An Information Theoretic Framework for Distributed Learning Algorithms

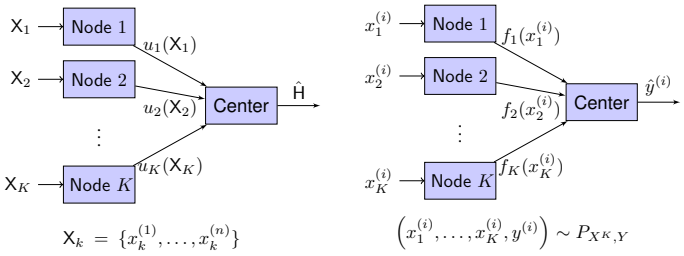
Xiangxiang Xu and Shao-Lun Huang
DSIT Research Center
Tsinghua–Berkeley Shenzhen Institute
Shenzhen, China 518055
Email: xiangxiangxu@ieee.org, shaolun.huang@sz.tsinghua.edu.cn

***Abstract*—Distributed learning is recently an important research topic, while the information theoretic optimality of the distributed learning algorithms is often not sufficiently addressed. This paper studies the distributed learning problems such that each node observes i.i.d. samples and sends a feature function of observed samples to the central machine for decision making. Both the binary hypothesis testing in information theory and the classification problems in machine learning are considered, and the optimal error exponent and the set of optimal features are characterized. By exploiting an information theoretic framework, we show that these two problems share the same set of optimal features, from which the information theoretic optimality of some machine learning algorithms can be established. Finally, we generalize our analyses to M-ary distributed hypothesis testing and classification problems.**

I. Introduction

The connections between information theory, statistics, and machine learning have recently been extensively explored, and the supervised learning is considered as a particularly successful area [1]. The key challenge in supervised learning is to extract informative low-dimensional features of data for effectively predicting the label by linear classifiers such as the support vector machine (SVM) [2] or softmax regression [3]. Such feature extraction problems have been studied by statistical and information theoretic tools [4], in which the fundamental connections between Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [5]–[7], Chernoff information [8], Wyner's common information [9], information bottleneck [10], and deep learning [11] are well established via an information geometric framework for single mode of data.

On the other hand, machine learning problems with distributed or multi-modal data have also gained much attentions recently in federated and multi-modal machine learning [12], [13]. Such problems were also extensively studied in information theory [14]–[17], but rather focused on deriving theoretical bounds in statistical models [18], [19], which leads to a gap between theory and practice. In this paper, our goal is to establish a theoretical connection between information theory and machine learning algorithms in distributed learning, by characterizing the optimal features in distributed supervised learning algorithms based on an information theoretic framework. Specifically, we consider two types of distributed learning setups for both information theory and machine learning scenarios:



(a) Distributed Hypothesis Testing (b) Distributed Classification

Fig. 1. Two Types of Distributed Learning Setups

Hypothesis testing setup: Suppose that there are K random variables $X^K \triangleq (X_1, \ldots, X_K)$, n observable samples $\left\{(x_1^{(\ell)}, \ldots, x_K^{(\ell)})\right\}_{\ell=1}^n$, and hypotheses $\mathsf{H} \in \{0, 1\}$, such that when $\mathsf{H} = i$, the n samples are generated in an independent, identically distributed (i.i.d.) manner from the distribution $P_{X^K}^{(i)}$. In addition, we assume that there are K distributed nodes, where the k-th node can only observe the samples $\mathsf{X}_k \triangleq \{x_k^{(1)}, \ldots, x_k^{(n)}\}$, and send a d_k-dimensional statistic $u_k(\mathsf{X}_k) = \frac{1}{n}\sum_{\ell=1}^n f_k(x_k^{(\ell)})$, for some d_k-dimensional function f_k, to a central machine. Then, the central machine collects the K statistics, and make a decision $\hat{\mathsf{H}}$ on the true hypothesis, as shown in Fig. 1a. The goal is to design the feature functions f_k such that when the central machine makes the maximum a posteriori (MAP) decision, the decision error rate is minimized.

Machine learning setup: Suppose that there are K data variables $X^K \triangleq (X_1, \ldots, X_K)$, and a binary label $Y \in \{0, 1\}$, and n training samples $(x_1^{(i)}, \ldots, x_K^{(i)}, y^{(i)})$, for $i = 1, \ldots, n$, with the empirical distribution $P_{X^K,Y}$. Again, we assume K distributed nodes, and the node k can only observe the data from the k-th data variable. In order to predict the label $y^{(i)}$, each node k sends a d_k-dimensional feature function $f_k(x_k^{(i)})$ to a central machine, and the central machine collects K features and employs a linear classifier for predicting the label $y^{(i)}$, as shown in Fig. 1b. The goal is to design the feature functions f_k, such that the training loss of the linear classifier of the central machine is minimized.

The contributions of this paper are summarized as follows. First, we show that for binary hypothesis testing, the optimal

error exponent can be achieved by one-dimensional functions for each node, i.e., $d_k = 1$, for all k, and the optimal feature functions can be computed by a Kullback-Leibler (K-L) divergence optimization problem. In addition, by applying the information geometric approach in [4], the hypotheses and the feature functions of each node can be modeled as vectors in the joint and marginal distribution spaces, respectively. Then, the optimal feature function of each node can be interpreted as a decomposition of the hypothesis vector in the joint distribution space into vectors in the marginal distribution spaces, where each decomposed component indicates the contribution of the corresponding node in making the inference. Moreover, for the machine learning setup, we show that when applying the softmax regression as the linear classifier in the local framework, the optimal feature functions coincide with the hypothesis testing setup, which demonstrates the fundamental connections between information theory and machine learning in the distributed learning problems. Finally, we generalize our analyses and show the coincidence for M-ary hypothesis testing and machine learning problems.

Related Works: The distributed hypothesis testing problems (also referred to as multiterminal hypothesis testing [14], [17], [18], or decentralized detection [20], [21]) have been extensively studied. When each node can observe single observation, and send an encoded message to the central machine, [20] showed that it is NP hard to determine the optimal coding scheme, and [21], [22] characterized the minimum decoding error rate and optimal coding scheme for conditionally independent nodes. In addition, when each node can observe n samples and send an encoded message to the central machine, [14], [16], [18], [23] characterized the optimal decoding error exponents for $K = 2$ nodes, and [24] generalized the results to $K > 2$ nodes. Moreover, [16] studied the Neyman–Pearson-like test, which further restricted the encoded messages being some empirical functional mean, and demonstrated the optimal functions for $K = 2$ nodes. Our result in Section III-A can be viewed as a generalization of such setup to $K > 2$ nodes. On the other hand, designing distributed learning algorithms is a well explored subject in machine learning, including distributed neural networks [12], [25], [26], distributed SVM [27], [28], and distributed LASSO (Least Absolute Shrinkage and Selection Operator) [29]. Despite of such massive studies, fundamental understandings of the optimality of the algorithms from the information theoretic perspective are often lacking. Our goal in this paper is to establish such connections between information theory and machine learning.

II. Problem Formulation and Preliminaries

In this section, we present the mathematical formulation of the distributed hypothesis testing and classification problem and introduce some related preliminaries.

A. Distributed Hypothesis Testing

In distributed hypothesis testing problem, we introduce a common assumption in distributed setup [18] that the generating distributions $P_{X^K}^{(0)}$ and $P_{X^K}^{(1)}$ satisfy $D(P_{X^K}^{(1)}\|P_{X^K}^{(0)}) < \infty, D(P_{X^K}^{(0)}\|P_{X^K}^{(1)}) < \infty$, to avoid the trivial irregularities.

