entropy

MDPI

Article

On the Optimal Error Exponent of Type-Based Distributed Hypothesis Testing †

Xinyi Tong [1], Xiangxiang Xu [2,‡] and Shao-Lun Huang [2,*]

[1] Tsinghua–Berkeley Shenzhen Institute, Shenzhen 518055, China; txy18@mails.tsinghua.edu.cn
[2] Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China; xuxx@mit.edu
* Correspondence: twn2gold@gmail.com
† This work was presented in part at the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Victoria, Australia, 12–20 July 2021.
‡ Current address: Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Abstract: Distributed hypothesis testing (DHT) has emerged as a significant research area, but the information-theoretic optimality of coding strategies is often typically hard to address. This paper studies the DHT problems under the type-based setting, which is requested from the popular federated learning methods. Specifically, two communication models are considered: (i) DHT problem over noiseless channels, where each node observes i.i.d. samples and sends a one-dimensional statistic of observed samples to the decision center for decision making; and (ii) DHT problem over AWGN channels, where the distributed nodes are restricted to transmit functions of the empirical distributions of the observed data sequences due to practical computational constraints. For both of these problems, we present the optimal error exponent by providing both the achievability and converse results. In addition, we offer corresponding coding strategies and decision rules. Our results not only offer coding guidance for distributed systems, but also have the potential to be applied to more complex problems, enhancing the understanding and application of DHT in various domains.

Keywords: hypothesis testing; distributed system; information theory; local geometry

Citation: Tong, X.; Xu, X.; Huang, S.-L. On the Optimal Error Exponent of Type-Based Distributed Hypothesis Testing. *Entropy* **2023**, *25*, 1434. https://doi.org/10.3390/e25101434

Academic Editors: T. Aaron Gulliver and Songze Li

Received: 4 August 2023
Revised: 1 October 2023
Accepted: 8 October 2023
Published: 10 October 2023

1. Introduction

Distributed hypothesis testing (DHT) is a significant problem in the field of information theory [1]. In this problem, each distributed node observes partial data generated from the joint distribution and transmits an encoded message through a communication channel to a decision center, aiming to detect the true hypothesis. The primary goal of DHT is to maximize the decision error exponent in the asymptotic regime, where many different communication models [2–6] were considered in the previous literature. The main challenges of the DHT arise in two respects. Firstly, due to the intricate distributed structures, most of the existing works have focused on demonstrating achievability results, with converse results being limited to specific cases, such as the 1-bit [3], $\log_2 3$-bit [7], and $O(\log_2 n)$-bit [1] communication channels. Secondly, many of the achievability results were established using random coding with auxiliary random variables [8], which are difficult to implemen in real systems.

Notice that the distributed encoders in many real applications are required to process high-dimensional data [9], such as images, texts, and audios. Consequently, many of the federated learning algorithms focus on computing the quantities, such as the statistics, empirical risks, and gradient of data [10], which can be viewed as certain functions of the empirical distribution (type) of the data (for example, given the data $x_1, \ldots, x_n$ and feature function $f(x)$, the statistic $\frac{1}{n}\sum_{i=1}^{n} f(x_i) = \sum_x \hat{P}_X(x) f(x)$ is a linear function of the empirical distribution $\hat{P}_X$).

Motivated by this observation, we investigate the optimal decision error exponent of DHT based on the empirical distributions (type-based) under two common communication models. The first problem considers a noiseless channel, which is the typical mathematical model in real federated learning scenarios. It comes from the reality that federated learning often assumes that the nodes and the center machine can exchange information precisely; however, the dimensionalities of the transmitted signals are limited [9]. Specifically, it is assumed that each node can only transmit the empirical mean of a one-dimensional feature, and such settings have gained significant attention recently in federated and multi-modal machine learning [9,11]. The second problem assumes that the signal of each node, encoded with the empirical distribution, is transmitted over an additive white Gaussian noise (AWGN) channel, which is a widely-used mathematical model for real-world channels [12]. The main goal of this paper is to establish the optimal error exponent for the aforementioned two problems by presenting: (i) the converse bound for the error exponent; and (ii) a practical coding strategy that achieves the converse bound.

The contributions of this paper are summarized as follows. First, in Section 4.1, we demonstrate the optimal error exponent for the type-based hypothesis testing over noiseless channels, where one-dimensional functions for all nodes and the corresponding decision rule are provided. Moreover, by applying the information geometric approach in [13], the hypotheses and the feature functions of each node can be modeled as vectors in the joint and marginal distribution spaces, respectively. In Section 4.3, 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.

Second, we establish the optimal achievable error of the type-based hypothesis testing over AWGN channels by presenting both the achievability and converse results. In particular, the achievability part is based on a mixture coding strategy of both the amplify-and-forward and decode-and-forward strategies. Specifically, when the observed empirical distribution at a distributed node is sufficiently close to one of the true marginal distributions with respect to the two hypotheses, the node is confident of the true hypothesis. Then, we apply the decode-and-forward strategy, which first estimates the true hypothesis based on the observed empirical distribution, and then we apply the binary phase shift keying (BPSK) to transmit the decoded bit to the decision center. On the other hand, when the observed empirical distribution is far from both true marginal distributions, we apply the amplify-and-forward strategy to encode and transmit the observed empirical distribution by the pulse amplitude modulation (PAM) to the decision center. By applying the proposed coding strategy and conducting the log-likelihood ratio test at the decision center, we show in Section 5.2 the achievable error exponent. Finally, we demonstrate the converse results of the error exponent in Section 5.3 based on a genie-aided approach. The main idea is to add additional information to the distributed nodes. By either leveraging the true hypothesis to the distributed nodes or eliminating the channel noises, we show that the error exponent in Section 5.2 is also an upper bound of the optimal error exponent, which establishes the optimality.

2. Problem Formulations

Suppose that there are K random variables $X^K \triangleq (X_1, \ldots, X_K)$. In this paper, we consider the binary hypothesis testing problem, and the two hypotheses H_0 and H_1 are defined as:

$$\begin{aligned} \mathsf{H}_0 &: (x_1^{(1)}, \cdots, x_K^{(1)}), \cdots, (x_1^{(n)}, \cdots, x_K^{(n)}) \overset{\text{i.i.d.}}{\sim} P_{X^K}^{(0)}, \\ \mathsf{H}_1 &: (x_1^{(1)}, \cdots, x_K^{(1)}), \cdots, (x_1^{(n)}, \cdots, x_K^{(n)}) \overset{\text{i.i.d.}}{\sim} P_{X^K}^{(1)}, \end{aligned} \tag{1}$$

where the observable data are i.i.d. generated according to either $P_{X^K}^{(0)}$ or $P_{X^K}^{(1)}$ from the alphabet set $(\mathcal{X}_1, \cdots, \mathcal{X}_K)$. In addition, we assume that there are K distributed nodes, where the k-th ($k = 1, \cdots, K$) node can only observe the samples $\mathsf{X}_k \triangleq \{x_k^{(1)}, \ldots, x_k^{(n)}\}$.

To facilitate clarity in our illustration, we concentrate on the discrete case, assuming that each alphabet $\mathcal{X}_k$ is discrete, and $\mathcal{X} \triangleq \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$. 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. We also denote $P^{(i)}_{X_1}, \cdots, P^{(i)}_{X_K}$ as the marginal distributions of $P^{(i)}_{X^K}$, for $i = 0, 1$. In the distributed hypothesis testing problem, we introduce a common assumption in the distributed setup [14] that the generating distributions $P^{(0)}_{X^K}$ and $P^{(1)}_{X^K}$ satisfy $D(P^{(1)}_{X^K} \| P^{(0)}_{X^K}) < \infty, D(P^{(0)}_{X^K} \| P^{(1)}_{X^K}) < \infty$, to avoid the trivial irregularities. Due to the type-based restriction, we further assume that $P^{(0)}_{X_k} \neq P^{(1)}_{X_k}$, $k = 1, \cdots, K$. Otherwise, the transmitted message as a function of the empirical distribution would be uninformative for distinguishing the hypotheses. In the following, we denote $\hat{P}_{X_k}$ as the empirical distributions of X_k, defined as:

$$\hat{P}_{X_k}(x_k) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{x_k = x_k^{(i)}\right\}. \tag{2}$$

2.1. Type-Based Hypothesis Testing over Noiseless Channels

As shown in Figure 1, node k ($k = 1, \cdots, K$) can encode the observed data X_k and transmit a scalar signal by function u_k. Due to the computational requirement as introduced in Section 1, we impose a restriction whereby the encoder u_k is explicitly dependent on the empirical distribution $\hat{P}_{X_k}$, i.e., $u_k \colon \mathcal{P}_{\mathcal{X}_k} \mapsto \mathbb{R}$, and $\mathcal{P}_{\mathcal{X}_k}$ denotes the set of probability distributions defined on the alphabet $\mathcal{X}_k$. For the most direct method, we can transmit the emprical distributions by encoding them into the real space, which can lead to computational difficulty for federated learning data. In this paper, we further consider one of the most commonly used approaches in federated learning [15,16] and assume that u_k computes a one-dimensional statistic

$$u_k(\hat{P}_{X_k}) = \frac{1}{n} \sum_{i=1}^{n} f_k(x_k^{(i)}) = \mathbb{E}_{\hat{P}_{X_k}}[f_k(X_k)], \tag{3}$$

where feature function $f_k \colon \mathcal{X}_k \mapsto \mathbb{R}$. Then, the decision center collects statistics $\{u_k(\hat{P}_{X_k})\}_{k=1}^{K}$, and makes a decision $\hat{\mathsf{H}}$ on the true hypothesis. We prove in Section 4 that the further restrictions of computing the empirical means of features are without a loss of generality, where we can make good decisions as we observe the types. Additionally, the error probability is defined as

$$\mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) \triangleq \sum_{i \in \{0,1\}} \mathbb{P}_{\mathsf{H}}(\mathsf{H}_i)\, \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H} | \mathsf{H} = \mathsf{H}_i),$$

where H denotes the true hypothesis, $\mathbb{P}_{\mathsf{H}}(\mathsf{H}_0)$ and $\mathbb{P}_{\mathsf{H}}(\mathsf{H}_1)$ are the prior distributions, and $\mathbb{P}_n(\cdot)$ is the probability measure defined from the data sampling process (1). In particular, we focus on the asymptotic error decaying rate, i.e., the error exponent, defined as

$$E \triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}), \tag{4}$$

where all logarithms are base e unless otherwise specified. The goal is to find the maximal error exponent of (4) and design the feature functions $f_1, \cdots, f_k$ and the detailed decision rule such that this error exponent can be achieved based on the log-likelihood ratio test (LLRT).

