfrak{i}^{\Delta[j]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l-i+j]}) + o(\epsilon) &(46)\\ &\quad = \sum_{j=1}^{l}\sum_{i=j}^{l} \mathfrak{i}^{\Delta[j]}_{\mathsf{seq}}(\overleftarrow{x}_{[l-i:l-i+j]}) + o(\epsilon) &(47)\\ &\quad = \sum_{j=1}^{l}\sum_{i=0}^{l-j} \mathfrak{i}^{\Delta[j]}_{\mathsf{seq}}(\overleftarrow{x}_{[i:i+j]}) + o(\epsilon) &(48)\\ &\quad = \sum_{i=1}^{l}\sum_{j=0}^{l-i} \mathfrak{i}^{\Delta[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[j:i+j]}) + o(\epsilon) &(49) \end{aligned}$$

where (46) follows from (28), and where to obtain (47) we have changed the order of summations. □

G. Proof of Lemma 2

Note that the first statement can be implies by the second one since

$$\begin{aligned} &\left\langle \mathfrak{i}^{[i]}_{\mathsf{seq}}(\overleftarrow{X}_{[l-i:l]}), \mathfrak{i}^{[i']}_{\mathsf{seq}}(\overleftarrow{X}_{[l-i':l]})\right\rangle_{R'} \\ &\quad = \sum_{j=1}^{i}\sum_{j'=1}^{i'} \left\langle \mathfrak{i}^{[j]}_{\mathsf{seq}}(\overleftarrow{X}_{[l-i:l]}), \mathfrak{i}^{[j']}_{\mathsf{seq}}(\overleftarrow{X}_{[l-i':l]})\right\rangle_{R'}. \end{aligned}$$

It remains only to establish the second statement. From the modal decomposition, we can write

$$\begin{aligned} \mathfrak{i}^{\Delta[i]}_{\mathsf{seq}}(\overleftarrow{x}_{[j:j+i]}) &= g^{[i]}(\overleftarrow{x}_j) \cdot f^{[i]}(\overleftarrow{x}_{[j+1:j+i]}), \\ \mathfrak{i}^{\Delta[i']}_{\mathsf{seq}}(\overleftarrow{x}_{[j':j'+i']}) &= g^{[i']}(\overleftarrow{x}_{j'}) \cdot f^{[i']}(\overleftarrow{x}_{[j'+1:j'+i']}). \end{aligned}$$

Suppose $j \neq j'$. Then, without loss of generality, we can assume $j < j'$. Since R' is a product distribution and $\mathbb{E}\left[g^{[j]}(\overleftarrow{X}_j)\right] = \mathbb{E}\left[g^{[j]}(\overleftarrow{X}_0)\right] = 0$, we have

$$\begin{aligned} &\left\langle \mathfrak{i}^{\Delta[i]}_{\mathsf{seq}}(\overleftarrow{X}_{[j:j+i]}), \mathfrak{i}^{\Delta[i']}_{\mathsf{seq}}(\overleftarrow{X}_{[j':j'+i']})\right\rangle_{R'} \\ &\quad = \mathbb{E}_{R'}\left[\left(g^{[i]}(\overleftarrow{X}_j) \cdot f^{[i]}(\overleftarrow{X}_{[j+1:j+i]})\right)\right. \\ &\qquad\quad \left.\cdot \left(g^{[i']}(\overleftarrow{X}_{j'}) \cdot f^{[i']}(\overleftarrow{X}_{[j'+1:j'+i']})\right)\right] \\ &\quad = 0. \end{aligned}$$

Therefore, it remains only to consider the case with $j = j'$ but $i \neq i'$. We can assume $i < i'$ without loss of generality. Note that since

$$\mathbb{E}_{P_{\overleftarrow{X}_{j+i'}}}\left[f^{[i']}(\overleftarrow{x}_{[j+1:j+i'-1]}, \overleftarrow{X}_{j+i'})\right] = 0,$$

we obtain

$$\left\langle \mathfrak{i}^{\Delta[i]}_{\mathsf{seq}}(\overleftarrow{X}_{[j:j+i]}), \mathfrak{i}^{\Delta[i']}_{\mathsf{seq}}(\overleftarrow{X}_{[j:j+i']})\right\rangle_{R'}$$

$$
\begin{aligned}
&= \mathbb{E}_{R'}\left[\left(g^{[i]}(\overleftarrow{X}_j)\cdot f^{[i]}(\overleftarrow{X}_{[j+1:j+i]})\right)\right.\\
&\qquad\left.\cdot\left(g^{[i']}(\overleftarrow{X}_j)\cdot f^{[i']}(\overleftarrow{X}_{[j+1:j+i']})\right)\right]\\
&= 0.
\end{aligned}
$$