The data samples $\left\{(x_1^{(\ell)},\ldots,x_K^{(\ell)})\right\}_{\ell=1}^n$ are i.i.d. generated from $P_{X^K}^{(\mathsf{H})}$, with the k-th node observing the n samples of X_k, i.e., $\mathsf{X}_k = \left\{x_k^{(1)},\ldots,x_k^{(n)}\right\}$, for $k = 1,\ldots,K$. Then, each node k sends the statistic

$$u_k(\mathsf{X}_k) = \frac{1}{n}\sum_{\ell=1}^{n} f_k\left(x_k^{(\ell)}\right) = \mathbb{E}_{\hat{P}_{\mathsf{X}_k}}\left[f_k(X_k)\right], \tag{1}$$

to the central machine, where $f_k\colon \mathcal{X}_k \to \mathbb{R}^{d_k}$ indicates the feature function and where $\hat{P}_{\mathsf{X}_k}$ denotes the empirical distribution of $\mathsf{X}_k = \left\{x_k^{(1)},\ldots,x_k^{(n)}\right\}$, i.e.,

$$\hat{P}_{\mathsf{X}_k}(x_k) \triangleq \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}_{\left\{x_k^{(\ell)}=x_k\right\}}, \quad \text{for all } x_k \in \mathcal{X}_k.$$

After receiving the features $u_1,\ldots,u_K$ sent by K nodes, the central machine then decides $\hat{\mathsf{H}}$ based on the MAP, which can be expressed as the log-likelihood ratio test

$$\log\frac{\mathbb{P}_n\left\{u_1,\ldots,u_K|\mathsf{H}=1\right\}}{\mathbb{P}_n\left\{u_1,\ldots,u_K|\mathsf{H}=0\right\}} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} \log\frac{P_{\mathsf{H}}(0)}{P_{\mathsf{H}}(1)}, \tag{2}$$

where $\mathbb{P}_n\{\cdot\}$ denotes the probability measured over all i.i.d. sampling process of $\left\{(x_1^{(\ell)},\ldots,x_K^{(\ell)})\right\}_{\ell=1}^n$.

Then, we characterize the decision error by the associated exponent

$$E \triangleq -\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n\left\{\hat{\mathsf{H}}\neq\mathsf{H}\right\}. \tag{3}$$

Our goal is to characterize the optimal achievable error exponent, defined as

$$E^* \triangleq -\min_{f_1,\ldots,f_K}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n\left\{\hat{\mathsf{H}}\neq\mathsf{H}\right\}, \tag{4}$$

and the corresponding optimal functions f_k, for $k = 1,\ldots,K$.

B. HGR Maximal Correlation and Regression

First, we introduce the HGR maximal correlation as follows.

Definition 1: Given a pair of discrete random variables X, Y with alphabets $\mathcal{X}$, $\mathcal{Y}$, respectively, their (generalized) HGR maximal correlation functions [4] are the solution of the optimization problem

$$\underset{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0\\ \mathbb{E}\left[f(X)f^{\mathrm{T}}(X)\right]=\mathbb{E}\left[g(Y)g^{\mathrm{T}}(Y)\right]=I_m}}{\text{maximize}} \quad \mathbb{E}\left[f^{\mathrm{T}}(X)g(Y)\right] \tag{5}$$

over m-dimensional functions $f\colon \mathcal{X}\to\mathbb{R}^m$, $g\colon \mathcal{Y}\to\mathbb{R}^m$, where I_m denotes the identity matrix of order m, and where the expectation is taken over the joint distribution $P_{X,Y}$ of X and Y.

It can be shown that the m-dimensional maximal correlation function corresponds to the top m singular vectors of the *canonical dependence matrix* (CDM) [4] $\tilde{B}_{X,Y} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ of X and Y, with entries

$$\tilde{B}_{X,Y}(x,y) \triangleq \frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}} \tag{6}$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$. In addition, it is shown in [30] that the HGR maximal correlation function can be computed via solving a low-rank approximation problem of CDM that minimizes

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left[\tilde{B}_{X,Y}(x,y) - f^{\mathrm{T}}(x)g(y)\sqrt{P_X(x)}\sqrt{P_Y(y)}\right]^2 \quad (7)$$

over all $f\colon \mathcal{X} \to \mathbb{R}^m, g\colon \mathcal{Y} \to \mathbb{R}^m$. This implies that the entry $\tilde{B}_{X,Y}(x,y)$ of CDM can be well-approximated by $f^{\mathrm{T}}(x)g(y)\sqrt{P_X(x)}\sqrt{P_Y(y)}$, and thus the posterior probability $P_{Y|X}(y|x)$ can be approximated by

$$P^{(f,g)}_{Y|X}(y|x) \triangleq P_Y(y)(1 + f^{\mathrm{T}}(x)g(y)). \quad (8)$$

This leads to a regression approach, which predicts the label $\hat{y}$ for newly observed x by the MAP decision with respect to the approximated conditional distribution (8) (cf. [31, Eq. (5)])

$$\hat{y}(x) = \arg\max_{y'} P^{(f,g)}_{Y|X}(y'|x). \quad (9)$$

It turns out that (9) is referred to as the maximal correlation regression (MCR) [31], and is equivalent to the softmax regression in the local regime, where X, Y are weakly dependent [31].

In the rest of this paper, we adopt the MCR in the central machine as the regression algorithm, where the feature f is concatenated by all f_i's received from K nodes, i.e.,

$$f(x_1,\ldots,x_K) \triangleq \begin{bmatrix} f_1(x_1) \\ \vdots \\ f_K(x_K) \end{bmatrix}. \quad (10)$$

Then, our goal is to characterize the optimal features $f_1,\ldots,f_K$ such that the approximation error (7) is minimized.

C. Local Information Geometry

Some useful notations and definitions in local information geometry [4] are introduced as follows.

Definition 2 (ϵ-neighborhood): Given a finite alphabet $\mathcal{Z}$ and let R_Z be a distribution supported on $\mathcal{Z}$ with all entries being positive, its ϵ-neighborhood $\mathcal{N}^{\mathcal{Z}}_{\epsilon}(R_Z)$ is defined as

$$\mathcal{N}^{\mathcal{Z}}_{\epsilon}(R_Z) \triangleq \left\{ P_Z \in \mathcal{P}^{\mathcal{Z}}\colon \sum_{z\in\mathcal{Z}} \frac{(P_Z(z) - R_Z(z))^2}{R_Z(z)} \le \epsilon^2 \right\},$$

with $\mathcal{P}^{\mathcal{Z}}$ denoting the set of all distributions supported on $\mathcal{Z}$.

Then, with R_Z used as the reference distribution, each distribution $P_Z \in \mathcal{P}^{\mathcal{Z}}$ can be equivalently expressed as a vector $\phi \in \mathbb{R}^{|\mathcal{Z}|}$ or a function $f\colon \mathcal{Z} \to \mathbb{R}$ with

$$\phi(z) \triangleq \frac{P_Z(z) - R_Z(z)}{\sqrt{R_Z(z)}},\ f(z) \triangleq \frac{\phi(z)}{\sqrt{R_Z(z)}}, \quad \forall\, z \in \mathcal{Z}, \quad (11)$$

referred to as the *information vector* and *feature function* associated with P_Z, respectively. This provides a three way correspondence $P_Z \leftrightarrow \phi \leftrightarrow f$, which will be useful in our derivations.

Notations: For the ease of illustration, we focus on the discrete case with the assumption that each alphabet $\mathcal{X}_k$ is discrete. We also define $d \triangleq \sum_{i=1}^K d_i$ and $\mathcal{X} \triangleq \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$, and use x^K to denote $(x_1,\ldots,x_K) \in \mathcal{X}$. In addition, for a joint distribution $Q_{X^K} \in \mathcal{P}^{\mathcal{X}}$, we use $[Q_{X^K}]_{X_k}$ to denote its marginal distribution with respect to X_k. Moreover, we define $\mathcal{F}$ as the set of all functions $h\colon \mathcal{X} \to \mathbb{R}$ that satisfy $h(x^K) = \sum_{k=1}^K f_i(x_i)$ for some $f_1,\ldots,f_K$ with $f_k\colon \mathcal{X}_k \to \mathbb{R}, k = 1,\ldots,K$. Finally, given a matrix $A \in \mathbb{R}^{m_1\times m_2}$, we use $A^{\dagger}$ to denote its Moore–Penrose inverse [32], and also define the associated column space $\mathcal{R}(A) \triangleq \{Ax\colon x \in \mathbb{R}^{m_2}\}$ and projection matrix $\Pi_A \triangleq AA^{\dagger}$.

III. Distributed Hypothesis Testing

A. Optimal Feature and Geometric Structure

The following definitions of exponential and linear families will be useful for delineating our results.

Definition 3 (Exponential family): Given distribution $P_Z(z)$, and a function $T\colon \mathcal{Z} \to \mathbb{R}$, we define the distribution $\tilde{P}^{(\lambda)}_Z(\,\cdot\,;T,P_Z)$ as

$$\tilde{P}^{(\lambda)}_Z(z;T,P_Z) \triangleq P_Z(z)\exp(\lambda T(z) - \alpha(\lambda)), \quad \text{for all } z \in \mathcal{Z}, \quad (12)$$

with $\alpha(\lambda) \triangleq \log\sum_{z'\in Z} P_Z(z')\exp(\lambda T(z'))$. In addition, we use $\mathcal{E}_{\mathcal{Z}}(T,P_Z) \triangleq \left\{\tilde{P}^{(\lambda)}_Z(\,\cdot\,;T,P_Z)\colon \lambda\in\mathbb{R}\right\}$ to denote the exponential family passing through P_Z with T being the natural statistic.