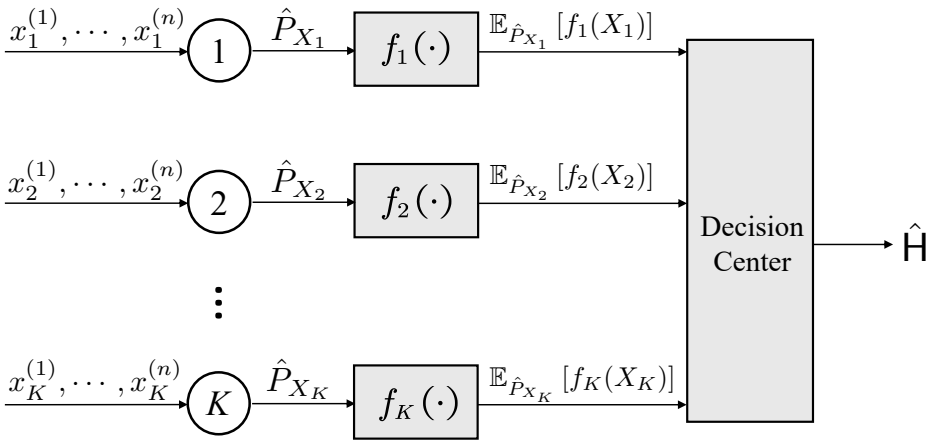

Figure 1. The transmission procedures for the type-based distributed hypothesis testing problem over noiseless channels.

2.2. Type-Based Hypothesis Testing over AWGN Channels

As depicted in Figure 2, we employ the identical hypothesis testing formulation as presented in (1). In this context, it is assumed that nodes 1 through K encode and transmit a length-m sequence using functions $g_1, \cdots, g_K$, which operate based on their respective observations through additive white Gaussian noise (AWGN) channels to the central decision center. To accommodate the computational constraints, we restrict that the encoder g_k ($k = 1, \cdots, K$) is a function of the empirical distribution $\hat{P}_{X_k}$, i.e.,

$$g_k \colon \mathcal{P}_{\mathcal{X}_k} \mapsto \mathbb{R}^m, \quad k = 1, \cdots, K. \tag{5}$$

Moreover, the averaged power constraints of the AWGN channels are:

$$\frac{1}{m}\mathbb{E}\left[\left\|g_k(\hat{P}_{X_k})\right\|^2\right] \leq p_k, \quad k = 1, \cdots, K, \tag{6}$$

where the expectations are taken over the data sampling process defined in (1). Then, the decision center makes a decision $\hat{\mathsf{H}}$ based on the received signals $g_1(\hat{P}_{X_1}) + \boldsymbol{Z}_1, \cdots, g_K(\hat{P}_{X_K}) + \boldsymbol{Z}_K$, where the noises are drawn from

$$\boldsymbol{Z}_k \sim \mathcal{N}\left(\mathbf{0}, \sigma_k^2 \mathbf{I}_m\right), \quad k = 1, \cdots, K, \tag{7}$$

and $\mathbf{I}_m$ denotes the $m \times m$ identity matrix.

Additionally, we make the following assumption to make the errors arising from the AWGN channels and the decision process comparable, so that the trade-off between them can be described. In detail, we assume that the sequence length m also increases with n, and there exists a positive constant μ such that

$$\lim_{n\to\infty} \frac{n}{m(n)} = \mu. \tag{8}$$

Our goal is to design the optimal encoders $g_1, \cdots, g_K$, subjected to the constraints (5) and (6), as well as the decision rule $\hat{\mathsf{H}}$, where we have assumed $\mathbb{P}_{\mathsf{H}}(\mathsf{H}_0) = \mathbb{P}_{\mathsf{H}}(\mathsf{H}_1) = \frac{1}{2}$ for explicit mathematical expression, such that the error exponent as defined in (4) is maximized.

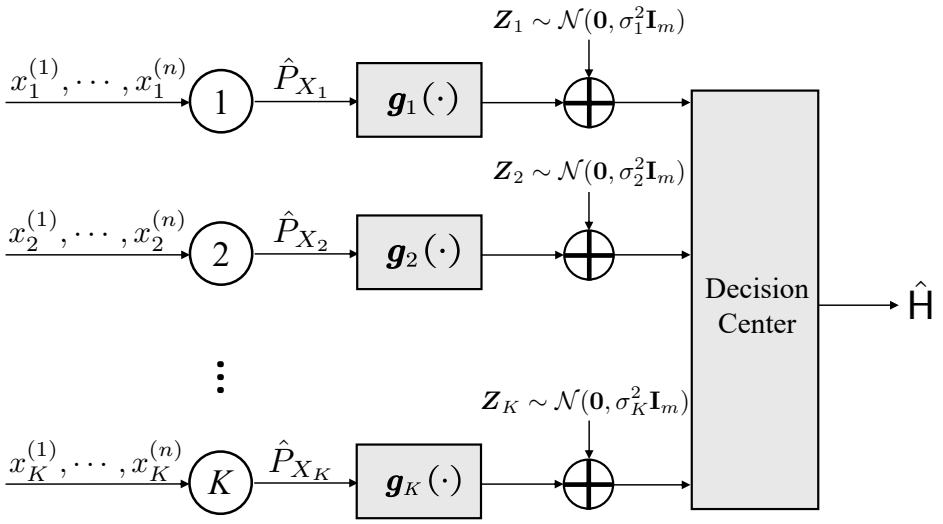

Figure 2. The transmission procedures for the type-based distributed hypothesis testing problem over AWGN channels.

3. Related Works

Distributed hypothesis testing problems, also known as multiterminal hypothesis testing [1,3,14] or decentralized detection [17,18], have been extensively explored in the literature. In scenarios where each node can observe a single observation and send an encoded message to the central machine, the authors of [17] demonstrated that determining the optimal coding scheme is NP-hard, while [18,19] provided characterizations for the minimum decoding error rate and the optimal coding scheme for conditionally independent nodes.

Furthermore, in situations where each node can observe n samples and transmit an encoded message to the decision center, [3,5,14,20] investigated the optimal decoding error exponents for the case of $K = 2$ nodes, with [21] generalizing the results to $K > 2$ nodes. Additionally, the author of [5] studied the Neyman–Pearson-like test, which further constrained the encoded messages to being an empirical functional mean, and provided optimal functions for the scenario with $K = 2$ nodes. The outcome presented in Section 4 can be perceived as a generalization of such setups to the case with $K > 2$ nodes.

On the other hand, DHT over noisy channels represents a novel and highly significant sub-problem within the broader context. While current research has primarily focused on transmission over discrete memoryless channels, certain aspects of this sub-problem have been investigated. For instance, some studies have explored scenarios involving side information [22] and cases that counteract independence assumptions [23]. Additionally, optimal Type-II error considerations have been examined [24], along with investigations into the optimal pairs of Type-I and Type-II errors [25].

Diverging from the existing literature, the present paper delves into the DHT problem in the context of widely considered AWGN channels while also addressing the implications of computational demands. This novel approach fills a critical research gap and extends the understanding of DHT to a broader set of channel conditions, thus contributing to the advancement of the field.

4. Type-Based Hypothesis Testing over Noiseless Channels

In this section, we present the optimal error exponent along with the corresponding decision rule for the type-based hypothesis testing over noiseless channels. We commence by introducing the optimal error exponent under the condition that the decision center has access to the empirical distributions from different nodes.

Definition 1. *The quantities $D_i^*(R_{X_1}, \cdots, R_{X_K})$, for $i = 0, 1$, are defined as*

$$D_i^*(R_{X_1}, \cdots, R_{X_K}) \triangleq \min_{Q_{X^K} \in \mathcal{S}} D(Q_{X^K} \| P_{X^K}^{(i)}), \tag{9}$$

where

$$\mathcal{S} \triangleq \big\{ Q_{X^K} \colon [Q_{X^K}]_{X_k} = R_{X_k},\ k = 1, \cdots, K \big\},$$

which represents the set of all distributions with given marginals $R_{X_1}, \cdots, R_{X_K}$.

The following result provides the operational meaning of (9), which can be proved by Sanov's theorem [12].

Lemma 1. *When H_i is the true hypothesis, the probability that nodes $1, \cdots, K$ observe the empirical distributions $\hat{P}_{X_1}, \cdots, \hat{P}_{X_K}$, respectively, is given by*

$$\mathbb{P}_n(\hat{P}_{X_1}, \cdots, \hat{P}_{X_K} | \mathsf{H} = \mathsf{H}_i) \doteq \exp\big(-n D_i^*(\hat{P}_{X_1}, \cdots, \hat{P}_{X_K})\big),\ i = 0, 1,$$

where $\doteq$ is the conventional dot-equal notation, i.e., we denote $f_n \doteq g_n$ when $\lim_{n\to\infty} \frac{1}{n} \log f_n = \lim_{n\to\infty} \frac{1}{n} \log g_n$. In addition, by applying the log-likelihood ratio test to detect the true hypothesis, the optimal decision error exponent based on the empirical distributions is

$$E^* \triangleq \min_{R_{X_1}, \cdots, R_{X_K}} \max_{i \in \{0,1\}} D_i^*(R_{X_1}, \cdots, R_{X_K}). \tag{10}$$

Note that the type-based hypothesis testing problem assumes that the signal from each node is a function of the empirical distribution. Hence, the optimal error exponent in (4) will not exceed E^*. In the following, we prove that error exponent E^* can be achieved and provide the corresponding decision rule.

4.1. Optimal Feature

First, we introduce the following definitions of exponential and linear families, which will be useful for delineating our results.

Definition 2 (Exponential family)**.** *Given distribution $P_Z(z)$, and a function $T \colon \mathcal{Z} \to \mathbb{R}$, we define the distribution $\tilde{P}_Z^{(\lambda)}(\,\cdot\,; T, P_Z)$ as*

$$\tilde{P}_Z^{(\lambda)}(z; T, P_Z) \triangleq P_Z(z) \exp(\lambda T(z) - \alpha(\lambda)), \quad \text{for all } z \in \mathcal{Z}, \tag{11}$$

with $\alpha(\lambda) \triangleq \log \sum_{z' \in \mathcal{Z}} P_Z(z') \exp(\lambda T(z'))$. In addition, we use

$$\mathcal{E}_{\mathcal{Z}}(T, P_Z) \triangleq \Big\{ \tilde{P}_Z^{(\lambda)}(\,\cdot\,; T, P_Z) \colon \lambda \in \mathbb{R} \Big\} \tag{12}$$

to denote the exponential family passing through P_Z with T being the natural statistic.

Definition 3 (Linear family)**.** *Given a function $h \colon \mathcal{Z} \to \mathbb{R}$, we define the linear family $\mathcal{L}_{\mathcal{Z}}(h)$ as*

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

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

$$\mathcal{S}_{\mathcal{Z}}^{(0)}(h) \triangleq \Big\{Q_Z \in \mathcal{P}^{\mathcal{Z}} \colon \mathbb{E}_{Q_Z}[h(Z)] \leq 0\Big\},$$
$$\mathcal{S}_{\mathcal{Z}}^{(1)}(h) \triangleq \Big\{Q_Z \in \mathcal{P}^{\mathcal{Z}} \colon \mathbb{E}_{Q_Z}[h(Z)] \geq 0\Big\}.$$

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

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

We 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 2. *For E^* as defined in (10), we have $\mathcal{D}(t) = \varnothing$ for all $t \in [0, E^*]$ and $\mathcal{D}(t) \neq \varnothing$ for all $t > E^*$. Additionally, a unique $(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) \in \mathcal{P}_{\mathcal{X}_1} \times \cdots \times \mathcal{P}_{\mathcal{X}_K}$ exists such that*

$$D_0^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = D_1^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = E^*. \tag{14}$$

Proof. See Appendix A. □

Based on Lemma 2, it follows from the separating hyperplane theorem (see, e.g., Section 2.5.1 of [26]) that functions $(f_1^*, \ldots, f_K^*)$, where $f_k^* \colon \mathcal{X}_k \to \mathbb{R}$, $k = 1, \cdots, K$ exist, such that for all $(R_{X_1}, \ldots, R_{X_K}) \in \mathcal{D}_0(E^*)$,

$$\sum_{i=1}^{K} \sum_{x_i \in \mathcal{X}_i} R_{X_i}(x_i) f_i^*(x_i) = \sum_{i=1}^{K} \mathbb{E}_{R_{X_i}}[f_i^*(X_i)] \leq 0, \tag{15}$$

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

$$\sum_{i=1}^{K} \mathbb{E}_{R_{X_i}}[f_i^*(X_i)] \geq 0. \tag{16}$$

Furthermore, we denote

$$h^*(x^K) \triangleq \sum_{i=1}^{K} f_i^*(x_i), \tag{17}$$

and then we have the following proposition. Given $P_Z \in \mathcal{P}_{\mathcal{Z}}$ and $\mathcal{S} \subset \mathcal{P}_{\mathcal{Z}}$, we adopt the notation [27,28] $D(\mathcal{S} \| P_Z) \triangleq \inf_{Q_Z \in \mathcal{S}} D(Q_Z \| P_Z)$, where $\mathcal{P}_{\mathcal{Z}}$ denotes the set of all distributions supported on $\mathcal{Z}$.

