

# Anonymous Heterogeneous Distributed Detection: Optimal Decision Rules, Error Exponents, and the Price of Anonymity

Wei-Ning Chen and I-Hsiang Wang

## Abstract

We explore the fundamental limits of heterogeneous distributed detection in an anonymous sensor network with  $n$  sensors and a single fusion center. The fusion center collects the single observation from each of the  $n$  sensors to detect a binary parameter. The sensors are clustered into multiple groups, and different groups follow different distributions under a given hypothesis. The key challenge for the fusion center is the anonymity of sensors – although it knows the exact number of sensors and the distribution of observations in each group, it does not know which group each sensor belongs to. It is hence natural to consider it as a composite hypothesis testing problem. First, we propose an optimal test called *mixture likelihood ratio test*, which is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under one hypothesis to that under the other hypothesis. Optimality is shown by first arguing that there exists an optimal test that is *symmetric*, that is, it does not depend on the order of observations across the sensors, and then proving that the mixture likelihood ratio test is optimal among all symmetric tests. Second, we focus on the Neyman-Pearson setting and characterize the error exponent of the worst-case type-II error probability as  $n$  tends to infinity, assuming the number of sensors in each group is proportional to  $n$ . Finally, we generalize our result to find the collection of all achievable type-I and type-II error exponents, showing that the boundary of the region can be obtained by solving a convex optimization problem. Our results elucidate the price of anonymity in heterogeneous distributed detection, and can be extended to  $M$ -ary hypothesis testing with heterogeneous observations generated according to hidden latent variables. The results are also applied to distributed detection under Byzantine attacks, which hints that the conventional approach based on simple hypothesis testing might be too pessimistic.

## I. INTRODUCTION

In wireless sensor networks, the cost of identifying individual sensors increases drastically as the number of sensors grows. For distributed detection [1], when the observations follow identical and independent distributions (i.i.d.) across all sensors, identifying individual sensors is not very important. When the fusion center can fully access

The material in this paper is presented in part at the IEEE International Symposium on Information Theory, Vail, Colorado, USA, June 2018.

W.-N. Chen is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (email: r05942078@ntu.edu.tw).

I.-H. Wang is with the Department of Electrical Engineering and the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (email: ihwang@ntu.edu.tw).

the observations, the empirical distribution (types) of the collected observation is a sufficient statistic. When the communication between each sensor and the fusion center is limited, for binary hypothesis testing it is asymptotically optimal to use the same local decision function at all sensors [2]. Hence, anonymity is not a critical issue for the classical (homogeneous) distributed detection problem.

However, when the joint distribution of the observations is *heterogeneous*, that is, marginal distributions of observations vary across sensors, sensor anonymity may deteriorate the performance of distributed detection, even for binary hypothesis testing. One such example is distributed detection under Byzantine attack [3], where a fixed number of sensors are compromised by malicious attackers and report fake observations following certain distributions. Even if the fusion center is aware of the number of compromised sensors and the attacking strategy that renders worst-case detection performance (the least favorable distribution as considered in [4]–[6]), it is more difficult to detect the hidden parameter when the fusion center does not know which sensors are compromised.

In this paper, we aim to quantify the performance loss due to sensor anonymity in heterogeneous distributed detection, with  $n$  sensors and a single fusion center. Each sensor (say sensor  $i$ ,  $i \in \{1, \dots, n\}$ ) has a single random observation  $X_i$ . The goal of the fusion center is to estimate the hidden parameter  $\theta \in \{0, 1\}$  (that is, binary hypothesis testing) from the collected observations. The distributions of the observations, however, are *heterogeneous* – observations at different sensors may follow different sets of distributions. In particular, we assume that these  $n$  sensors are clustered into  $K$  groups  $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ , and group  $\mathcal{I}_k \subseteq \{1, \dots, n\}$  comprises  $n_k$  sensors, for  $k = 1, \dots, K$ . Under hypothesis  $\mathcal{H}_\theta$ ,  $\theta \in \{0, 1\}$ ,

$$X_i \sim P_{\theta;k}, \text{ for } i \in \mathcal{I}_k.$$

Moreover, the sensors are *anonymous*, that is, the collected observations at the fusion center are *unordered*. In other words, although the fusion center is fully aware of the *heterogeneity* of its observation, including the set of distributions  $\{P_{\theta;k} \mid \theta \in \{0, 1\}, k = 1, \dots, K\}$  and  $\{n_k \mid k = 1, \dots, K\}$ , it does not know what distribution each individual sensor will follow.