Definition 4: Given a function $h\colon \mathcal{Z} \to \mathbb{R}$, we define the linear family $\mathcal{L}_{\mathcal{Z}}(h)$ as

$$\mathcal{L}_{\mathcal{Z}}(h) \triangleq \left\{Q_Z \in \mathcal{P}^{\mathcal{Z}}\colon \mathbb{E}_{Q_Z}[h(Z)] = 0\right\}. \quad (13)$$

In addition, we define the half-spaces $\mathcal{S}^{(0)}_{\mathcal{Z}}(h)$ and $\mathcal{S}^{(1)}_{\mathcal{Z}}(h)$ as

$$\mathcal{S}^{(0)}_{\mathcal{Z}}(h) \triangleq \left\{Q_Z \in \mathcal{P}^{\mathcal{Z}}\colon \mathbb{E}_{Q_Z}[h(Z)] \le 0\right\},$$
$$\mathcal{S}^{(1)}_{\mathcal{Z}}(h) \triangleq \left\{Q_Z \in \mathcal{P}^{\mathcal{Z}}\colon \mathbb{E}_{Q_Z}[h(Z)] \ge 0\right\}.$$

Then, the following result characterizes the optimal exponent E^* and the corresponding feature functions.

Theorem 1: The optimal exponent E^* in (4) is[1]

$$E^* = D\big(\mathcal{S}^{(0)}_{\mathcal{X}}(h^*)\big\|P^{(1)}_{X^K}\big) = D\big(\mathcal{S}^{(1)}_{\mathcal{X}}(h^*)\big\|P^{(0)}_{X^K}\big) \quad (14)$$

and can be achieved by one-dimensional features $f^*_i\colon \mathcal{X}_i \to \mathbb{R}$, where the function $h^*\colon \mathcal{X} \to \mathbb{R}$ and f^*_i's satisfy $h^*(x^K) = \sum_{i=1}^K f^*_i(x_i)$ and the equations

$$D\big(Q^{(0)}_{X^K}\big\|P^{(0)}_{X^K}\big) = D\big(Q^{(1)}_{X^K}\big\|P^{(1)}_{X^K}\big), \quad (15a)$$
$$\big[Q^{(0)}_{X^K}\big]_{X_i} = \big[Q^{(1)}_{X^K}\big]_{X_i}, \text{ for all } i = 1,\ldots,K, \quad (15b)$$
$$\mathbb{E}_{Q^{(0)}_{X^K}}\big[h^*(X^K)\big] = 0, \quad (15c)$$
$$\mathbb{E}_{P^{(0)}_{X^K}}\big[h^*(X^K)\big] < 0, \quad \mathbb{E}_{P^{(1)}_{X^K}}\big[h^*(X^K)\big] > 0, \quad (15d)$$

[1]For given $P_Z \in \mathcal{P}^{\mathcal{Z}}$ and $\mathcal{S} \subset \mathcal{P}^{\mathcal{Z}}$, we adopt the notation [33], [34] $D(\mathcal{S}\|P_Z) \triangleq \inf_{Q_Z\in\mathcal{S}} D(Q_Z\|P_Z)$.

where we have defined $Q^{(0)}_{X^K} \triangleq \tilde{P}^{(\frac{1}{2})}_{X^K}(\,\cdot\,; h^*, P^{(0)}_{X^k})$ and $Q^{(1)}_{X^K} \triangleq \tilde{P}^{(\lambda)}_{X^K}(\,\cdot\,; h^*, P^{(1)}_{X^k})$ with λ being an auxiliary parameter determined by the equations.

Proof: The proof is organized in two parts. In the first part, we show the existence of the optimal error exponent E^* and the one-dimensional features $f_1^*, \ldots, f_K^*$, and illustrate the optimal features satisfy equations (14) and (15). In the second part, we prove that each solution of (15) corresponds to the optimal error exponent and features.

To begin, for given marginal distributions $R_{X_k} \in \mathcal{P}^{\mathcal{X}_k}$, $k = 1, \ldots, K$ and $i = 0, 1$, we use $D_i^*(R_{X_1}, \ldots, R_{X_K})$ to denote the optimal value of the following K-L divergence minimization problem

$$\begin{aligned} \text{minimize} \quad & D(Q_{X^K} \| P^{(i)}_{X^K}) && \text{(16a)} \\ \text{subject to} \quad & [Q_{X^K}]_{X_i} = R_{X_i}, \quad i = 1, \ldots, K. && \text{(16b)} \end{aligned}$$

Then, for $i = 0, 1$ and $t > 0$, we define the sets

$$\mathcal{D}_i(t) \triangleq \{(R_{X_1}, \ldots, R_{X_K})\colon D_i^*(R_{X_1}, \ldots, R_{X_K}) < t\},$$

and also define $\mathcal{D}(t) \triangleq \mathcal{D}_0(t) \cap \mathcal{D}_1(t)$. It can be verified that, for all $t \geq 0$, both $\mathcal{D}_0(t)$ and $\mathcal{D}_1(t)$ are convex subsets of $\mathcal{P}^{\mathcal{X}_1} \times \cdots \times \mathcal{P}^{\mathcal{X}_K}$, and thus $\mathcal{D}(t)$ is also convex.

In addition, we have the following lemma.

Lemma 1: If $D(P^{(1)}_{X^K} \| P^{(0)}_{X^K}) < \infty, D(P^{(0)}_{X^K} \| P^{(1)}_{X^K}) < \infty$, there exists $t_0 > 0$ such that $\mathcal{D}(t) = \varnothing$ for all $t \in [0, t_0]$ and $\mathcal{D}(t) \neq \varnothing$ for all $t > t_0$. In addition, there exists a unique $(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) \in \mathcal{P}^{\mathcal{X}_1} \times \cdots \times \mathcal{P}^{\mathcal{X}_K}$ such that

$$D_0^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = D_1^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = t_0. \quad (17)$$

Proof: Note that since

$$D_0^*(P^{(1)}_{X_1}, \ldots, P^{(1)}_{X_K}) \leq D(P^{(1)}_{X^K} \| P^{(0)}_{X^K})$$

and

$$D_1^*(P^{(0)}_{X_1}, \ldots, P^{(0)}_{X_K}) \leq D(P^{(0)}_{X^K} \| P^{(1)}_{X^K}),$$

we have $\mathcal{D}(\tilde{t}) \neq \varnothing$, where

$$\tilde{t} \triangleq \min\{D(P^{(0)}_{X^K} \| P^{(1)}_{X^K}), D(P^{(1)}_{X^K} \| P^{(0)}_{X^K})\}.$$

Moreover, from $D(0) = \varnothing$ and

$$\mathcal{D}(t_1) \subset \mathcal{D}(t_2), \quad \text{for all } 0 \leq t_1 \leq t_2, \quad (18)$$

we can define

$$t_0 \triangleq \sup\{t \geq 0\colon \mathcal{D}(t) = \varnothing\}. \quad (19)$$

In addition, we have

$$\mathcal{D}(t) \neq \varnothing \implies \mathcal{D}(t - \epsilon) \neq \varnothing \text{ for some } \epsilon > 0. \quad (20)$$

Indeed, since $\mathcal{D}(t)$ is non-empty, there exists $(R_{X_1}, \ldots, R_{X_K})$ and $\epsilon > 0$ such that $D_i^*(R_{X_1}, \ldots, R_{X_K}) < t - \epsilon$ for $i = 0, 1$, and thus $\mathcal{D}(t - \epsilon)$ is non-empty.

Therefore, from (18) – (20) we obtain $\mathcal{D}(t) \neq \varnothing$ for all $t > t_0$ and $\mathcal{D}(t) = \varnothing$ for all $t \leq t_0$.

Furthermore, to prove (17), we define

$$\bar{\mathcal{D}}_i(t) \triangleq \{(R_{X_1}, \ldots, R_{X_K})\colon D_i^*(R_{X_1}, \ldots, R_{X_K}) \leq t\},$$

and $\bar{\mathcal{D}}(t) \triangleq \bar{\mathcal{D}}_0(t) \cap \bar{\mathcal{D}}_1(t)$. Then, for all $t > t_0$, we have

$$\begin{aligned} &\min_{R_{X_1}, \ldots, R_{X_K}} \max_{i \in \{0,1\}} D_i^*(R_{X_1}, \ldots, R_{X_K}) \\ &= \min_{(R_{X_1}, \ldots, R_{X_K}) \in \bar{\mathcal{D}}(t)} \max_{i \in \{0,1\}} D_i^*(R_{X_1}, \ldots, R_{X_K}) \in [t_0, t], \end{aligned}$$

where the second minimum exists since $\bar{\mathcal{D}}(t)$ is closed and bounded.