□

H. Proof of Proposition 8

For each $i \in [l]$, we have

$$
\begin{aligned}
&\tilde{\imath}_{\overleftarrow{X}_{l-i};\overleftarrow{X}_{[l+1-i:l]}}(\overleftarrow{x}_{[l-i:l]})\\
&\qquad= \log\frac{P_{\overleftarrow{X}_{[l-i:l]}}(\overleftarrow{x}_{[l-i:l]})}{P_{\overleftarrow{X}_{l-i}}(\overleftarrow{x}_{l-i})P_{\overleftarrow{X}_{[l+1-i:l]}}(\overleftarrow{x}_{[l+1-i:l]})}\\
&\qquad= \log\frac{P_{\overleftarrow{X}_{[l-i:l]}}(\overleftarrow{x}_{[l-i:l]})}{P_{\overleftarrow{X}_{[l-(i-1):l]}}(\overleftarrow{x}_{[l-(i-1):l]})} - \log P_{\overleftarrow{X}_{l-i}}(\overleftarrow{x}_{l-i}),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&\sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_{l-i};\overleftarrow{X}_{[l+1-i:l]}}(\overleftarrow{x}_{[l-i:l]})\\
&= \log P_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]}) - \sum_{i=0}^{l}\log P_{\overleftarrow{X}_i}(\overleftarrow{x}_i) = \tilde{\imath}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]}).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\tilde{\imath}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]}) &= \sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_{l-i};\overleftarrow{X}_{[l+1-i:l]}}(\overleftarrow{x}_{[l-i:l]}) && (50)\\
&= \sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{[i]}}(\overleftarrow{x}_{[l-i:l]}), && (51)
\end{aligned}
$$

which gives (37). To obtain (38), note that from (50), we have

$$
\begin{aligned}
&\tilde{\imath}_{\overleftarrow{X}_{[0:l]}}(\overleftarrow{x}_{[0:l]})\\
&\quad= \sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_{l-i};\overleftarrow{X}_{[l+1-i:l]}}(\overleftarrow{x}_{[l-i:l]})\\
&\quad= \sum_{i=1}^{l}\tilde{\imath}_{\overleftarrow{X}_{i-1};\overleftarrow{X}_{[i:l]}}(\overleftarrow{x}_{[i-1:l]})\\
&\quad= \sum_{i=1}^{l}\sum_{j=0}^{l-i}\tilde{\imath}_{\overleftarrow{X}_{i-1};\overleftarrow{X}_{i+j}|\overleftarrow{X}_{[i:i+j-1]}}(\overleftarrow{x}_{i-1},\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[i:i+j-1]})\\
&\quad= \sum_{i=1}^{l}\sum_{j=0}^{l-i}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{j+1}|\overleftarrow{X}_{[j]}}(\overleftarrow{x}_{i-1},\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[i:i+j-1]})\\
&\quad= \sum_{j=0}^{l-1}\sum_{i=1}^{l-j}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{j+1}|\overleftarrow{X}_{[j]}}(\overleftarrow{x}_{i-1},\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[i:i+j-1]})\\
&\quad= \sum_{i=0}^{l-1}\sum_{j=1}^{l-i}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{i+1}|\overleftarrow{X}_{[i]}}(\overleftarrow{x}_{j-1},\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[j:i+j-1]})\\
&\quad= \sum_{i=1}^{l}\sum_{j=0}^{l-i}\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}}(\overleftarrow{x}_j,\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[j+1:i+j-1]}),
\end{aligned}
$$

where to obtain the third equality we have used the fact that [cf. (9)]

$$
\begin{aligned}
&\tilde{\imath}_{\overleftarrow{X}_{i-1};\overleftarrow{X}_{[i:l]}}(\overleftarrow{x}_{[i-1:l]})\\
&\qquad= \sum_{j=0}^{l-i}\tilde{\imath}_{\overleftarrow{X}_{i-1};\overleftarrow{X}_{i+j}|\overleftarrow{X}_{[i:i+j-1]}}(\overleftarrow{x}_{i-1},\overleftarrow{x}_{i+j}|\overleftarrow{x}_{[i:i+j-1]}),
\end{aligned}
$$

and where to obtain the fourth equality we have used the stationarity of $\{X_n\}_{n\in\mathbb{Z}}$.

Finally, (39) and (40) can be obtained by taking expectations of (37) and (38), respectively, via noting the facts that $\mathbb{E}\left[\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_{[i]}}(\overleftarrow{X}_{[l-i:l]})\right] = I(\overleftarrow{X}_0;\overleftarrow{X}_{[i]})$ and

$$
\begin{aligned}
\mathbb{E}&\left[\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}}(\overleftarrow{X}_j,\overleftarrow{X}_{i+j}|\overleftarrow{X}_{[j+1:i+j-1]})\right]\\
&= \mathbb{E}\left[\tilde{\imath}_{\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}}(\overleftarrow{X}_0,\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]})\right]\\
&= I(\overleftarrow{X}_0;\overleftarrow{X}_i|\overleftarrow{X}_{[i-1]}).
\end{aligned}
$$

□

References

[1] Y. Bengio *et al.*, "Markovian models for sequential data," *Neural computing surveys*, vol. 2, no. 199, pp. 129–162, 1999.

[2] T. G. Dietterich, "Machine learning for sequential data: A review," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada, August 6–9, 2002 Proceedings*. Springer, 2002, pp. 15–30.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[7] S.-I. Amari, "Information geometry on hierarchy of probability distributions," *IEEE transactions on information theory*, vol. 47, no. 5, pp. 1701–1711, 2001.

[8] 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.

[9] 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.

[10] X. Xu and L. Zheng, "Multivariate feature extraction," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–8.

[11] 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.

[12] 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.

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

[15] 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.

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

[17] X. Xu and L. Zheng, "A geometric framework for neural feature learning," *CoRR*, vol. abs/2309.10140, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.10140

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