To address the lack of knowledge about the exact distributions of the observations, we formulate the detection problem as a *composite hypothesis testing* problem, where the vector observation of length  $n$  follows a product distribution within a finite class of  $n$ -letter product distributions under a given parameter  $\theta$ . The class consists of  $\binom{n}{n_1, \dots, n_K}$  possible product distributions, each of which follows one of the  $\binom{n}{n_1, \dots, n_K}$  possible partitions of the sensors. The fusion center takes all the possible partitions into consideration when detecting the hidden parameter. We mainly focus on a Neyman-Pearson setting, where the goal is to minimize the worst-case type-II error probability such that the worst-case type-I error probability is not larger than a constant. Towards the end of this paper, we also extend our results to a Bayesian setting, where a binary prior distribution is laid on  $\mathcal{H}_0$  and  $\mathcal{H}_1$ .

Our main contribution comprises three parts. First, we develop an optimal test, termed *mixture likelihood ratio test* (MLRT), for the anonymous heterogeneous distributed detection problem. MLRT is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under hypothesis  $\mathcal{H}_1$  to the uniform mixture of those under  $\mathcal{H}_0$ . To prove the optimality, we first argue that there exists an optimal test that is *symmetric*, that is, it does not depend on the order of observations across the sensors, and thus we only need to consider tests

which depend on the histogram of observations. In other words, the histogram of observations contains sufficient information for optimal detection. Moreover, all possible distributions over the space of observations  $\mathcal{X}^n$  under  $\mathcal{H}_0$  (or  $\mathcal{H}_1$ ) turn out to be the same one over the space of its histogram, so if we test the hypothesis according to the histogram, the original composite hypothesis testing problem boils down to a simple hypothesis testing problem. The one-to-one correspondence between symmetric tests and tests defined on the histogram is the key to derive optimal test. This result extends to  $M$ -ary hypothesis testing with heterogeneous observations generated according to hidden latent variables, each of which is associated to a observation, but the decision maker only knows the histogram of the latent variables.

Second, for the case that the alphabet  $\mathcal{X}$  is a finite set, we characterize the error exponent of the minimum worst-case type-II error probability as  $n \rightarrow \infty$  with the ratios  $\frac{n_k}{n} \rightarrow \alpha_k \forall k = 1, \dots, K$ . The optimal error exponent turns out to be the minimization of a linear combination of Kullback-Leibler divergences (KL divergences) with the  $k$ -th term being  $D(U_k \| P_{1;k})$  and  $\alpha_k$  being the coefficient, for  $k = 1, \dots, K$ . The minimization is over all possible distributions  $U_1, \dots, U_K$  such that  $\sum_{k=1}^K \alpha_k U_k = \sum_{k=1}^K \alpha_k P_{0;k}$ . In a simple hypothesis testing problem with i.i.d. observations, a standard approach to derive the type-II error exponent is invoking a strong converse lemma (see, for example, Chapter 12 in [7]) to relate the type-I and type-II error probability of an optimal test, and then applying the large deviation toolkit on the optimal test to single-letterize and find the exponent. In contrast, in our problem, neither can the mixture distributions in the optimal test be decomposed into a product form, nor can the acceptance region be bounded by a large deviation event, making this approach fail to characterize the error exponent. To circumvent the difficulties, we turn to the method of types and use bounds on types (empirical distributions) for single-letterization.

For achievability, instead of the optimal MLRT which is difficult to single-letterize, we employ a simpler test that resemble Hoeffding's test [8]. For the converse, we use an argument based on the method of types. We propose a generalized divergence  $D_{\alpha_1, \dots, \alpha_K}(P_1, \dots, P_K; Q_1, \dots, Q_K)$  from a group of distributions  $\{Q_1, \dots, Q_K\}$  to another group of distributions  $\{P_1, \dots, P_K\}$ , which plays a similar role as KL divergence in simple hypothesis testing problems. The key to the characterization of the optimal error exponent is to prove a generalized Sanov Theorem for the composite setting we considered. Based on the characterized error exponent, given the number of bits that a sensor can send to the fusion center, one can also formulate an optimization problem to find the best local decision functions, as in the homogeneous case [2].