This implies that

$$t_0 = \min_{R_{X_1}, \ldots, R_{X_K}} \max_{i \in \{0,1\}} D_i^*(R_{X_1}, \ldots, R_{X_K}). \quad (21)$$

Hence, there exist marginal distributions $\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}$ such that

$$D_i^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = t_0, \quad i = 0, 1. \quad (22)$$

Finally, to illustrate the uniqueness of $(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K})$, suppose that (17) also holds for $(\tilde{R}'_{X_1}, \ldots, \tilde{R}'_{X_K}) \neq (\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K})$. Let $\tilde{R}''_{X_k} \triangleq (\tilde{R}_{X_k} + \tilde{R}'_{X_k})/2$ for $k = 1, \ldots, K$, then it follows from the strong convexities of $D_i^*(i = 0, 1)$ that

$$D_i^*(\tilde{R}''_{X_1}, \ldots, \tilde{R}''_{X_K}) < t_0, \quad i = 0, 1,$$

which contradicts (21). Therefore, $(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K})$ is unique. ∎

We then illustrate that the optimal error exponent $E^* = t_0$.

To see this, first note that from the Markov relation $\mathsf{H} - (\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K}) - (u_1, \ldots, u_K)$, the minimum possible decision error can be obtained when we choose the empirical distributions $\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K}$ themselves as the statistics $u_1, \ldots, u_K$.

Then, it follows from Sanov's theorem (see, e.g., [8, Theorem 11.4.1]) that

$$\mathbb{P}_n\left\{\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K} \middle| \mathsf{H} = i\right\} \doteq \exp(-nD_i^*(\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K})),$$

for $i = 0, 1$, where "$\doteq$" is the conventional dot-equal notation.[2] Therefore, in the asymptotic regime, the corresponding MAP decision rule (2) based on the empirical distributions $\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K}$ can be expressed as

$$D_0^*(\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K}) \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} D_1^*(\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K}), \quad (23)$$

with the decision error

$$\begin{aligned} &\mathbb{P}_n\left\{\hat{\mathsf{H}} \neq \mathsf{H}\right\} \\ &= P_{\mathsf{H}}(0) \cdot \mathbb{P}_n\left\{\hat{\mathsf{H}} = 1 \middle| \mathsf{H} = 0\right\} + P_{\mathsf{H}}(1) \cdot \mathbb{P}_n\left\{\hat{\mathsf{H}} = 0 \middle| \mathsf{H} = 1\right\} \\ &\doteq \exp\left(-n \min_{\hat{P}_{\mathsf{X}_i}, \ldots, \hat{P}_{\mathsf{X}_K}} \max_{i \in \{0,1\}} D_i^*(\hat{P}_{\mathsf{X}_1}, \ldots, \hat{P}_{\mathsf{X}_K})\right) && \text{(24)} \\ &\doteq \exp\left(-n \min_{R_{X_1}, \ldots, R_{X_K}} \max_{i \in \{0,1\}} D_i^*(R_{X_1}, \ldots, R_{X_K})\right) && \text{(25)} \\ &= \exp\left(-nt_0\right). && \text{(26)} \end{aligned}$$

[2] In particular, we use $a_n \doteq \exp(nb)$ to denote

$$\lim_{n \to \infty} \frac{1}{n} \log a_n = b.$$

where to obtain (25) we have used the fact that the set of all empirical distributions are dense in the space $\mathcal{P}^{\mathcal{X}_1}\times\cdots\times\mathcal{P}^{\mathcal{X}_K}$, and where the last equality follows from (21).

Then, it follows from separating hyperplane theorem (see, e.g., [35, Section 2.5.1]) that there exist functions $(f_1^*,\ldots,f_K^*)$ with $f_k^*\colon \mathcal{X}_k\to\mathbb{R}$ such that

$$\sum_{i=1}^K\sum_{x_i\in\mathcal{X}_i} R_{X_i}(x_i)f_i^*(x_i)=\sum_{i=1}^K\mathbb{E}_{R_{X_i}}\left[f_i^*(X_i)\right]\le 0$$

for all $(R_{X_1},\ldots,R_{X_K})\in\mathcal{D}_0(E^*)$ and

$$\sum_{i=1}^K\mathbb{E}_{R_{X_i}}\left[f_i^*(X_i)\right]\ge 0$$

for all $(R_{X_1},\ldots,R_{X_K})\in\mathcal{D}_1(E^*)$.

Therefore, we have $\mathcal{D}_i(E^*)\subset\mathcal{S}_{\mathcal{X}}^{(i)}(h^*)$, for $i=0,1$, where we have defined $h^*\colon\mathcal{X}\to\mathbb{R}$ as

$$h^*(x^K)\triangleq\sum_{k=1}^K f_k^*(x_k). \tag{27}$$

This implies that $\mathcal{S}_{\mathcal{X}}^{(i)}(h^*)\subset\mathcal{D}_{1-i}^{\mathsf{c}}(E^*)$, where for $t\ge 0$ and $i=0,1$, we have defined $\mathcal{D}_i^{\mathsf{c}}(t)\triangleq(\mathcal{P}^{\mathcal{X}_1}\times\cdots\times\mathcal{P}^{\mathcal{X}_K})\setminus\mathcal{D}_i(t)$.

Moreover, let $(\tilde R_{X_1},\ldots,\tilde R_{X_K})\in\mathcal{P}^{\mathcal{X}_1}\times\cdots\times\mathcal{P}^{\mathcal{X}_K}$ be as defined in Lemma 1, then we have

$$(\tilde R_{X_1},\ldots,\tilde R_{X_K})\in\mathcal{L}_{\mathcal{X}}(h^*)=\mathcal{S}_{\mathcal{X}}^{(0)}(h^*)\cap\mathcal{S}_{\mathcal{X}}^{(1)}(h^*). \tag{28}$$

As a result, for $i=0,1$, we have

$$\begin{aligned}E^*&=D_i^*(\tilde R_{X_1},\ldots,\tilde R_{X_K})\\&\ge D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)\big\|P_{X^K}^{(i)}\big)\\&=\min_{(R_{X_1},\ldots,R_{X_K})\in\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)}D_i^*(R_{X_1},\ldots,R_{X_K})\\&\ge\min_{(R_{X_1},\ldots,R_{X_K})\in\mathcal{D}_i^{\mathsf{c}}(E^*)}D_i^*(R_{X_1},\ldots,R_{X_K})\\&\ge E^*,\end{aligned} \tag{29}$$

which implies (14).

In addition, from [33, Corollary 3.1], there exist $\lambda_0,\lambda_1\in\mathbb{R}$ such that

$$Q_{X^K}^{(i)}\triangleq\tilde P_{X^K}^{(\lambda_i)}(\,\cdot\,;h^*,P_{X^K}^{(i)})\quad i=0,1, \tag{30}$$

satisfy

$$D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)\big\|P_{X^K}^{(i)}\big)=D\big(Q_{X^K}^{(i)}\big\|P_{X^K}^{(i)}\big), \tag{31}$$

where $\tilde P_{X^K}^{(\lambda_i)}(\,\cdot\,;h^*,P_{X^K}^{(i)}),i=0,1$ are as defined in (12). Since for each $i=0,1$, $Q_{X^K}^{(i)}$ depends only on the product $\lambda_i h^*$, we may assume $\lambda_0=\frac{1}{2}$ and simply use λ to denote λ_1.

Furthermore, for each $i=0,1$, from (29) and (31) we have,

$$\begin{aligned}&D_i^*(\tilde R_{X_1},\ldots,\tilde R_{X_K})\\&=\min_{(R_{X_1},\ldots,R_{X_K})\in\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)}D_i^*(R_{X_1},\ldots,R_{X_K})\\&=D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)\big\|P_{X^K}^{(i)}\big)\\&=D\big(Q_{X^K}^{(i)}\big\|P_{X^K}^{(i)}\big).\end{aligned}$$

Then, for each $i=0,1$, from the strong convexities of $D_i^*(\cdot)$ and K-L divergence $D(\,\cdot\,\|P_{X^K}^{(i)})$, the marginal distributions of $Q_{X^K}^{(i)}$ must be $\tilde R_{X_k},k=1,\ldots,K$, which implies (15b). As a result, the optimal error exponent and features satisfy (15).