Proposition 1. *The optimal exponent E^* as defined in (10) satisfies*

$$E^* = D\big(\mathcal{S}_{\mathcal{X}}^{(0)}(h^*) \big\| P_{X^K}^{(1)}\big) = D\big(\mathcal{S}_{\mathcal{X}}^{(1)}(h^*) \big\| P_{X^K}^{(0)}\big). \tag{18}$$

Proof. See Appendix B. □

Consequently, we establish the optimality of E^* and provide the corresponding decision rule.

Theorem 1. *Let $f_1^*, \ldots, f_K^*$ denote the features as defined in (15) and (16). The optimal error exponent of (4) is given by*

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) = E^*, \tag{19}$$

where E^* is defined in (10). In addition, the corresponding decision rule $\hat{\mathsf{H}}$ is

$$\sum_{k=1}^{K} \mathbb{E}_{\hat{P}_{X_k}}[f_k^*(X_k)] \underset{\hat{\mathsf{H}}=\mathsf{H}_0}{\overset{\hat{\mathsf{H}}=\mathsf{H}_1}{\gtrless}} 0. \tag{20}$$

Proof. See Appendix C. □

4.2. General Geometric Structure

The geometry associated with Proposition 1 and Theorem 1 is depicted in Figure 3. In this figure, each point represents a distribution in $\mathcal{P}_{\mathcal{X}}$, and the decision boundary (20) corresponds to the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$ defined as in (13). In addition, from Corollary 3.1 of [27], $\lambda_0, \lambda_1 \in \mathbb{R}$ exist such that

$$Q_{X^K}^{(i)} \triangleq \tilde{P}_{X^K}^{(\lambda_i)}(\,\cdot\,; h^*, P_{X^K}^{(i)}), \quad i = 0, 1, \tag{21}$$

satisfy

$$D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*) \| P_{X^K}^{(i)}\big) = D\big(Q_{X^K}^{(i)} \| P_{X^K}^{(i)}\big), \tag{22}$$

where $\tilde{P}_{X^K}^{(\lambda_i)}(\,\cdot\,; h^*, P_{X^K}^{(i)}), i = 0, 1$ are as defined in (11). In this context, $Q_{X^K}^{(0)}$ and $Q_{X^K}^{(1)}$ in (21) are the I-projections [27] of $P_{X^K}^{(0)}$ and $P_{X^K}^{(1)}$ 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. Additionally, all the points in $\mathcal{D}_0(E^*)$ and $\mathcal{D}_1(E^*)$ are divided by the the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$.

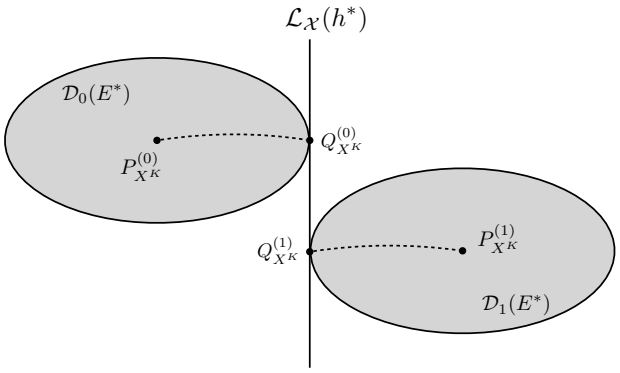

Figure 3. The geometric structure in distributed hypothesis testing, with $Q_{X^K}^{(i)}$ denoting the I-projection of $P_{X^K}^{(i)}$ onto the linear family $\mathcal{L}_{\mathcal{X}}(h^*)$, $i = 0, 1$, and $\mathcal{L}_{\mathcal{X}}(h^*)$ can devide $\mathcal{D}_0(E^*)$ and $\mathcal{D}_1(E^*)$ in different half spaces.

4.3. Local Information Geometric Analysis

Although an explicit information geometry has been shown, we apply the local information geometric framework [13] to provide fundamental insights into this problem. Some useful notations and definitions in local information geometry are introduced as follows.

Definition 4 (ϵ-neighborhood). *Given a finite alphabet $\mathcal{Z}$, and letting R_Z be a distribution supported on $\mathcal{Z}$ with all entries being positive, its ϵ-neighborhood $\mathcal{N}_\epsilon^{\mathcal{Z}}(R_Z)$ is defined as*

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

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

$$\phi(z) \triangleq \frac{P_Z(z) - R_Z(z)}{\sqrt{R_Z(z)}}, \; f(z) \triangleq \frac{\phi(z)}{\sqrt{R_Z(z)}}, \quad \forall\, z \in \mathcal{Z}, \tag{23}$$

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.

Based on Definition 4, we introduce the local assumption that

$$P_{X^k}^{(i)} \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_{X^k}), \quad \text{for } i = 0, 1, \tag{24}$$

We use $\psi^{(i)} \leftrightarrow P_{X^K}^{(i)}, i = 0, 1$ to represent the corresponding information vectors [cf. (23)]. For each $k = 1, \ldots, K$, and given feature $f_k\colon \mathcal{X}_k \to \mathbb{R}$, we define the corresponding information vector $\phi_k \in \mathbb{R}^{|\mathcal{X}_k|}$, where $P_{X_k} \triangleq [P_{X^K}]_{X_k}$ is used as the reference distribution. Note that for $i = 0, 1$, the correspondence $B_k^{\mathrm{T}}\psi^{(i)} \leftrightarrow P_{X_k}^{(i)}$ exists, where $P_{X_k}^{(i)} \triangleq [P_{X^K}^{(i)}]_{X_k}$ represents the corresponding marginal distributions. Specifically, B_k is an $|\mathcal{X}| \times |\mathcal{X}_k|$ dimensional matrix with entries [29]

$$B_k(x^K, \hat{x}_k) \triangleq \sqrt{\frac{P_{X^K}(x^K)}{P_{X_k}(\hat{x}_k)}}\delta_{x_k \hat{x}_k}, \tag{25}$$

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

Moreover, the feature f_k defined on $\mathcal{X}_k$, when considered as a mapping from $\mathcal{X}$ to $\mathbb{R}$, corresponds to the information vector $B_k\phi_k$ in $\mathbb{R}^{|\mathcal{X}|}$. Leveraging this correspondence, we can further establish the information vector for $h(x^K) = \sum_{k=1}^K f_k(x_k)$ as

$$\sum_{i=1}^{K} B_i\phi_i = B_0\phi_0 \in \mathbb{R}^{|\mathcal{X}|}, \tag{26}$$

where we have defined

$$B_0 \triangleq \begin{bmatrix} B_1 & \cdots & B_K \end{bmatrix} \quad \text{and} \quad \phi_0 \triangleq \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_K \end{bmatrix}, \tag{27}$$

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

Additionally, given a matrix $A \in \mathbb{R}^{m_1 \times m_2}$, we use $A^\dagger$ to denote its Moore–Penrose inverse [30], and we define the associated column space $\mathcal{R}(A) \triangleq \{Ax\colon x \in \mathbb{R}^{m_2}\}$ and projection matrix $\Pi_A \triangleq AA^\dagger$. Then, we can establish the local counterpart of E^* in Theorem 1 as follows.

Theorem 2. *Under the local assumption (24), let $\psi^{(i)} \leftrightarrow P_{X^K}^{(i)}$, $i = 0, 1$ denote the corresponding information vectors. Then, for h^* as defined in (17), we have the correspondence $h^* \leftrightarrow B_0\phi_0^*$, where*

$$\phi_0^* \triangleq B_0^\dagger(\psi^{(1)} - \psi^{(0)}), \tag{28}$$

and where B_0 is defined in (27). In addition, the optimal exponent E^* in (10) can be expressed as

$$E^* = \frac{1}{8}\left\|B_0\phi_0^*\right\|^2 + o(\epsilon^2). \tag{29}$$

Proof. See Appendix D. □

Note that from Theorem 2, we have

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

where Π_{B_0} is the projection matrix associated with the subspace $\mathcal{R}(B_0)$. The optimal feature $B_0\phi_0^*$ in (26) corresponds to the projection of the sufficient statistic $f_{\text{LLR}} \leftrightarrow (\psi^{(1)} - \psi^{(0)})$ onto the function space that encompasses all possible h's satisfying the form $h(x^K) = \sum_{k=1}^K f_k(x_k)$. In other words, $B_0\phi_0^*$ represents the best approximation of f_{LLR} within the function space of interest, which leads to the optimal decision error exponent E^* as shown in (29).

Moreover, from (26), this optimal feature can be decomposed to K components in subspaces $\mathcal{R}(B_k)$, for $k = 1, \ldots, K$,

$$B_0\phi_0^* = \sum_{k=1}^{K} B_k\phi_k^*, \tag{30}$$

where ϕ_0^* is stacked by $\phi_k^* \in \mathbb{R}^{|\mathcal{X}_k|}$, $k = 1, \ldots, K$, as in (27). This decomposition structure can be depicted as Figure 4 for the case $K = 2$.

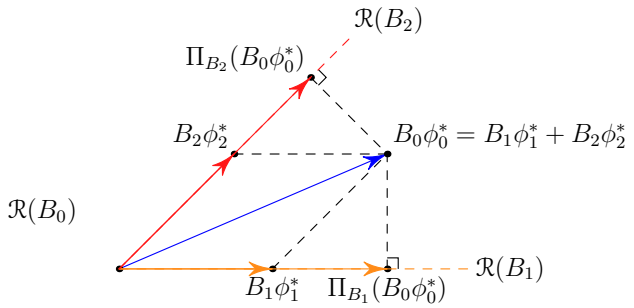

Figure 4. The information decomposition structure in distributed hypothesis testing with $K = 2$ nodes, compared with the orthogonal decompositions on the subspace $\mathcal{R}(B_k)$ for each node $k = 1, 2$.