Finally, we extend our results from the Neyman-Pearson setting to a Bayesian setting, minimizing the average probability of error (that is, combining type-I and type-II error). It can be shown that the optimal test is computationally infeasible, since it involves summation over all possible permutations. To overcome the complexity issue, we propose an asymptotically optimal test based on information geometry, which achieves the same error exponent of the average probability of error. We also study the exponent region  $\mathcal{R}$ , the collection of all pairs of achievable type-I and type-II error exponents. In particular, we propose a way to parametrize the contour of  $\mathcal{R}$  based on information projection. However, the closed-form expression of  $\mathcal{R}$  involves an explicit solution of a convex optimization problem, which remains unsettled.

As a by-product, we apply our results for  $K = 2$  to the distributed detection problem under Byzantine attack

and further obtain bounds on the worst-case type-II error exponent. Compared with the worst-case exponent in an alternative Bayesian formulation [3] where the observation of sensors are assumed to be i.i.d. according to a mixture distribution, it is shown that the worst-case exponent in the composite testing formulation is strictly larger. This hints that the conventional approach taken in [3] might be too pessimistic.

### *Related Works*

Decentralized detection is a classical topic, and attracts extensive attention in recent years due to its application in wireless sensor networks. See, for example, [1], [2], [6], [9]. Most works in decentralized detection are focused on finding optimal local decision function in both Neyman-Pearson and Bayesian regime. Under some assumptions on the distribution of a given hypothesis, optimal design criteria of local decision function and the decision rule at the fusion center are given. Unlike the anonymous setting considered in our work, the above-mentioned classical works assume fusion centers, as well as the local sensors, have perfect knowledge about the joint distribution, and hence the decision rules are designed according to it. This is termed an “informed” setting in our paper and is used as a baseline to compare with and see the price of anonymity. On the other hand, in our setting, the fusion center collects observations without knowing the exact index of each one, and thus the problem is formulated into a composite hypothesis testing problem.

Composite hypothesis testing is a long-standing problem in statistics, and is notoriously difficult to find an optimal test. In general, the uniform most powerful (UMP) test does not exist, see, for example, Section 8.3 in [10]. Even if we relax the performance evaluation to the *minimax* regime, the general form of the optimal test is still unknown, except for some special case. For example, [5] considered the case that the composite hypothesis class  $\mathcal{H}_\theta$  is formed by all  $\epsilon$ -contaminated distributions of  $P_\theta$ , that is,  $\{(1 - \epsilon)P_\theta + \epsilon Q \mid \forall \text{ possible distributions } Q\}$ . Under this structure, Huber showed that a censored version of likelihood ratio test is optimal in the minimax regime. Other works such as [8], [11] followed the idea of Hoeffding’s test [8] and proposed an universal asymptotically optimal test when the null hypothesis is simple. Meanwhile, in our setting, neither the parameter space of the considered distributions is continuous, nor the null hypothesis is simple, making their approaches hard to extend. Another common test for composite hypothesis testing is the *generalized likelihood ratio test* (GLRT). The optimality of GLRT is guaranteed under some circumstances, see, for example, [12]. However, the results in [12] hold only for simple null and composite alternative. In contrast, our result indicates that GLRT is not optimal in our setting.

The concept of Byzantine attack can be traced back to [13] (known as the “Byzantine Generals Problem”), in which reliability of a computer system with malfunctioned components is studied. After that, Byzantine model is developed and generalized by several research areas, especially in communication security. For example, the distributed detection with Byzantine attack is studied under the Neyman-Pearson formulation in [3] and under the Bayesian setting in [14]. In their settings, each sensor is assumed to be compromised with probability  $\alpha$ , so the observation turns out to be drawn identically and independently from an mixture distribution, making the hypothesis testing problem simple, and thus Neyman-Pearson lemma can be applied. In contrast, in our work we assume the number of Byzantine sensors is fixed and is  $\alpha n$ , where  $n$  is the total number of sensors, and thus the problem falls into a composite hypothesis testing instead of the mixture setting.

This work is presented in part at ISIT 2018. In the conference version [15], upper and lower bounds on the type-II error exponent were given, where the lower bound (achievability) is based on an modified version of Hoeffding's test, and the upper bound (converse) is derived by relaxing the original problem into a simple hypothesis testing. In this journal version, we show that the achievability bound in the conference version is indeed tight, closing the gap between the upper and lower bounds.