For the second part of the proof, let E' denote the value of (15a), i.e.,

$$E'=D\big(Q_{X^K}^{(0)}\big\|P_{X^K}^{(0)}\big)=D\big(Q_{X^K}^{(1)}\big\|P_{X^K}^{(1)}\big). \tag{32}$$

Then, from (15b) and (15c), it can be verified that (see, e.g., [33, Corollary 3.1]) for each $i=0,1$, $Q_{X^K}^{(i)}$ is the unique intersection of the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$ and the exponential family $\mathcal{E}_{\mathcal{X}}\big(h^*,P_{X^K}^{(i)}\big)$, and satisfies (31).

Moreover, note that for each $i=0,1$, we have $D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)\big\|P_{X^K}^{(i)}\big)\le E^*$ if $(\tilde R_{X_1},\ldots,\tilde R_{X_K})\in\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)$. As a result, we obtain

$$E'=\min\big\{D\big(\mathcal{S}_{\mathcal{X}}^{(1)}(h^*)\big\|P_{X^K}^{(0)}\big),D\big(\mathcal{S}_{\mathcal{X}}^{(0)}(h^*)\big\|P_{X^K}^{(1)}\big)\big\}\le E^*.$$

In addition, let $\tilde Q_{X_k}\triangleq[Q_{X^K}^{(0)}]_{X_k}=[Q_{X^K}^{(1)}]_{X_k}$ denote the marginal distributions in (15b), then from the definition of $D_i^*(\cdot)$, we have [cf. (16)]

$$D_i^*(\tilde Q_{X_1},\ldots,\tilde Q_{X_K})\le D\big(Q_{X^K}^{(i)}\big\|P_{X^K}^{(i)}\big)=E',\quad\text{for } i=0,1.$$

Therefore, it follows from the definition of $\mathcal{D}(\cdot)$ that

$$(\tilde Q_{X_1},\ldots,\tilde Q_{X_K})\in\mathcal{D}_0(E')\cap\mathcal{D}_1(E')=\mathcal{D}(E'),$$

which implies that $\mathcal{D}(E')\neq\varnothing$, and thus $E'\ge E^*$.

Hence, we have $E'=E^*$ and (14) can be readily obtained from (31) and (32). ∎

While the computation of the log-likelihood function (2) for the distributed setting is typically intractable, the following proposition shows that the optimal error exponent E^* can be achieved by an efficiently computable decision rule.

Proposition 1: Let $f_1^*,\ldots,f_K^*$ denote the optimal features as defined in Theorem 1, then the exponent E^* can be achieved under the distributed decision settings where

1) The k-th node computes the feature u_k according to $u_k(\mathsf{X}_k)=\mathbb{E}_{\hat P_{\mathsf{X}_k}}\left[f_k^*(X_k)\right]$, for all $k=1,\ldots,K$;
2) The central machine decides $\hat{\mathsf{H}}$ based on

$$\sum_{k=1}^K u_k\underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}}0. \tag{33}$$

Proof: From Theorem 1, the error exponents associated with the type I error $\mathbb{P}_n\left\{\hat{\mathsf{H}}=1\middle|\mathsf{H}=0\right\}$ and the type II error $\mathbb{P}_n\left\{\hat{\mathsf{H}}=0\middle|\mathsf{H}=1\right\}$ are $D\big(\mathcal{S}_{\mathcal{X}}^{(1)}(h^*)\big\|P_{X^K}^{(0)}\big)$ and $D\big(\mathcal{S}_{\mathcal{X}}^{(0)}(h^*)\big\|P_{X^K}^{(1)}\big)$, respectively. From (14), both exponents are E^*, and thus the error exponent for $\mathbb{P}_n\left\{\hat{\mathsf{H}}\neq\mathsf{H}\right\}$ is also E^*. ∎

The geometry associated with Theorem 1 and Proposition 1 is depicted in Fig. 2. In this figure, each point represents a distribution in $\mathcal{P}^{\mathcal{X}}$, and the decision boundary (33) corresponds to the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$ defined as in (13). In addition, the $Q_{X^K}^{(0)}$ and $Q_{X^K}^{(1)}$ in (14) are the I-projections [33] of $P_{X^K}^{(0)}$ and

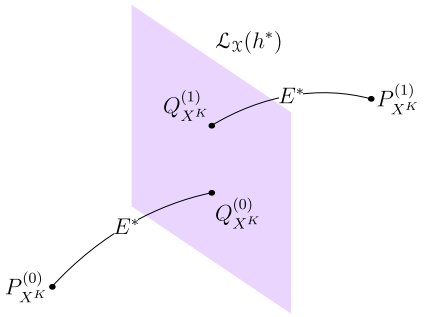


Fig. 2. The geometric structure in distributed hypothesis testing, with $Q_{X^K}^{(i)}$ denoting the I-projection of $P_{X^K}^{(i)}$ onto the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$, $i = 0, 1$.

$P_{X^K}^{(1)}$ onto this linear family, respectively, which also induces the two exponential families $\mathcal{E}_{\mathcal{X}}\big(h^*, P_{X^K}^{(0)}\big)$ and $\mathcal{E}_{\mathcal{X}}\big(h^*, P_{X^K}^{(1)}\big)$ with h^* as their common natural statistic.

B. Local Information Geometric Analysis

In this section, we apply local information geometric framework [4] to provide fundamental insights of this problem.

To begin, we introduce the local assumption that

$$P_{X^k}^{(i)} \in \mathcal{N}_{\epsilon}^{\mathcal{X}}(P_{X^k}) \quad \text{for } i = 0, 1 \tag{34}$$

for some small $\epsilon > 0$ and a reference distribution P_{X^K} with $P_{X^K}(x^K) > 0$ for all $x^K \in \mathcal{X}$, where $\mathcal{N}_{\epsilon}^{\mathcal{X}}(\cdot)$ is as defined in Definition 2. Then, we use $\psi^{(i)} \leftrightarrow P_{X^K}^{(i)}, i = 0, 1$, to denote the corresponding information vectors [cf. (11)]. For each $k = 1, \ldots, K$, and given feature $f_k \colon \mathcal{X}_k \to \mathbb{R}$, we define the corresponding information vector $\phi_k \in \mathbb{R}^{|\mathcal{X}_k|}$, with $P_{X_k} \triangleq [P_{X^K}]_{X_k}$ used as the reference distribution. Note that for $i = 0, 1$, we have the correspondence $B_k^{\mathrm{T}} \psi^{(i)} \leftrightarrow P_{X_k}^{(i)}$ with $P_{X_k}^{(i)} \triangleq [P_{X^K}^{(i)}]_{X_k}$ denoting the corresponding marginal distributions, where B_k is an $|\mathcal{X}| \times |\mathcal{X}_k|$ dimensional matrix with entries [36]

$$B_k(x^K, \hat{x}_k) \triangleq \sqrt{\frac{P_{X^K}(x^K)}{P_{X_k}(\hat{x}_k)}} \delta_{x_k \hat{x}_k}, \tag{35}$$

where $\delta_{x_k \hat{x}_k}$ represents the Kronecker delta.

Moreover, the feature f_k defined on $\mathcal{X}_k$, when regarded as a mapping from $\mathcal{X}$ to $\mathbb{R}$, has the corresponding information vector $B_k \phi_k$.

Using this correspondence, we can further establish the information vector for each for functions in $\mathcal{F}$. In particular, given $h \in \mathcal{F}$ with $h(x^K) = \sum_{k=1}^K f_k(X_k)$ for some one-dimensional features $f_k \colon \mathcal{X}_k \to \mathbb{R}$, $k = 1, \ldots, K$, the information vector corresponding to h is

$$\sum_{i=1}^K B_i \phi_i = B_0 \phi_0 \in \mathbb{R}^{|\mathcal{X}|}, \tag{36}$$

where we have defined

$$B_0 \triangleq \begin{bmatrix} B_1 & \cdots & B_K \end{bmatrix} \quad \text{and} \quad \phi_0 \triangleq \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_K \end{bmatrix}, \tag{37}$$

and where for each $k = 1, \ldots, K$, $\phi_k \in \mathbb{R}^{|\mathcal{X}_k|}$ denotes the information vector corresponding to f_k.

Then, we can establish the local counterpart of Theorem 1 as follows.