Remark 1. *The vectors $B_i\phi_k^*$ are not simply the orthogonal projections of $B_0\phi_0^*$ onto the subspaces $\mathcal{R}(B_k)$ since these subspaces, for $k = 1, \ldots, K$, are not mutually orthogonal. Therefore, the decomposition of $B_0\phi_0^*$ will depend on the Gram matrix [30] of the subspaces $\mathcal{R}(B_k)$, as illustrated in Figure 4. Furthermore, it is noteworthy 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 solely with the observations of X_k [12]. When the subspaces $\mathcal{R}(B_k)$ are orthogonal to each other, the optimal inference approach is straightforward, involving the extraction of the optimal information from each node by orthogonal projection. However, when the subspaces $\mathcal{R}(B_k)$ are not orthogonal, different nodes may share various forms of common information. Our result fundamentally demonstrates how to handle this shared information and extract the optimal features through the decomposition of the information vector over non-orthogonal subspaces. This insight provides a novel approach to address the challenges posed by the non-orthogonal subspaces and reveals how to extract the most informative features effectively, ultimately leading to improved performance in the distributed hypothesis testing problem.*

5. Type-Based Hypothesis Testing over AWGN Channels

This section presents the optimal error exponent of the type-based hypothesis testing problem over AWGN channels, along with the corresponding coding strategy. To begin, we introduce several notations that will help in the presentation of the results.

Definition 5. *Let $[K] \triangleq \{1, 2, \cdots, K\}$, and for subset $\omega \subseteq [K]$, $i = 0, 1$, we define*

$$D_i^{\omega}(\{R_{X_k}\}_{k\in\omega}) \triangleq \min_{Q_{X^K} \in \mathcal{S}_\omega} D(Q_{X^K} \| P_{X^K}^{(i)}), \tag{31}$$

where

$$\mathcal{S}_\omega \triangleq \{Q_{X^K} \colon [Q_{X^K}]_{X_k} = R_{X_k},\ k \in \omega\}.$$

It would be easy to find that $D_i^{[K]}(\cdot) = D_i^*(\cdot)$, and $D_i^*(\cdot)$ is as defined in (9). Moreover, we define the following error exponent with respect to $\omega \subseteq [K]$.

$$E_\omega \triangleq \min_{\{R_{X_k}\}_{k\in\omega}, \{\theta_k\}_{k\in[K]\setminus\omega}} \max\left\{ D_0^\omega(\{R_{X_k}\}_{k\in\omega}) + \sum_{k\in[K]\setminus\omega} \frac{(\theta_k - \sqrt{p_k})^2}{2\mu\sigma_k^2}, \right.$$
$$\left. D_1^\omega(\{R_{X_k}\}_{k\in\omega}) + \sum_{k\in[K]\setminus\omega} \frac{(\theta_k + \sqrt{p_k})^2}{2\mu\sigma_k^2} \right\}, \tag{32}$$

where we have used $A \setminus B$ to represent the relative complement of set B in set A, and where μ is as defined in (8). We can also find $E_{[K]} = E^*$ and E^* is as defined in (10). Finally, we define the quantity $E^\odot$, which will be shown as the optimal error exponent

$$E^\odot \triangleq \min_{\omega\in\Im([K])} E_\omega, \tag{33}$$

where $\Im([K])$ denotes the power set of $[K]$.

Theorem 3. *The optimal error exponent of* (4) *is given by*

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) = E^\odot. \tag{34}$$

In the following, we prove Theorem 3 by both the achievability and converse result.

5.1. The Coding Strategy for Distributed Nodes

First, we define the different regimes of empirical distributions, for each $k = 1, \cdots, K$ and for some $\gamma \in (0, 1)$. Basically, the specific choice of γ does not effect the achievable error exponent as long as $\gamma \in (0, 1)$. It helps conduct the decode-and-forward and amplify-and-forward coding strategies as introduced in Section 1.

Decode-and-forward regime:

$$\mathcal{M}_k^{(0)} \triangleq \left\{ R_{X_k} \colon D(R_{X_k} \| P_{X_k}^{(0)}) < n^{-\gamma} \right\},$$
$$\mathcal{M}_k^{(1)} \triangleq \left\{ R_{X_k} \colon D(R_{X_k} \| P_{X_k}^{(1)}) < n^{-\gamma} \right\}.$$

Amplify-and-forward regime:

$$\mathcal{M}_k^{\mathsf{c}} \triangleq \left\{ R_{X_k} \colon \min\left\{ D(R_{X_k} \| P_{X_k}^{(0)}), D(R_{X_k} \| P_{X_k}^{(1)}) \right\} \geq n^{-\gamma} \right\}. \tag{35}$$

Note that for each $k = 1, \cdots, K$, the probability of the empirical distribution $\hat{P}_{X_k}$ in $\mathcal{M}_k^{\mathsf{c}}$ is $\exp\left(-n^{1-\gamma}\right)$. Consequently, in the amplify-and-forward regime, we can transmit such empirical distributions with exponentially large power by Pulse Amplitude Modulation (PAM) while still satisfying the power constraint. Specifically, let $\mathcal{P}_{\mathcal{X}_k}^{(n)}$ be the set of all possible empirical distributions of X_k with n samples, and denote $\eta_k \triangleq |\mathcal{P}_{\mathcal{X}_k}^{(n)} \cap \mathcal{M}_k^{\mathsf{c}}|$. We define the bijective function $\xi_k \colon \mathcal{P}_{\mathcal{X}_k}^{(n)} \cap \mathcal{M}_k^{\mathsf{c}} \mapsto \{1, \ldots, \eta_k\}$ as the indices of empirical distributions. Then, according to the observed empirical distribution, the encoder of node k ($k = 1, \cdots, K$) is designed to transmit the signal

$$\boldsymbol{Q}_k(\hat{P}_{X_k}) \triangleq \xi_k(\hat{P}_X) \cdot \exp\left(n^{\frac{1-\gamma}{2}}\right). \tag{36}$$

Furthermore, if the empirical distributions are in the decode-and-forward regimes, we initially detect the true hypothesis and then transmit the bit using Binary Phase Shift Keying (BPSK) with the appropriate power. By employing these strategies, the achievability result can be obtained through repeated transmissions from all the distributed nodes. In other words, the resulting encoder for node k is defined as follows:

$$\boldsymbol{g}_k^* = [g_k^*, \cdots, g_k^*], \quad k = 1, \cdots, K, \tag{37}$$

where

$$g_k^*(\hat{P}_{X_k}) \triangleq \begin{cases} \sqrt{p_k - \delta(n,\gamma)}, & \text{if } \hat{P}_{X_k} \in \mathcal{M}_k^{(0)} \\ -\sqrt{p_k - \delta(n,\gamma)}, & \text{if } \hat{P}_{X_k} \in \mathcal{M}_k^{(1)} \\ \boldsymbol{Q}_k(\hat{P}_{X_k}), & \text{if } \hat{P}_{X_k} \in \mathcal{M}_k^{\mathsf{c}} \end{cases}, \tag{38}$$

and where

$$\delta(n,\gamma) \triangleq \max_{k\in[K]} \frac{\mathbb{P}_n(\hat{P}_{X_k} \in \mathcal{M}_k^{\mathsf{c}})}{\mathbb{P}_n(\hat{P}_{X_k} \notin \mathcal{M}_k^{\mathsf{c}})} \cdot (n+1)^{2|\mathcal{X}_k|} \cdot \exp\left(2n^{\frac{1-\gamma}{2}}\right). \tag{39}$$

Proposition 2. *The encoders as defined in* (38) *satisfy the power constraint* (6)*, and*

$$\lim_{n\to\infty} \delta(n,\gamma) = 0. \tag{40}$$

Proof. See Appendix E. □

5.2. Decision Rule and Achievable Error Exponent

After the decision center receives the output signals $\boldsymbol{g}_1^*(\hat{P}_{X_1}) + \boldsymbol{Z}_1, \cdots, \boldsymbol{g}_K^*(\hat{P}_{X_K}) + \boldsymbol{Z}_K$, we then compute

$$\theta_k \triangleq \frac{1}{m} \sum_{i=1}^{m} [\boldsymbol{g}_k^*(\hat{P}_{X_k}) + \boldsymbol{Z}_k]_i, \quad k = 1, \cdots, K,$$

where $[\cdot]_i$ denotes the i-th entry of a given vector. Then, we conduct the log-likelihood ratio test (LLRT) to detect the true hypothesis:

$$\log \frac{\mathbb{P}_n(\theta_1, \cdots, \theta_K | \mathsf{H} = \mathsf{H}_0)}{\mathbb{P}_n(\theta_1, \cdots, \theta_K | \mathsf{H} = \mathsf{H}_1)} \underset{\hat{\mathsf{H}}=\mathsf{H}_1}{\overset{\hat{\mathsf{H}}=\mathsf{H}_0}{\gtrless}} 0. \tag{41}$$

Note that exponentially large power is allocated for the empirical distributions in the amplify-and-forward regime (cf. (35), (36)); the decision center can correctly detect the coding regime of the nodes with super-exponentially high probability, i.e., for $k = 1, \cdots, K$,

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_n\left(\hat{P}_{X_k} \in \mathcal{M}_k^{\mathsf{c}} \middle| \theta_k \le \exp\left(n^{\frac{1-\gamma}{4}}\right)\right) &= \infty, \\ \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_n\left(\hat{P}_{X_k} \notin \mathcal{M}_k^{\mathsf{c}} \middle| \theta_k > \exp\left(n^{\frac{1-\gamma}{4}}\right)\right) &= \infty. \end{aligned} \tag{42}$$

Therefore, we can assume that the decision center knows the coding regime of the nodes and define the following regime of the received signals with respect to subset $\omega \subseteq [K]$.

$$\Theta_\omega \triangleq \left\{(\theta_1, \cdots, \theta_K) \colon \theta_k > \exp\left(n^{\frac{1-\gamma}{4}}\right), \forall k \in \omega, \text{ and } \theta_{k'} \le \exp\left(n^{\frac{1-\gamma}{4}}\right), \forall k' \in [K] \setminus \omega\right\},$$

for all $\omega \in \Im([K])$. When the received signals $(\theta_1, \cdots, \theta_K) \in \Theta_\omega$, the decision center can recover the empirical distributions $\hat{P}_{X_k}$ $(k \in \omega)$ from the received signals θ_k by the decoder:

$$Q_k^{-1}(\theta_k) \triangleq \xi_k^{-1}\left(\left\lfloor \theta_k / \exp\left(n^{\frac{1-\gamma}{2}}\right) + 0.5 \right\rfloor\right), \tag{43}$$

where $\lfloor \cdot \rfloor$ denotes the floor function [31]. The following result shows that decoding error of (43) can be neglected.

Proposition 3. *For all* $\hat{P}_{X_k} \in \mathcal{P}_{\mathcal{X}_k}^{(n)} \cup \mathcal{M}_k^{\mathsf{c}}$, $k = 1, \cdots, K$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}(Q_k^{-1}(\theta_k) \neq \hat{P}_{X_k}) = \infty. \tag{44}$$

Proof. See Appendix F. □

In the following, we denote $p_k' \triangleq p_k - \delta$, for $k = 1, \cdots, K$ and discuss the decision error exponent when the received signals are in Θ_ω. For $k \in \omega$, the empirical distribution $\hat{P}_{X_k}$ can be recovered by (43), and for $k \in [K] \setminus \omega$, node k detects the hypothesis according to the observed empirical distribution and transmits the detected bit by BPSK (cf. (38)) through the AWGN channel. Then, the decision center detects the true hypothesis from the received signals by LLRT (41), which can be reduced to

$$\tilde{E}_0^\omega(\theta_1, \cdots, \theta_K) \underset{\hat{\mathsf{H}}=\mathsf{H}_0}{\overset{\hat{\mathsf{H}}=\mathsf{H}_1}{\gtrless}} \tilde{E}_1^\omega(\theta_1, \cdots, \theta_K), \tag{45}$$

where for $i = 0, 1$,

$$\begin{aligned}&\tilde{E}_i^\omega(\theta_1, \cdots, \theta_K)\\ &\triangleq \min_{\bar{\omega} \in \Im([K]\setminus\omega)} D_i^*(\bar{P}_{X_1}, \cdots, \bar{P}_{X_K}) + \sum_{k\in\bar{\omega}} \frac{(\theta_k - \sqrt{p_k'})^2}{2\mu\sigma_k^2} + \sum_{k'\in[K]\setminus(\omega\cup\bar{\omega})} \frac{(\theta_{k'} + \sqrt{p_{k'}'})^2}{2\mu\sigma_{k'}^2},\end{aligned}$$

where $\Im([K] \setminus \omega)$ denotes the power set of $[K] \setminus \omega$, and where for $k = 1, \cdots, K$,

$$\bar{P}_{X_k} \triangleq \begin{cases} Q_k^{-1}(\theta_k), & \text{if } k \in \omega \\ P_{X_k}^{(0)}, & \text{if } k \in \bar{\omega} \\ P_{X_k}^{(1)}, & \text{if } k \in [K] \setminus (\omega \cup \bar{\omega}) \end{cases}. \tag{46}$$

Consequently, the decision error exponent is characterized by the following proposition.

Proposition 4. *For any* $\epsilon > 0$ *and* $\omega \in \Im([K])$, *the decision error exponent by the decision rule* (45) *satisfies*

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_n\left(\hat{\mathsf{H}} \neq \mathsf{H}, (\theta_1, \cdots, \theta_K) \in \Theta_\omega\right) \geq E^{\odot} - \epsilon, \tag{47}$$

where $E^{\odot}$ *is as defined in* (33).

Proof. See Appendix G. □

Noticing that the overall decision error probability is

$$\mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) = \sum_{\omega\in\Im([K])} \mathbb{P}(\hat{\mathsf{H}} \neq \mathsf{H}, (\theta_1, \cdots, \theta_K) \in \Theta_\omega),$$

the following proposition establishes the achievable error exponent by the coding strategy (38).

Proposition 5. *By using the encoders* $\boldsymbol{g}_1^*, \cdots, \boldsymbol{g}_K^*$ *as defined in* (38)*, and the decision rules* $\hat{\mathsf{H}}$ *from* (41)*, the achievable error exponent is given by* $E^{\odot}$*, i.e.,*

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) \geq E^{\odot}, \tag{48}$$

where $E^{\odot}$ *is as defined in* (33).

5.3. The Converse Result

In this section, we show that $E^{\odot}$ is indeed an upper bound of (4), which establishes Theorem 3. Our main technique is to apply a genie-aided approach, which provides different kinds of additional information to both nodes and computes the corresponding error exponents under additional information. As depicted in Figure 5, given index set $\omega \in \Im([K])$, suppose that for all $k \in \omega$, node k can know and cancel the channel noise in advance; then, the channel is noiseless, and the decision center can perfectly receive the empirical distribution $\hat{P}_{X_k}$. On the other hand, suppose that for all $k' \in [K] \setminus \omega$, we can leverage the true hypothesis H to node k'; then, with such additional information, we can establish the following upper bound of (4) (cf. (33)).

Proposition 6. *Given index set* $\omega \in \Im([K])$*, suppose that for all* $k \in \omega$*, the decision center can obtain* $\hat{P}_{X_k}$ *perfectly. Additionally, for all* $k' \in [K] \setminus \omega$*, node* k' *can obtain the true hypothesis* H*. The resulting optimal decision error exponent is*

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) = E_{\omega}, \tag{49}$$

where E_{ω} *is as defined in* (32).

Proof. See Appendix H. □

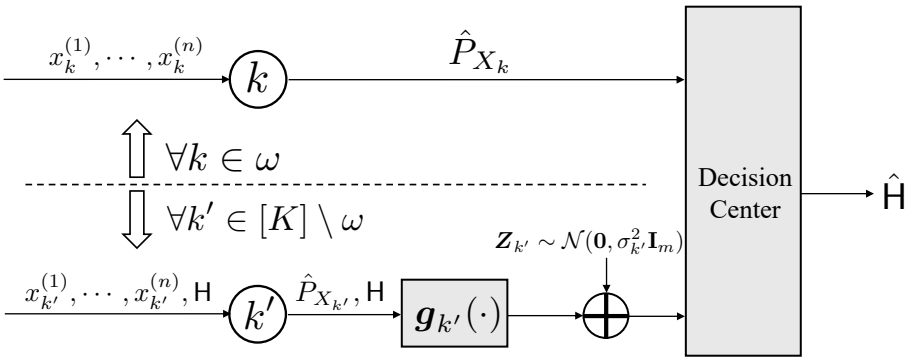

Figure 5. A geometric explanation of the genie-aided approach, which can lead to E_{ω} as the upper bound of the error exponent in (4).

Notice that Proposition 6 is verified for all $\omega \in \Im([K])$, and we cannot obtain a better performance than Proposition 6 for the DHT over AWGN channels without the additional information. We then conclude the following error exponent upper bound.

Proposition 7. *For all possible encodes $\boldsymbol{g}_1, \cdots, \boldsymbol{g}_K$ under the power constraint (6), the corresponding error exponent with respect to the LLRT decision rule satisfies*

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H}) \leq E^{\odot}, \tag{50}$$

where $E^{\odot}$ is as defined in (33).

Finally, by combining Propositions 5 and 7, Theorem 3 is proved.

Remark 2 (Local-geometric interpretation)**.** *Note that the expression of the optimal error exponent $E^{\odot}$ as defined in (33) is quite intricate, which could limit our understanding. To simplify the analysis, we introduce the local geometry assumption as given in (24). In Appendix I, we demonstrate that the error exponent corresponds to a more manageable expression*

$$E^{\odot} = \min_{\omega\in\Im([K])} \frac{1}{8}\left\|B_\omega B_\omega^{\dagger}\left(\psi_\omega^{(1)} - \psi_\omega^{(0)}\right)\right\|^2 + \sum_{k\in[K]\setminus\omega} \frac{p_k}{2\mu\sigma_k^2} + o(\epsilon^2), \tag{51}$$

where for $\omega = \{i_1, \cdots, i_{|\omega|}\}$, we have defined

$$B_\omega \triangleq \begin{bmatrix} B_{i_1} & \cdots & B_{i_{|\omega|}} \end{bmatrix}, \tag{52}$$

and $\psi_\omega^{(i)} \leftrightarrow [P_{X^K}^{(i)}]_{X_{i_1}\cdots X_{i_{|\omega|}}}, i = 0, 1$. Given $\omega \in \Im([K])$, the first term in (51) represents the optimal error exponent (cf. (29)) when the decision center can access the empirical distributions $\hat{P}_{X_k}, k \in \omega$. The second term corresponds to the optimal error exponent when node k, $k \in [K] \setminus \omega$ can know the true hypothesis H and transmit the bit using BPSK modulation. The total error exponent is the sum of these two parts, and $E^{\odot}$ aims to determine the minimum sum among all possible splits of the index set $[K]$. In other words, $E^{\odot}$ finds the optimal trade-off between accessing empirical distributions at the decision center and having individual nodes transmit bits with BPSK modulation.

6. Discussion

This paper discusses the DHT problem over two communication models. The first is the noiseless channel, which is mostly considered in current distributed learning and federated learning systems [9,11]. For the noiseless channels, we show that by using one-dimensional statistics from different nodes, it is possible to achieve the same error exponent when the decision center has knowledge of the corresponding empirical distributions. This result is significant as it simplifies the coding process at distributed nodes, allowing them to transmit only the necessary statistics rather than the entire empirical distribution, which provides a practical implementation of the result in [5]. This finding proves the rationality of transmitting statistics as the most widely-used strategy in distributed learning and federated learning [11].

For the AWGN channels, this paper introduces a novel coding strategy, which cleverly combines decode-and-forward and amplify-and-forward techniques. The underlying concept of this coding strategy is based on the observation that the probability of the empirical distribution deviating significantly from the true marginal distribution diminishes exponentially. Consequently, by employing sufficiently large power, we can transmit the empirical distribution almost perfectly to the decision center while satisfying the averaged power constraint. When the prior distributions are not $1/2$, the strategy still work for the optimal error exponent, and the only difference is to adjust the BPSK points for two hypotheses according to the power constaint. The demonstrated optimality of the achieved decision error exponent further indicates that the proposed coding strategy is highly effective and successfully approaches the theoretical limit within the given constraints of the problem.

7. Conclusions

This paper focuses on investigating DHT problems over both noiseless channels and AWGN channels, where the distributed nodes are constrained to encoding the received empirical distributions, driven by practical computational considerations. In the first problem, we demonstrate that utilizing one-dimensional statistics of distributed nodes and simply summing them up as the decision rule can lead to the optimal error exponent. For the second problem, we propose a coding strategy that combines decode-and-forward and amplify-and-forward techniques. We further introduce a genie-aided approach to establish the optimality of the achieved decision error exponent. Overall, our findings offer valuable insights into coding techniques for distributed nodes, and the established strategies can be extended to more general scenarios, broadening the applicability of DHT in diverse settings.

Author Contributions: X.T., X.X. and S.-L.H. contributed to the conceptualization, methodology, and writing of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202 and the Shenzhen Science and Technology Program under Grant KQTD20170810150821146.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DHT	Distributed hypothesis testing
AWGN	Additive white Gaussian noise
BPSK	Binary phase shift keying
LLRT	Log-likelihood ratio test
PAM	Pulse amplitude modulation

Appendix A. Proof of Lemma 2

We have the following facts:

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

and

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

from which we know $\mathcal{D}(\tilde{t}) \neq \varnothing$, where $\tilde{t} \triangleq \min\{D(P_{X^K}^{(0)} \| P_{X^K}^{(1)}), D(P_{X^K}^{(1)} \| P_{X^K}^{(0)})\}$. Moreover, from the facts $D(0) = \varnothing$ and

$$\mathcal{D}(t_1) \subset \mathcal{D}(t_2), \quad \text{for all } 0 \leq t_1 \leq t_2, \tag{A1}$$

we define

$$t_0 \triangleq \sup\{t \geq 0 \colon \mathcal{D}(t) = \varnothing\}. \tag{A2}$$

We also have

$$\mathcal{D}(t) \neq \varnothing \implies \mathcal{D}(t - \epsilon) \neq \varnothing \text{ for some } \epsilon > 0. \tag{A3}$$

Indeed, since $\mathcal{D}(t)$ is non-empty, $(R_{X_1}, \ldots, R_{X_K})$ and $\epsilon > 0$ exist 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.

To sum up, from (A1)–(A3) 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 (14), 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 $t_0 = E^*$ (cf. (10)). Hence, marginal distributions $\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}$ exist such that

$$D_i^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) = E^*, \quad i = 0, 1. \tag{A4}$$

Finally, to illustrate the uniqueness of $(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K})$, suppose that (14) 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 $D_0^*(\cdot)$ and $D_1^*(\cdot)$ that

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

which contradicts (A2).