Theorem 2: Under the local assumption (34), let $\psi^{(i)} \leftrightarrow P_{X^K}^{(i)}$, $i = 0, 1$, denote the corresponding information vectors. Then, we have the correspondence $h^* \leftrightarrow B_0 \phi_0^*$, where h^* is as defined in Theorem 1, where B_0 is as defined in (37), and where ϕ_0^* is given by

$$\phi_0^* \triangleq B_0^{\dagger}\big(\psi^{(1)} - \psi^{(0)}\big). \tag{38}$$

In addition, the optimal exponent E^* can be expressed as

$$E^* = \frac{1}{8}\big\|B_0 \phi_0^*\big\|^2 + o(\epsilon^2).$$

Proof: To begin, we define $\psi \triangleq \psi^{(1)} - \psi^{(0)}$. Then, for given $f_k \colon \mathcal{X}_k \to \mathbb{R},$, it follows from [4, Lemma 17] that the exponent based on the feature $h(x^K) = \sum_{k=1}^K f_k(x_i)$ is

$$E = \frac{1}{8} \cdot \frac{\langle \psi, \xi \rangle^2}{\|\xi\|^2} + o(\epsilon^2)$$

where we have defined $\xi \triangleq B_0 \phi_0 \in \mathcal{R}(B_0)$ and where ϕ_0 is as defined in (37).

Then, note that the projection matrix Π_{B_0} satisfies $\Pi_{B_0} = (\Pi_{B_0})^2$ and $\xi = \Pi_{B_0} \xi$. Therefore, from Cauchy–Schwarz inequality we have

$$\frac{\langle \psi, \xi \rangle^2}{\|\xi\|^2} = \frac{(\psi^{\mathrm{T}} \Pi_{B_0} \xi)^2}{\|\xi\|^2} = \frac{\langle \Pi_{B_0} \psi, \xi \rangle^2}{\|\xi\|^2} \le \big\|\Pi_{B_0} \psi\big\|^2$$

where the inequality holds with equality if and only if ξ takes the optimal values

$$\xi^* = c \cdot \Pi_{B_0} \psi,$$

or equivalently, $B_0 \phi_0^* = c \cdot B_0 B_0^{\dagger} \psi$ for some constant scalar $c \neq 0$.

To determine the value of c, note that we have $\xi^* \leftrightarrow h^*$ where h^* is the optimal feature as defined in Theorem 1.Then, from (15) we have

$$\begin{aligned}
&Q_{X^K}^{(0)}(x^K) \\
&= \tilde{P}_{X^K}^{(\frac{1}{2})}(x^K; h^*, P_{X^K}^{(0)}) \\
&= P_{X^K}^{(0)}(x^K) \left[1 + \frac{1}{2}\left(h^*(x^K) - \mathbb{E}_{P_{X^K}^{(0)}}\big[h^*(X^K)\big]\right)\right] + o(\epsilon) \\
&= \left(P_{X^K}(x^K) + \sqrt{P_{X^K}(x^K)}\psi^{(0)}(x^K)\right) \\
&\quad \cdot \left[1 + \frac{1}{2}\sqrt{P_{X^K}(x^K)}\xi(x^K)\right] + o(\epsilon) \\
&= P_{X^K}(x^K) + \sqrt{P_{X^K}(x^K)} \cdot \left(\psi^{(0)}(x^K) + \frac{1}{2}\xi(x^K)\right) \\
&\quad + o(\epsilon)
\end{aligned}$$

which implies the correspondence

$$Q_{X^K}^{(0)}(x^K) \leftrightarrow \left(\psi^{(0)} + \frac{1}{2}\xi + o(\epsilon)\right).$$

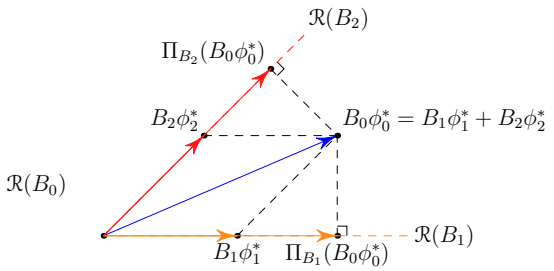


Fig. 3. The information decomposition structure in distributed hypothesis testing with $K = 2$ nodes, compared with the orthogonal decompositions on the subspace $\mathcal{R}(B_k)$ for each node $k = 1, 2$.

Similarly, we have

$$Q^{(1)}_{X^K}(x^K) \leftrightarrow \left(\psi^{(1)} + \lambda\xi + o(\epsilon)\right).$$

Then, it follows from the second-order Taylor series expansion of the K-L divergence that (see, e.g., [4, Lemma 10])

$$D\big(Q^{(0)}_{X^K}\big\|P^{(0)}_{X^K}\big) = \frac{1}{8}\|\xi\|^2 + o(\epsilon^2), \tag{39a}$$

$$D\big(Q^{(1)}_{X^K}\big\|P^{(1)}_{X^K}\big) = \frac{\lambda^2}{2}\|\xi\|^2 + o(\epsilon^2). \tag{39b}$$

Moreover, note that since (cf. [4, Lemma 9])

$$\mathbb{E}_{Q^{(0)}_{X^K}}\left[h^*(X^K)\right] = \left\langle \psi^{(0)} + \frac{1}{2}\xi, \xi\right\rangle + o(\epsilon^2),$$

$$\mathbb{E}_{Q^{(1)}_{X^K}}\left[h^*(X^K)\right] = \left\langle \psi^{(1)} + \lambda\xi, \xi\right\rangle + o(\epsilon^2),$$

it follows from (15c) and (15b) that

$$\begin{aligned}
0 &= \mathbb{E}_{Q^{(1)}_{X^K}}\left[h^*(X^K)\right] - \mathbb{E}_{Q^{(0)}_{X^K}}\left[h^*(X^K)\right] \\
&= \left\langle \psi + \left(\lambda - \frac{1}{2}\right)\xi, \xi\right\rangle + o(\epsilon^2) \\
&= c\left\langle \psi + \left(\lambda - \frac{1}{2}\right) c\cdot\Pi_{B_0}\psi, \Pi_{B_0}\psi\right\rangle + o(\epsilon^2) \\
&= c\cdot\left[1 + \left(\lambda - \frac{1}{2}\right)c\right]\cdot\|\Pi_{B_0}\psi\|^2 + o(\epsilon^2)
\end{aligned} \tag{40}$$

As a result, it follows from (39), (15a), and (40) that $c = 1, \lambda = -\frac{1}{2}$. Then, we obtain

$$\xi^* = \Pi_{B_0}\psi = B_0B_0^\dagger\psi = B_0\phi_0^*,$$

where $\phi_0^* \triangleq B_0^\dagger\psi$.

Finally, the optimal error exponent is

$$\begin{aligned}
E^* &= \frac{1}{8}\cdot\left\|\Pi_{B_0}\psi\right\|^2 + o(\epsilon^2) \\
&= \frac{1}{8}\cdot\left\|B_0\phi_0^*\right\|^2 + o(\epsilon^2).
\end{aligned}$$

■

Note that from Theorem 2, we have

$$h^* \leftrightarrow B_0B_0^\dagger(\psi^{(1)} - \psi^{(0)}) = \Pi_{B_0}(\psi^{(1)} - \psi^{(0)}),$$

where Π_{B_0} is the projection matrix associated to the subspace $\mathcal{R}(B_0)$. Thus, the optimal feature $B_0\phi_0^*$ (36) is the projection of the sufficient statistic $f_{\text{LLR}} \leftrightarrow (\psi^{(1)} - \psi^{(0)})$ onto the function space $\mathcal{F}$.

Moreover, from (36), this optimal feature can be decomposed to K components in subspaces $\mathcal{R}(B_k)$, for $k = 1, \ldots, K$,

$$B_0\phi_0^* = \sum_{k=1}^K B_k\phi_k^*, \tag{41}$$

where ϕ_0^* is stacked by $\phi_k^* \in \mathbb{R}^{|\mathcal{X}_k|}$, $k = 1, \ldots, K$, as in (37). This decomposition structure can be depicted as Fig. 3 for the case $K = 2$.

Remark 1: Note that $B_i\phi_k^*$ are not simply the orthogonal projections of $B_0\phi_0^*$ onto the subspaces $\mathcal{R}(B_k)$ since the subspaces $\mathcal{R}(B_k)$, for $k = 1, \ldots, K$ are not mutually orthogonal to each other. Thus, the decomposition of $B_0\phi_0^*$ will depend on the Gram matrix [32] of the subspaces $\mathcal{R}(B_k)$ as shown in Fig. 3. Moreover, it can be shown that the orthogonal projection of $B_0\phi_0^*$ onto the subspaces $\mathcal{R}(B_k)$ can be interpreted as characterizing the optimal error exponent of the binary hypothesis testing problem with only the observations of X_k [8]. When the subspaces $\mathcal{R}(B_k)$ are orthogonal to each other, the optimal inference approach is simply extracting the optimal information from each node by the orthogonal projection. However, when $\mathcal{R}(B_k)$ are not orthogonal, different nodes may share different kinds of common information. Our result essentially shows how to deal with the common information and extract the optimal features by the decomposition of the information vector over non-orthogonal subspaces.

IV. Distributed Classification

In this section, we provide the information theoretic optimality for MCR in distributed problems.

To begin, let $P_{X^K} \triangleq [P_{X^K,Y}]_{X^K}$ denote the marginal distribution for the distributed data variables. Then, for each $k = 1, \ldots, K$, we define an $|\mathcal{X}_k| \times d_k$ matrix Φ_k corresponding to the d_k dimensional feature function $f_k = (f_k^{(1)}, \ldots, f_k^{(d_k)})^{\mathrm{T}}$, such that

$$\Phi_k = \left[\phi_k^{(1)} \quad \cdots \quad \phi_k^{(d_k)}\right],$$

where $\phi_k^{(j)} \leftrightarrow f_k^{(j)}$, $j = 1, \ldots, d_k$ are the corresponding information vectors as defined in (11), and where we have used the marginal distribution $P_{X_k} \triangleq [P_{X^K}]_{X_k}$ as the reference distribution. Furthermore, we define the $|\mathcal{X}| \times d$ matrix Φ as

$$\Phi \triangleq \left[B_1\Phi_1 \quad \cdots \quad B_K\Phi_K\right], \tag{42}$$

where $B_k, k = 1, \ldots, K$ are as defined in (35).

Then, it has been shown in [31] that the optimal MCR feature $f(X^K)$ in predicting Y is the feature that maximizes the (single-sided) H-score [37] of f, defined as

$$H(f) \triangleq \frac{1}{2}\left\|\tilde{B}_{Y,X^K}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2, \tag{43}$$

where $\tilde{B}_{Y,X^K} \in \mathbb{R}^{2\times|\mathcal{X}|}$ denotes the CDM of Y and X^K [cf. (6)], and where $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm.

In general, the optimal function f corresponds to the top singular vectors of $\tilde{B}_{Y,X^K}$ [4]. However, the function f is restricted to be a concatenated form due to (10). The following proposition shows that such constraint leads to the same projection and decomposition structure as in Section III-B.

Proposition 2: The feature f of the form (10) maximizes the H-score (43) if and only if $B_0\phi_0^* \in \mathcal{R}(\Phi)$, where B_0 and Φ are as defined in (37) and (42), respectively, where ϕ_0^* is given by

$$\phi_0^* \triangleq B_0^\dagger \left(\psi^{(1)} - \psi^{(0)}\right),$$

and where for each $y \in \{0,1\}$, $\psi^{(y)} \leftrightarrow P_{X^K|Y=y}$ represents the information vector for corresponding conditional distribution. In addition, the resulting optimal H-score is

$$H(f) = \frac{P_Y(0)P_Y(1)}{2} \cdot \|B_0\phi_0^*\|^2.$$

Proof: To begin, note that Φ can be written as

$$\Phi = \begin{bmatrix} B_1\phi_1 & \cdots & B_K\phi_K \end{bmatrix} = B_0 \begin{bmatrix} \phi_1 & & \\ & \ddots & \\ & & \phi_K \end{bmatrix}.$$

In addition, it follows from the definition of CDM (6) that

$$\begin{aligned} \tilde{B}_{Y,X^K}(y,x^K) &= \frac{P_{X^K,Y}(x^K,y) - P_{X^K}(x^K)P_Y(y)}{\sqrt{P_{X^K}(x^K)P_Y(y)}} \\ &= \sqrt{P_Y(y)} \cdot \frac{P_{X^K|Y=y}(x^K) - P_{X^K}(x^K)}{\sqrt{P_{X^K}(x^K)}} \\ &= \sqrt{P_Y(y)} \cdot \psi^{(y)}(x^K) \end{aligned} \tag{44}$$

where we have used the correspondence

$$\psi^{(y)} \leftrightarrow P_{X^K|Y=y}, \quad \text{for } y = 0, 1.$$

Moreover, let $\psi \triangleq \psi^{(1)} - \psi^{(0)}$, then we obtain

$$B_0^{\mathrm{T}}\psi = B_0^{\mathrm{T}}B_0B_0^\dagger\psi = B_0^{\mathrm{T}}B_0\phi_0^*,$$

where we have used the fact $B_0^{\mathrm{T}} = B_0^{\mathrm{T}}B_0B_0^\dagger$ of the Moore–Penrose inverse.

Therefore, we obtain

$$\begin{aligned} \Phi^{\mathrm{T}}\psi &= \begin{bmatrix} \phi_1 & & \\ & \ddots & \\ & & \phi_K \end{bmatrix}^{\mathrm{T}} B_0^{\mathrm{T}}\psi \\ &= \begin{bmatrix} \phi_1 & & \\ & \ddots & \\ & & \phi_K \end{bmatrix}^{\mathrm{T}} B_0^{\mathrm{T}}B_0\phi_0^* = \Phi^{\mathrm{T}}B_0\phi_0^*, \end{aligned}$$

and thus

$$\begin{aligned} H(f) &= \frac{1}{2} \cdot \left\|\tilde{B}_{Y,X^K}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\ &= \frac{1}{2} \sum_{y\in\{0,1\}} P_Y(y) \left\|\left(\psi^{(y)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 && (45) \\ &= \frac{P_Y(0)P_Y(1)}{2} \cdot \left\|\left(\psi^{(1)} - \psi^{(0)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 && (46) \\ &= \frac{P_Y(0)P_Y(1)}{2} \cdot \left\|(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\Phi^{\mathrm{T}}\left(\psi^{(1)} - \psi^{(0)}\right)\right\|^2 && (47) \\ &= \frac{P_Y(0)P_Y(1)}{2} \cdot \left\|(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\Phi^{\mathrm{T}}B_0\phi^*\right\|^2 && (48) \\ &\le \frac{P_Y(0)P_Y(1)}{2} \cdot \left\|(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\Phi^{\mathrm{T}}\right\|_{\mathrm{s}}^2 \|B_0\phi^*\|^2 && (49) \\ &= \frac{P_Y(0)P_Y(1)}{2} \cdot \|B_0\phi^*\|^2, && (50) \end{aligned}$$

where $\|\cdot\|_{\mathrm{s}}$ denotes the spectral norm, where to obtain (46) we have used the fact that

$$P_Y(0)\psi^{(0)} + P_Y(1)\psi^{(1)} = 0,$$

and where to obtain (49), we have used the submultiplicativity of the spectral norm. Finally, note that the inequality holds with equality if and only if $B_0\phi_0^* \in \mathcal{R}(\Phi)$. ∎

From Proposition 2, the optimality of f can be achieved by one-dimensional features $f_1, \ldots, f_K$ satisfying $f_i \leftrightarrow \phi_i, i = 1, \ldots, K$ and $\left[\phi_1^{\mathrm{T}}, \ldots, \phi_K^{\mathrm{T}}\right]^{\mathrm{T}} = \phi_0^*$. Therefore, when we have the correspondences $P_{X^K|Y=\ell} = P_{X^K}^{(\ell)}$, for $\ell = 0, 1$, the optimal feature for distributed classification is also optimal in hypothesis testing, which establishes the desired connection between information theory and machine learning.

V. Generalization to M-ary Settings

The above discussion can be further generalized to M-ary settings, where both the hypotheses space $\mathcal{H}$ the label space $\mathcal{Y}$ can take M values, e.g., $\{0, 1, \ldots, M-1\}$. For convenience, we assume that the distributions in these two settings have the correspondences $P_{\mathsf{H}} = P_Y$ and $P_{X^K}^{(\ell)} = P_{X^K|Y=\ell}$, for all $\ell \in \mathcal{H} = \mathcal{Y}$, with $P_{\mathsf{H}}(\cdot)$ denoting the prior of the hypotheses. We also introduce the local assumption $P_{X^K}^{(\ell)} \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_{X^K})$ for $\ell \in \mathcal{Y}$, with $P_{X^K} \triangleq [P_{X^K,Y}]_{X^K}$ denoting the corresponding marginal distribution.