Appendix B. Proof of Proposition 1

We know that $\mathcal{D}_i(E^*) \subset \mathcal{S}_{\mathcal{X}}^{(i)}(h^*)$, for $i = 0, 1$. This implies that $\mathcal{S}_{\mathcal{X}}^{(i)}(h^*) \subset \mathcal{D}_{1-i}^{\mathsf{c}}(E^*)$, where for $t \geq 0$ 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 2; then, we have

$$(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) \in \mathcal{L}_{\mathcal{X}}(h^*) = \mathcal{S}_{\mathcal{X}}^{(0)}(h^*) \cap \mathcal{S}_{\mathcal{X}}^{(1)}(h^*). \tag{A5}$$

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

$$\begin{aligned} E^* &= D_i^*(\tilde{R}_{X_1}, \ldots, \tilde{R}_{X_K}) \\ &\geq D\big(\mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*) \big\| P_{X^K}^{(i)}\big) \\ &= \min_{(R_{X_1}, \ldots, R_{X_K}) \in \mathcal{S}_{\mathcal{X}}^{(1-i)}(h^*)} D_i^*(R_{X_1}, \ldots, R_{X_K}) \\ &\geq \min_{(R_{X_1}, \ldots, R_{X_K}) \in \mathcal{D}_i^{\mathsf{c}}(E^*)} D_i^*(R_{X_1}, \ldots, R_{X_K}) \\ &\geq E^*, \end{aligned} \tag{A6}$$

which implies (18).

Appendix C. Proof of Theorem 1

On the one hand, note that from the Markov relation

$$\mathsf{H} - (\hat{P}_{X_1}, \ldots, \hat{P}_{X_K}) - (u_1(\hat{P}_{X_1}), \ldots, u_K(\hat{P}_{X_K})),$$

the minimum possible decision error can be obtained when we choose the empirical distributions $\hat{P}_{X_1}, \ldots, \hat{P}_{X_K}$ themselves as the statistics.

One the other hand, from Proposition 1, the error exponents associated with the type I error and the type II error are $D\big(\mathcal{S}^{(1)}_{\mathcal{X}}(h^*)\big\|P^{(0)}_{X^K}\big)$ and $D\big(\mathcal{S}^{(0)}_{\mathcal{X}}(h^*)\big\|P^{(1)}_{X^K}\big)$, respectively. From (18), both exponents are E^*, and thus the error exponent for $\mathbb{P}_n(\hat{\mathsf{H}} \neq \mathsf{H})$ is also E^*.

Appendix D. Proof of Theorem 2

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 Lemma 17 of [13] that the exponent based on the feature $h(x^K) = \sum_{k=1}^K f_k(x_k)$ is

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

where we have defined $\zeta \triangleq B_0\phi_0 \in \mathcal{R}(B_0)$, and where ϕ_0 is as defined in (27).

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

$$\frac{\langle \psi, \zeta \rangle^2}{\|\zeta\|^2} = \frac{(\psi^{\mathrm{T}}\Pi_{B_0}\zeta)^2}{\|\zeta\|^2} = \frac{\langle \Pi_{B_0}\psi, \zeta \rangle^2}{\|\zeta\|^2} \leq \|\Pi_{B_0}\psi\|^2,$$

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

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

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 $\zeta^* \leftrightarrow h^*$, where h^* is the optimal feature as defined in (17). Note that in (21), for each $i = 0, 1$, $Q^{(i)}_{X^K}$ depends only on the product $\lambda_i h^*$; we may assume $\lambda_0 = 1/2$ and simply use λ to denote λ_1. Then, we have

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

which implies the correspondence

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

Similarly, we have

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

Then, it follows from the second-order Taylor series expansion of the K-L divergence that (see, e.g., Lemma 10 of [13])

$$\begin{aligned}
D\big(Q^{(0)}_{X^K}\big\|P^{(0)}_{X^K}\big) &= \frac{1}{8}\|\zeta\|^2 + o(\epsilon^2), \\
D\big(Q^{(1)}_{X^K}\big\|P^{(1)}_{X^K}\big) &= \frac{\lambda^2}{2}\|\zeta\|^2 + o(\epsilon^2).
\end{aligned} \tag{A7}$$

Moreover, note that since (cf. Lemma 9 of [13])

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

we have

$$\begin{aligned} 0 &= \mathbb{E}_{Q_{X^K}^{(1)}}\left[h^*(X^K)\right] - \mathbb{E}_{Q_{X^K}^{(0)}}\left[h^*(X^K)\right] \\ &= \left\langle \psi + \left(\lambda - \frac{1}{2}\right)\zeta, \zeta \right\rangle + o(\epsilon^2) \\ &= c\left\langle \psi + \left(\lambda - \frac{1}{2}\right)c \cdot \Pi_{B_0}\psi, \Pi_{B_0}\psi \right\rangle + o(\epsilon^2) \\ &= c \cdot \left[1 + \left(\lambda - \frac{1}{2}\right)c\right] \cdot \|\Pi_{B_0}\psi\|^2 + o(\epsilon^2). \end{aligned} \tag{A8}$$

As a result, it follows from $D\left(Q_{X^K}^{(0)} \middle\| P_{X^K}^{(0)}\right) = D\left(Q_{X^K}^{(1)} \middle\| P_{X^K}^{(1)}\right)$ and (A8) that $c = 1, \lambda = -\frac{1}{2}$. Then, we obtain

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

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

Finally, the optimal error exponent is

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

Appendix E. Proof of Proposition 2

According to Sanov's theorem, $\mathbb{P}_n(\hat{P}_{X_k} \in \mathcal{M}_k^c) \doteq \exp\left(-n^{1-\gamma}\right)$, and $\mathbb{P}_n(\hat{P}_{X_k} \notin \mathcal{M}_k^c) \doteq 1$. Then, we have

$$\frac{\mathbb{P}_n(\hat{P}_{X_k} \in \mathcal{M}_k^c)}{\mathbb{P}_n(\hat{P}_{X_k} \notin \mathcal{M}_k^c)} \cdot (n+1)^{2|\mathcal{X}_k|} \cdot \exp\left(2n^{\frac{1-\gamma}{2}}\right) \doteq \exp\left(-n^{1-\gamma}\right),$$

which will converge to 0 as $n \to 0$. Additionally, for the power constraint,

$$\begin{aligned} \mathbb{E}[g_k^{*2}(\hat{P}_{X_k})] &\le (p_k - \delta(n,\gamma)) \cdot \mathbb{P}_n(\hat{P}_{X_k} \notin \mathcal{M}_k^c) + \left(|\mathcal{M}_k^c| \cdot \exp\left(n^{\frac{1-\gamma}{2}}\right)\right)^2 \cdot \mathbb{P}(\hat{P}_X \in \mathcal{M}_k^c) \\ &\le p_k - \delta(n,\gamma) \cdot \mathbb{P}_n(\hat{P}_{X_k} \notin \mathcal{M}_k^c) + (n+1)^{2|\mathcal{X}_k|} \cdot \exp\left(2n^{\frac{1-\gamma}{2}}\right) \cdot \mathbb{P}(\hat{P}_X \in \mathcal{M}_k^c) \\ &\le p_k. \end{aligned}$$

Appendix F. Proof of Proposition 3

Note that equivalently,

$$\theta_k = g_k^*(\hat{P}_{X_k}) + \tilde{Z}_k, \tag{A9}$$

where $\tilde{Z}_k \sim \mathcal{N}(0, \sigma_k^2/m)$. We then apply the typical result for Gaussian tail [32], i.e., for any $\alpha > 0$,

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\tilde{Z}_k > \alpha) = \frac{\alpha^2}{2\mu\sigma_k^2},$$

which implies that

$$\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\Big(\boldsymbol{Q}_k^{-1}(\boldsymbol{Q}_k(\hat{P}_{X_k})+\tilde{Z}_k)\neq\hat{P}_{X_k}\Big)\geq\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\bigg(|\tilde{Z}_k|>\frac{1}{2}\exp\Big(n^{\frac{1-\gamma}{2}}\Big)\bigg)=\infty.$$

Appendix G. Proof of Proposition 4

Note that

$$\begin{aligned}
&\mathbb{P}_n((\theta_1,\cdots,\theta_K),(\theta_1,\cdots,\theta_K)\in\Theta_\omega|\mathsf{H}=\mathsf{H}_0)\\
&\doteq\mathbb{P}_n\Big((\theta_1,\cdots,\theta_K),\hat{P}_{X_k}\in\mathcal{M}_k^{\mathsf{c}},\forall k\in\omega,\hat{P}_{X_{k'}}\notin\mathcal{M}_{k'}^{\mathsf{c}},\forall k'\in[K]\setminus\omega\Big|\mathsf{H}=\mathsf{H}_0\Big) \qquad \text{(A10)}\\
&=\sum_{\bar{\omega}\in\Im([K]\setminus\omega)}\Bigg\{\prod_{k'\in\bar{\omega}}\mathbb{P}\Big(\theta_{k'}\Big|\hat{P}_{X_{k'}}\in\mathcal{M}_k^{(0)}\Big)\cdot\prod_{k''\in[K]\setminus(\omega\cup\bar{\omega})}\mathbb{P}\Big(\theta_{k''}\Big|\hat{P}_{X_{k''}}\in\mathcal{M}_k^{(1)}\Big)\\
&\quad\cdot\prod_{k\in\omega}\sum_{\hat{P}_{X_k}\in\mathcal{P}_{\mathcal{X}_k}^{(n)}}\mathbb{P}(\theta_k|\hat{P}_{X_k})\,\mathbb{P}_n\Big(\hat{P}_{X_k},\hat{P}_{X_k}\in\mathcal{M}_k^{\mathsf{c}},\forall k\in\omega,\hat{P}_{X_{k'}}\in\mathcal{M}_{k'}^{(0)},\forall k'\in\bar{\omega},\\
&\quad\hat{P}_{X_{k''}}\in\mathcal{M}_{k''}^{(1)},\forall k''\in[K]\setminus(\omega\cup\bar{\omega})\Big|\mathsf{H}=\mathsf{H}_0\Big)\Bigg\}, \qquad \text{(A11)}
\end{aligned}$$

where (A10) comes from (42). By decoding the empirical distributions from θ_k with $\boldsymbol{Q}_k^{-1}(\cdot)$ for $k\in\omega$ and Proposition 3, we have

$$\begin{aligned}
&\sum_{\hat{P}_{X_k}\in\mathcal{P}_{\mathcal{X}_k}^{(n)}}\mathbb{P}(\theta_k|\hat{P}_{X_k})\,\mathbb{P}_n\Big(\hat{P}_{X_k},\hat{P}_{X_k}\in\mathcal{M}_k^{\mathsf{c}},\forall k\in\omega,\hat{P}_{X_{k'}}\in\mathcal{M}_{k'}^{(0)},\forall k'\in\bar{\omega},\hat{P}_{X_{k''}}\in\mathcal{M}_{k''}^{(1)},\\
&\quad\forall k''\in[K]\setminus(\omega\cup\bar{\omega})\Big|\mathsf{H}=\mathsf{H}_0\Big)\\
&\doteq\mathbb{P}(\theta_k|\hat{P}_{X_k}=\boldsymbol{Q}_k^{-1}(\theta_k))\,\mathbb{P}_n\Big(\hat{P}_{X_k}=\boldsymbol{Q}_k^{-1}(\theta_k),\hat{P}_{X_k}\in\mathcal{M}_k^{\mathsf{c}},\forall k\in\omega,\hat{P}_{X_{k'}}\in\mathcal{M}_{k'}^{(0)},\forall k'\in\bar{\omega},\\
&\quad\hat{P}_{X_{k''}}\in\mathcal{M}_{k''}^{(1)},\forall k''\in[K]\setminus(\omega\cup\bar{\omega})\Big|\mathsf{H}=\mathsf{H}_0\Big)\\
&\doteq\mathbb{P}(\theta_k|\hat{P}_{X_k}=\boldsymbol{Q}_k^{-1}(\theta_k))\cdot\exp\big(-n\cdot D_0^*(\bar{P}_{X_1},\cdots,\bar{P}_{X_k})\big).
\end{aligned}$$