Note that in the M-ary hypothesis testing, we can define the pairwise error exponents E_{ij}, for $i \neq j$, as

$$E_{ij} \triangleq \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_n\left\{\hat{\mathsf{H}} \neq \mathsf{H} \middle| \mathsf{H} \in \{i,j\}\right\},$$

where the decision $\hat{\mathsf{H}}$ is obtained according to MAP decision (2). In the following, we focus on characterizing the weighted sum of these error exponents, defined as

$$\bar{E} \triangleq \sum_{i\in\mathcal{H}} \sum_{j\in\mathcal{H}\setminus\{i\}} P_{\mathsf{H}}(i)P_{\mathsf{H}}(j)E_{ij}. \tag{51}$$

Then, we have the following result.

Proposition 3: For given feature f as defined in (10), the weighted error exponent $\bar{E}$ as defined in (51) satisfies

$$\bar{E} = \frac{1}{2}H(f) + o(\epsilon^2),$$

where $H(f)$ denotes the H-score of f, as defined in (43).

Proof: From [4, Lemma 17], For each (i, j) pair, the associated exponent E_{ij} is given by

$$E_{ij} = \frac{1}{8}\left\|\left(\psi^{(i)} - \psi^{(j)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 + o(\epsilon^2).$$

Therefore, we have

$$\begin{aligned}
E &= \sum_{i\in\mathcal{H}}\sum_{j\in\mathcal{H}\setminus\{i\}} P_{\mathsf{H}}(i)P_{\mathsf{H}}(j)E_{ij} \\
&= \frac{1}{8}\sum_{i\in\mathcal{Y}}\sum_{j\in\mathcal{Y}} P_Y(i)P_Y(j)\left\|\left(\psi^{(i)} - \psi^{(j)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 \\
&\quad + o(\epsilon^2) \\
&= \frac{1}{4}\sum_{i\in\mathcal{Y}} P_Y(i)\left\|\left(\psi^{(i)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 \\
&\quad - \frac{1}{4}\left(\sum_{i\in\mathcal{Y}} P_Y(i)\psi^{(i)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\left(\sum_{j\in\mathcal{Y}} P_Y(j)\psi^{(j)}\right) \\
&\quad + o(\epsilon^2) \\
&= \frac{1}{4}\sum_{i\in\mathcal{Y}} P_Y(i)\left\|\left(\psi^{(i)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2 + o(\epsilon^2),
\end{aligned}$$

where the last equality follows from the fact that

$$\sum_{y\in\mathcal{Y}} P_Y(y)\psi^{(y)} = 0.$$

Then, the conclusion can be readily obtained by noting that

$$\begin{aligned}
H(f) &= \frac{1}{2}\left\|\tilde{B}_{Y,X^K}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\
&= \frac{1}{2}\sum_{y\in\mathcal{Y}} P_Y(y)\left\|\left(\psi^{(y)}\right)^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right\|^2,
\end{aligned}$$

where to obtain the second equality we have used (44). ∎

Therefore, the optimal feature f^* for predicting the label Y coincides with the optimal feature that achieves the optimal exponent $\bar{E}^*$ in M-ary hypothesis testing, which again illustrates the fundamental connection between information theory and algorithms.

Acknowledgment

The work of Shao-Lun Huang was supported in part by the National Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146, and in part by the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

Appendix
Discussions on Decomposition (41)

As illustrated in Section III-B, the optimal feature $B_k\phi_k^*$ for node k in (41) is in general different from the orthogonal projection of $B_0\phi_0^*$ onto $\mathcal{R}(B_i)$, denoted by $\Pi_{B_k}(B_0\phi_0^*)$. The following proposition provides an interpretation for orthogonal projections $\Pi_{B_k}(B_0\phi_0^*), k = 1, \ldots, K$, and also demonstrates their optimality when $P_{X^K}^{(0)}$ and $P_{X^K}^{(1)}$ follow nearly independent structures.

Proposition 4: Let $\tilde{\phi}_k \triangleq B_k^{\mathrm{T}}(B_0\phi_0^*), k = 1, \ldots, K$, with ϕ_0^* as defined in (38), and let $f_k \leftrightarrow \tilde{\phi}_k$ denote the corresponding function as defined in (11), for $k = 1, \ldots, K$. Then, f_k is the optimal feature that optimizes the error exponent for the local MAP decision

$$\log\frac{\mathbb{P}\{u_k|\mathsf{H}=1\}}{\mathbb{P}\{u_k|\mathsf{H}=0\}} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} \log\frac{P_{\mathsf{H}}(0)}{P_{\mathsf{H}}(1)}, \tag{52}$$

with the optimal error exponent being

$$\tilde{E}_k^* = \frac{1}{8}\left\|\tilde{\phi}_k\right\|^2 + o(\epsilon^2) = \frac{1}{8}\left\|B_k\tilde{\phi}_k\right\|^2 + o(\epsilon^2). \tag{53}$$

In addition, when $P_{X^K} \in \mathcal{N}_{\mathcal{X}}^{\epsilon}(P_{X_1}\ldots P_{X_K})$, we have

$$\phi_k^* = \tilde{\phi}_k + o(\epsilon), \quad \text{for all } k = 1, \ldots, K, \tag{54}$$

$$\langle B_j\phi_j^*, B_k\phi_k^*\rangle = o(\epsilon^2), \quad \text{for all } 1 \le j < k \le K. \tag{55}$$

Proof: First, note that since $B_0^{\mathrm{T}}(I - \Pi_{B_0}) = O$, for $k = 1, \ldots, K$ we have $B_k^{\mathrm{T}}(I - \Pi_{B_0}) = O$, and thus

$$B_k^{\mathrm{T}}(\psi^{(1)} - \psi^{(0)}) = B_k^{\mathrm{T}}\Pi_{B_0}(\psi^{(1)} - \psi^{(0)}) = \tilde{\phi}_k. \tag{56}$$

Then, the optimality of $\tilde{\phi}_k$ and the exponent (53) can be readily obtained from [4, Lemma 17].

Finally, (54) and (55) can be obtained by noting that, when $P_{X^K} \in \mathcal{N}_{\mathcal{X}}^{\epsilon}(P_{X_1}\ldots P_{X_K})$, for all $j \neq k$, we have $P_{X_j,X_k} \in \mathcal{N}_{\mathcal{X}}^{\epsilon}(P_{X_j}P_{X_k})$ and $\tilde{B}_{X_j,X_k} = O(\epsilon)$, where $\tilde{B}_{X_j,X_k}$ denotes the corresponding CDM as defined in (6). ∎

References

[1] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

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

[6] 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.

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

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

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

[10] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[13] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2949–2980, 2014.

[14] T. S. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 2–14, 1989.

[15] S.-I. Amari and T. S. Han, "Statistical inference under multiterminal rate restrictions: a differential geometric approach," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 217–227, 1989.

[16] S. Watanabe, "Neyman–pearson test for zero-rate multiterminal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4923–4939, 2017.

[17] Te Sun Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.

[18] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE transactions on information theory*, vol. 33, no. 6, pp. 759–772, 1987.

[19] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE transactions on information theory*, vol. 32, no. 4, pp. 533–542, 1986.

[20] J. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Transactions on Automatic Control*, vol. 30, no. 5, pp. 440–446, 1985.

[21] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals and Systems*, vol. 1, no. 2, pp. 167–182, 1988.

[22] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic systems*, no. 4, pp. 501–510, 1981.

[23] H. M. Shalaby and A. Papamarcou, "Multiterminal detection with zero-rate data compression," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 254–267, 1992.

[24] W. Zhao and L. Lai, "Distributed testing with zero-rate compression," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 2792–2796.

[25] S. Scardapane, D. Wang, M. Panella, and A. Uncini, "Distributed learning for random vector functional-link networks," *Information Sciences*, vol. 301, pp. 271–284, 2015.

[26] L. Georgopoulos and M. Hasler, "Distributed machine learning in networks by consensus," *Neurocomputing*, vol. 124, pp. 2–12, 2014.

[27] A. Navia-Vázquez, D. Gutierrez-Gonzalez, E. Parrado-Hernández, and J. Navarro-Abellan, "Distributed support vector machines," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, p. 1091, 2006.

[28] Y. Lu, V. Roychowdhury, and L. Vandenberghe, "Distributed parallel support vector machines in strongly connected networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 7, pp. 1167–1178, 2008.

[29] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.

[30] 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, 2019, pp. 5281–5288.

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

[32] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[33] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[34] I. Csiszár, "The method of types [information theory]," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[36] S.-L. Huang, X. Xu, and L. Zheng, "An information-theoretic approach to unsupervised feature selection for high-dimensional data," *IEEE Journal on Selected Areas in Information Theory*, 2020.

[37] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1984–1988.