With

$$\mathbb{P}(\theta_{k'}|\hat{P}_{X_{k'}}\in\mathcal{M}_k^{(0)})\doteq\exp\left(-n\cdot\frac{(\theta_{k'}-\sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2}\right),$$

and

$$\mathbb{P}(\theta_{k''}|\hat{P}_{X_{k''}}\in\mathcal{M}_k^{(1)})\doteq\exp\left(-n\cdot\frac{(\theta_{k''}-\sqrt{p'_{k''}})^2}{2\mu\sigma_{k''}^2}\right),$$

we have

$$\begin{aligned}
&\mathbb{P}_n((\theta_1,\cdots,\theta_K),(\theta_1,\cdots,\theta_K)\in\Theta_\omega|\mathsf{H}=\mathsf{H}_0)\\
&\doteq\sum_{\bar{\omega}\in\Im([K]\setminus\omega)}\Bigg\{\prod_{k\in\omega}\cdot\mathbb{P}(\theta_k|\hat{P}_{X_k}=\boldsymbol{Q}_k^{-1}(\theta_k))\cdot\exp\big(-n\cdot\tilde{E}_0^{\omega}(\theta_1,\cdots,\theta_K)\big)\Bigg\}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}_n((\theta_1,\cdots,\theta_K),(\theta_1,\cdots,\theta_K)\in\Theta_\omega|\mathsf{H}=\mathsf{H}_1)\\
&\doteq\sum_{\bar{\omega}\in\Im([K]\setminus\omega)}\Bigg\{\prod_{k\in\omega}\mathbb{P}(\theta_k|\hat{P}_{X_k}=\boldsymbol{Q}_k^{-1}(\theta_k))\cdot\exp\big(-n\cdot\tilde{E}_1^{\omega}(\theta_1,\cdots,\theta_K)\big)\Bigg\}.
\end{aligned}$$

Note that $\mathbb{P}(\theta_k|\hat{P}_{X_k} = \boldsymbol{Q}_k^{-1}(\theta_k))$ is not related to $\bar{\omega}$ and H, and then we can derive the decision rule (45) with LLRT. To compute the error exponent, we use Proposition 3 and the fact that $\mathbb{P}(\theta_k|\hat{P}_{X_k} = \boldsymbol{Q}_k^{-1}(\theta_k)) \doteq 1$ when $\theta_k = \boldsymbol{Q}(\hat{P}_{X_k})$. Then, the optimal error exponent corresponds to

$$\min_{\{\hat{P}_{X_k}\}_{k\in\omega},\{\theta_{k'}\}_{k'\in[K]\setminus\omega}} \max_{i=0,1} \min_{\bar{\omega}\in\Im([K]\setminus\omega)} D_i^*(\bar{R}_{X_1}^{\bar{\omega}},\cdots,\bar{R}_{X_K}^{\bar{\omega}}) + \sum_{k\in\bar{\omega}} \frac{(\theta_k - \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in[K]\setminus(\omega\cup\bar{\omega})} \frac{(\theta_{k'} + \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2}, \tag{A12}$$

where for $k = 1,\cdots,K$, and $\bar{\omega} \in \Im([K]\setminus\omega)$,

$$\bar{R}_{X_k}^{\bar{\omega}} \triangleq \begin{cases} \hat{P}_{X_k}, & \text{if } k \in \omega \\ P_{X_k}^{(0)}, & \text{if } k \in \bar{\omega} \\ P_{X_k}^{(1)}, & \text{if } k \in [K]\setminus(\omega\cup\bar{\omega}) \end{cases}. \tag{A13}$$

To finish the proof, we introduce the following lemma.

Lemma A1. *For arbitrary functions $v_1,\cdots,v_\ell\colon \mathcal{Z}\mapsto\mathbb{R}$ and $w_1,\cdots,w_{\ell'}\colon \mathcal{Z}\mapsto\mathbb{R}$, where $\mathcal{Z}$ is a given set, we have*

$$\begin{aligned} &\min_{z\in\mathcal{Z}} \max\big\{\min\{v_1(z),\cdots,v_\ell(z)\},\min\{w_1(z),\cdots,w_{\ell'}(z)\}\big\} \\ &= \min_{i\in\{1,\cdots,\ell\},j\in\{1,\cdots,\ell'\}} \min_{z\in\mathcal{Z}} \max\big\{v_i(z),w_j(z)\big\}. \end{aligned} \tag{A14}$$

With Lemma A1, we only need to compare each component in (A12), i.e.,

$$\begin{aligned} &\min_{\bar{\omega},\bar{\omega}'\in\Im([K]\setminus\omega)} \min_{\{\hat{P}_{X_k}\}_{k\in\omega},\{\theta_{k'}\}_{k'\in[K]\setminus\omega}} \max \\ &\Bigg\{ D_0^*(\bar{R}_{X_1}^{\bar{\omega}},\cdots,\bar{R}_{X_K}^{\bar{\omega}}) + \sum_{k\in\bar{\omega}} \frac{(\theta_k - \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in[K]\setminus(\omega\cup\bar{\omega})} \frac{(\theta_{k'} + \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2}, \\ &D_1^*(\bar{R}_{X_1}^{\bar{\omega}'},\cdots,\bar{R}_{X_K}^{\bar{\omega}'}) + \sum_{k\in\bar{\omega}'} \frac{(\theta_k - \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in[K]\setminus(\omega\cup\bar{\omega}')} \frac{(\theta_{k'} + \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2} \Bigg\}. \end{aligned} \tag{A15}$$

Given $\bar{\omega}$ and $\bar{\omega}'$, let $\tilde{\omega} = \bar{\omega}\cap\bar{\omega}'$. By selecting $\theta_k = \sqrt{p'_k}$ for $k \in \tilde{\omega}$ and $\theta_k = -\sqrt{p'_k}$ for $k \in [K]\setminus(\omega\cup(\bar{\omega}\cup\bar{\omega}'))$ in the minimization of (A15), (A15) equals

$$\begin{aligned} &\min_{\bar{\omega},\bar{\omega}'\in\Im([K]\setminus\omega)} \min_{\{\hat{P}_{X_k}\}_{k\in\omega},\{\theta_{k'}\}_{k'\in\bar{\omega}\cup\bar{\omega}'\setminus\tilde{\omega}}} \max \\ &\Bigg\{ D_0^*(\bar{R}_{X_1}^{\bar{\omega}},\cdots,\bar{R}_{X_K}^{\bar{\omega}}) + \sum_{k\in\bar{\omega}\setminus\tilde{\omega}} \frac{(\theta_k - \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in\bar{\omega}'\setminus\tilde{\omega}} \frac{(\theta_{k'} + \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2}, \\ &D_1^*(\bar{R}_{X_1}^{\bar{\omega}'},\cdots,\bar{R}_{X_K}^{\bar{\omega}'}) + \sum_{k\in\bar{\omega}\setminus\tilde{\omega}} \frac{(\theta_k + \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in\bar{\omega}'\setminus\tilde{\omega}} \frac{(\theta_{k'} - \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2} \Bigg\}. \end{aligned} \tag{A16}$$

In the following, we denote $\Omega \triangleq [K]\setminus(\omega\cup(\bar{\omega}\cup\bar{\omega}'))$. For those indices $k \in \tilde{\omega}$ or $k \in \Omega$, although they do not contribute to the Gaussian-like error exponents, they restrict that

$\bar{R}^{\bar{\omega}}_{X_k} = \bar{R}^{\bar{\omega}'}_{X_k} = P^{(0)}_{X_k}$ or $\bar{R}^{\bar{\omega}}_{X_k} = \bar{R}^{\bar{\omega}'}_{X_k} = P^{(1)}_{X_k}$. By letting $\bar{R}^{\bar{\omega}}_{X_k} = \bar{R}^{\bar{\omega}'}_{X_k} = \hat{P}_{X_k}$ ($k \in \tilde{\omega}$ or $k \in \Omega$) that can be optimized, we find the lower bound of (A15).

$$\begin{aligned}(\text{A15}) \geq & \min_{\bar{\omega},\bar{\omega}' \in \Im([K]\setminus\omega)} \ \min_{\{\hat{P}_{X_k}\}_{k\in\omega\cup\tilde{\omega}\cup\Omega}, \{\theta_{k'}\}_{k'\in\tilde{\omega}\cup\tilde{\omega}'\setminus\tilde{\omega}}} \ \max \\ & \left\{ D_0^*(\bar{R}^{\bar{\omega}}_{X_1}, \cdots, \bar{R}^{\bar{\omega}}_{X_K}) + \sum_{k\in\bar{\omega}\setminus\tilde{\omega}} \frac{(\theta_k - \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in\bar{\omega}'\setminus\tilde{\omega}} \frac{(\theta_{k'} + \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2}, \right. \\ & \left. D_1^*(\bar{R}^{\bar{\omega}'}_{X_1}, \cdots, \bar{R}^{\bar{\omega}'}_{X_K}) + \sum_{k\in\bar{\omega}\setminus\tilde{\omega}} \frac{(\theta_k + \sqrt{p'_k})^2}{2\mu\sigma_k^2} + \sum_{k'\in\bar{\omega}'\setminus\tilde{\omega}} \frac{(\theta_{k'} - \sqrt{p'_{k'}})^2}{2\mu\sigma_{k'}^2} \right\} \\ = & \min_{\bar{\omega},\bar{\omega}' \in \Im([K]\setminus\omega)} E_{\omega\cup\tilde{\omega}\cup\Omega} - \epsilon \geq E^{\odot} - \epsilon, \end{aligned} \tag{A17}$$

where we have used the fact that $\lim_{n\to\infty} p'_k = p_k$,

$$D_0^*(\bar{R}^{\bar{\omega}}_{X_1}, \cdots, \bar{R}^{\bar{\omega}}_{X_K}) = D_0^{\omega\cup\tilde{\omega}\cup\Omega}\big(\{\hat{P}_{X_k}\}_{k\in\omega\cup\tilde{\omega}\cup\Omega}\big),$$

$$D_1^*(\bar{R}^{\bar{\omega}'}_{X_1}, \cdots, \bar{R}^{\bar{\omega}'}_{X_K}) = D_1^{\omega\cup\tilde{\omega}\cup\Omega}\big(\{\hat{P}_{X_k}\}_{k\in\omega\cup\tilde{\omega}\cup\Omega}\big),$$

and have substituted $-\theta_{k'}$ for $\theta_{k'}$.

Appendix H. Proof of Proposition 6

Let the encoders for $k \in [K] \setminus \omega$ be functions of H and $\hat{P}_{X_k}$. The upper bound comes from the fact that the type is also generated from the hypothesis H. Therefore, the encoder on both the hypothesis and the type is just a function of the true hypothesis. Suppose that $\boldsymbol{\rho}_k \colon \{0,1\} \mapsto \mathbb{R}^m$ ($k \in [K] \setminus \omega$) satisfying $\frac{1}{m}\mathbb{E}[\|\boldsymbol{\rho}_k(\mathsf{H})\|^2] \leq p_k$. Let $\rho_k^{(i)}$ denote the i-th entry of $\boldsymbol{\rho}_k$, and

$$\rho_k^{(i)}(\mathsf{H}) \triangleq \begin{cases} \kappa_k^{(i)}, & \text{if } \mathsf{H} = \mathsf{H}_0 \\ \bar{\kappa}_k^{(i)}, & \text{if } \mathsf{H} = \mathsf{H}_1 \end{cases}, \tag{A18}$$

where $\frac{1}{2}\kappa_k^{(i)2} + \frac{1}{2}\bar{\kappa}_k^{(i)2} = p_k^{(i)}$ and $\sum_{i=1}^m p_k^{(i)} = p_k$. The error exponent with respect to the LLRT is

$$\begin{aligned} \min_{\{R_{X_k}\}_{k\in\omega}, \{\theta_k^{(i)}\}_{k\in[K]\setminus\omega, i=1,\cdots,m}} \ \max \Big\{ & \frac{1}{n}\sum_{k\in[K]\setminus\omega}\sum_{i=1}^m \frac{(\theta_k^{(i)} - \kappa_k^{(i)})^2}{2\sigma_k^2} + D_0^{\omega}(\{R_{X_k}\}_{k\in\omega}), \\ & \frac{1}{n}\sum_{k\in[K]\setminus\omega}\sum_{i=1}^m \frac{(\theta_k^{(i)} - \bar{\kappa}_k^{(i)})^2}{2\sigma_k^2} + D_1^{\omega}(\{R_{X_k}\}_{k\in\omega}) \Big\}. \end{aligned} \tag{A19}$$

Here, we explain the optimality of $\bar{\kappa}_k^{(i)} = -\kappa_k^{(i)} = -\sqrt{p_k^{(i)}}$, under which let $R^*_{X_k}, \theta_k^{(i)*}$ be the solution to problem (A19). For other pairs of $(\bar{\kappa}_k^{(i)}, \kappa_k^{(i)})$, $|\bar{\kappa}_k^{(i)} - \kappa_k^{(i)}| < 2\sqrt{p_k^{(i)}}$. Let $\tilde{\theta}_k^{(i)*} = \kappa_k^{(i)} + (\bar{\kappa}_k^{(i)} - \kappa_k^{(i)}) \cdot \frac{\theta_k^{(i)*} + \sqrt{p_k^{(i)}}}{2\sqrt{p_k^{(i)}}}$. Then, we have

$$\frac{(\theta_k^{(i)*} - \sqrt{p_k^{(i)}})^2}{2\sigma_k^2} \geq \frac{(\tilde{\theta}_k^{(i)*} - \kappa_k^{(i)})^2}{2\sigma_k^2},$$

and

$$\frac{(\theta_k^{(i)*}+\sqrt{p_k^{(i)}})^2}{2\sigma_k^2} \geq \frac{(\tilde{\theta}_k^{(i)*}-\bar{\kappa}_k^{(i)})^2}{2\sigma_k^2},$$

which will lead to a smaller error exponent (cf. (A19)) and the optimality is proved. The solution to problem (A19) is

$$\begin{aligned}
&\lim_{n\to\infty} \min_{\{R_{X_k}\}_{k\in\omega},\{\theta_k^{(i)}\}_{k\in[K]\setminus\omega,i=1,\cdots,m}} \max\Big\{\frac{1}{n}\sum_{k\in[K]\setminus\omega}\sum_{i=1}^{m}\frac{(\theta_k^{(i)}-\sqrt{p_k^{(i)}})^2}{2\sigma_k^2}+D_0^{\omega}(\{R_{X_k}\}_{k\in\omega}),\\
&\qquad\qquad \frac{1}{n}\sum_{k\in[K]\setminus\omega}\sum_{i=1}^{m}\frac{(\theta_k^{(i)}+\sqrt{p_k^{(i)}})^2}{2\sigma_k^2}+D_1^{\omega}(\{R_{X_k}\}_{k\in\omega})\Big\}\\
&= \min_{\{R_{X_k}\}_{k\in\omega},\{\theta_k\}_{k\in[K]\setminus\omega}} \max\Big\{D_0^{\omega}(\{R_{X_k}\}_{k\in\omega})+\sum_{k\in[K]\setminus\omega}\frac{(\theta_k-\sqrt{p_k})^2}{2\mu\sigma_k^2},\\
&\qquad\qquad D_1^{\omega}(\{R_{X_k}\}_{k\in\omega})+\sum_{k\in[K]\setminus\omega}\frac{(\theta_k+\sqrt{p_k})^2}{2\mu\sigma_k^2}\Big\}\\
&= E_{\omega}. \qquad\qquad (A20)
\end{aligned}$$

Appendix I

Based on the results in Appendix D, E_ω as defined in (32) satisfies

$$\begin{aligned}
E_{\omega} = \min_{\phi_\omega\in\mathbb{R}^{k_\omega},\{\theta_k\}_{k\in[K]\setminus\omega}} \max\Big\{&\frac{1}{8}\big\|B_\omega(B_\omega^{\dagger}\psi_\omega^{(0)}-\phi_\omega)\big\|^2+\sum_{k\in[K]\setminus\omega}\frac{(\theta_k-\sqrt{p_k})^2}{2\mu\sigma_k^2},\\
&\frac{1}{8}\big\|B_\omega(B_\omega^{\dagger}\psi_\omega^{(1)}-\phi_\omega)\big\|^2+\sum_{k\in[K]\setminus\omega}\frac{(\theta_k+\sqrt{p_k})^2}{2\mu\sigma_k^2}\Big\}+o(\epsilon^2), \qquad (A21)
\end{aligned}$$

where $k_\omega \triangleq \sum_{k\in\omega}|\mathcal{X}_k|$, and then the result can be easily verified using Lagrangian multipliers.

References

1. Han, T.S.; Amari, S. Statistical inference under multiterminal data compression. *IEEE Trans. Inf. Theory* **1998**, *44*, 2300–2324. [CrossRef]
2. Ahlswede, R.; Csiszár, I. Hypothesis testing with communication constraints. *IEEE Trans. Inf. Theory* **1986**, *32*, 533–542. [CrossRef]
3. Han, T.S.; Kobayashi, K. Exponential-type error probabilities for multiterminal hypothesis testing. *IEEE Trans. Inf. Theory* **1989**, *35*, 2–14. [CrossRef]
4. Amari, S.I.; Han, T.S. Statistical inference under multiterminal rate restrictions: A differential geometric approach. *IEEE Trans. Inf. Theory* **1989**, *35*, 217–227. [CrossRef]
5. Watanabe, S. Neyman–Pearson test for zero-rate multiterminal hypothesis testing. *IEEE Trans. Inf. Theory* **2017**, *64*, 4923–4939. [CrossRef]
6. Shimokawa, H.; Han, T.S.; Amari, S. Error bound of hypothesis testing with data compression. In Proceedings of the 1994 IEEE International Symposium on Information Theory, Trondheim, Norway, 27 June–1 July 1994; p. 114. [CrossRef]
7. Xu, X.; Huang, S.L. On Distributed Learning with Constant Communication Bits. *IEEE J. Sel. Areas Inf. Theory* **2022**, *3*, 125–134. [CrossRef]
8. Sreekumar, S.; Gündüz, D. Strong Converse for Testing Against Independence over a Noisy channel. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 1283–1288. [CrossRef]
9. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 1273–1282.
10. Vapnik, V. Principles of Risk Minimization for Learning Theory. In Proceedings of the 4th International Conference on Neural Information Processing Systems, San Francisco, CA, USA, 2–5 December 1991; pp. 831–838.

11. Srivastava, N.; Salakhutdinov, R. Multimodal learning with deep boltzmann machines. *J. Mach. Learn. Res.* **2014**, *15*, 2949–2980.
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
13. Huang, S.L.; Makur, A.; Wornell, G.W.; Zheng, L. On universal features for high-dimensional learning and inference. *arXiv* **2019**, arXiv:1911.09105.
14. Han, T.S. Hypothesis testing with multiterminal data compression. *IEEE Trans. Inf. Theory* **1987**, *33*, 759–772. [CrossRef]
15. Scardapane, S.; Wang, D.; Panella, M.; Uncini, A. Distributed learning for random vector functional-link networks. *Inf. Sci.* **2015**, *301*, 271–284.
16. Georgopoulos, L.; Hasler, M. Distributed machine learning in networks by consensus. *Neurocomputing* **2014**, *124*, 2–12. [CrossRef]
17. Tsitsiklis, J.; Athans, M. On the complexity of decentralized decision making and detection problems. *IEEE Trans. Autom. Control* **1985**, *30*, 440–446. [CrossRef]
18. Tsitsiklis, J.N. Decentralized detection by a large number of sensors. *Math. Control. Signals Syst.* **1988**, *1*, 167–182. [CrossRef]
19. Tenney, R.R.; Sandell, N.R. Detection with distributed sensors. *IEEE Trans. Aerosp. Electron. Syst.* **1981**, *AES-17*, 501–510. [CrossRef]
20. Shalaby, H.M.; Papamarcou, A. Multiterminal detection with zero-rate data compression. *IEEE Trans. Inf. Theory* **1992**, *38*, 254–267. [CrossRef]
21. Zhao, W.; Lai, L. Distributed testing with zero-rate compression. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 2792–2796.
22. Sreekumar, S.; Gündüz, D. Distributed hypothesis testing over noisy channels. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 983–987.
23. Zaidi, A. Hypothesis Testing Against Independence Under Gaussian Noise. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 1289–1294. [CrossRef]
24. Salehkalaibar, S.; Wigger, M.A. Distributed hypothesis testing over a noisy channel. In Proceedings of the International Zurich Seminar on Information and Communication (IZS 2018), Zurich, Switzerland, 21–23 February 2018; pp. 25–29.
25. Weinberger, N.; Kochman, Y.; Wigger, M. Exponent trade-off for hypothesis testing over noisy channels. In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019; pp. 1852–1856.
26. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
27. Csiszár, I.; Shields, P.C. *Information Theory and Statistics: A Tutorial*; Now Publishers Inc.: Delft, The Netherlands, 2004.
28. Csiszár, I. The method of types [information theory]. *IEEE Trans. Inf. Theory* **1998**, *44*, 2505–2523. [CrossRef]
29. Huang, S.L.; Xu, X.; Zheng, L. An information-theoretic approach to unsupervised feature selection for high-dimensional data. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 157–166. [CrossRef]
30. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
31. Graham, R.L.; Knuth, D.E.; Patashnik, O.; Liu, S. Concrete mathematics: A foundation for computer science. *Comput. Phys.* **1989**, *3*, 106–107. [CrossRef]
32. Blair, J.; Edwards, C.; Johnson, J.H. Rational Chebyshev approximations for the inverse of the error function. *Math. Comput.* **1976**, *30*, 827–830. [CrossRef]