The rest of this paper is organized as follows. In Section II, we formulate the composite hypothesis testing problem for anonymous heterogeneous distributed detection and provide some background. In Section III, the main results are provided, where the proofs are delegated to Section IV and V. In Section VI, we generalize the results to the Bayesian setting, and in Section VII, we briefly discuss the case when  $\mathcal{X}$  is not finite, and the case when partial information about the group assignment is available at the fusion center. Finally, we conclude the paper with some further directions and open questions in Section VIII.

## II. PROBLEM FORMULATION AND PRELIMINARIES

### A. Problem Setup

Following the description of the setting in Section I, let us formulate the composite hypothesis testing problem. Let  $\sigma(i)$  denote the label of the group that sensor  $i$  belongs to. This labeling  $\sigma(\cdot)$ , however, is not revealed to the fusion center. Hence, the fusion center needs to consider all  $\binom{n}{n_1, \dots, n_K}$  possible  $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$  satisfying

$$|\{i \mid \sigma(i) = k\}| = n_k, \quad \forall k = 1, \dots, K, \quad (1)$$

and decides whether the hidden  $\theta$  is 0 or 1. For notational convenience, let  $\boldsymbol{\nu}$  denote the vector  $[n_1 \dots n_K]^T$ , and let  $\mathcal{S}_{n, \boldsymbol{\nu}}$  denote the collection of all labelings satisfying (1).

Hence, the fusion center is faced with the following *composite* hypothesis testing problem, where the goal is to infer the parameter  $\theta$ :

$$\mathcal{H}_\theta : X^n \sim \mathbb{P}_{\theta; \sigma} \triangleq \prod_{i=1}^n P_{\theta; \sigma(i)}, \quad \text{for some } \sigma \in \mathcal{S}_{n, \boldsymbol{\nu}}.$$

As mentioned in Section I, throughout the paper we consider binary hypothesis testing, that is,  $\theta \in \{0, 1\}$ .

Let each single observation take values from some measurable space  $(\mathcal{X}, \mathcal{F})$ , where  $\mathcal{F}$  is a  $\sigma$ -algebra on  $\mathcal{X}$ . Hence  $P_{\theta; k} \in \mathcal{P}_{\mathcal{X}}$  for all  $\theta \in \{0, 1\}$  and  $k \in \{1, \dots, K\}$ , where  $\mathcal{P}_{\mathcal{X}}$  denotes the collection of all possible distributions over  $(\mathcal{X}, \mathcal{F})$ . The vector observation  $x^n$  is defined on the space  $(\mathcal{X}^n, \mathcal{F}^{\otimes n})$ , where  $\mathcal{F}^{\otimes n}$  is the *tensor product*  $\sigma$ -algebra of  $\mathcal{F}$ , that is, the smallest  $\sigma$ -algebra contains the following collection of events:

$$\{\mathcal{E}_1 \times \mathcal{E}_2 \times \dots \times \mathcal{E}_n \mid \mathcal{E}_i \in \mathcal{F}\}.$$

A (randomized) test is a measurable function  $\phi : (\mathcal{X}^n, \mathcal{F}^{\otimes n}) \rightarrow ([0, 1], \mathfrak{B})$ , where  $\mathfrak{B}$  denotes the Borel  $\sigma$ -field on  $\mathbb{R}$ . The worst-case type-I and type-II error probabilities of a decision rule  $\phi$  are defined as

$$\begin{aligned} P_F^{(n)}(\phi) &\triangleq \max_{\sigma \in \mathcal{S}_{n, \boldsymbol{\nu}}} \mathbb{E}_{\mathbb{P}_{0; \sigma}} [\phi(X^n)] \quad (\text{Type I}) \\ P_M^{(n)}(\phi) &\triangleq \max_{\sigma \in \mathcal{S}_{n, \boldsymbol{\nu}}} \mathbb{E}_{\mathbb{P}_{1; \sigma}} [1 - \phi(X^n)] \quad (\text{Type II}). \end{aligned}$$

Our focus is on the Neyman-Pearson setting: find a decision rule  $\phi$  satisfying  $P_F^{(n)}(\phi) \leq \epsilon$  such that  $P_M^{(n)}(\phi)$  is minimized. Let  $\beta^{(n)}(\epsilon, \boldsymbol{\nu})$  denote the minimum type-II error probability.

For the asymptotic regime, we assume that the ratio  $\frac{n_k}{n} \rightarrow \alpha_k$  as  $n \rightarrow \infty$  for all  $k = 1, \dots, K$ , and  $\sum_{k=1}^K \alpha_k = 1$ . We aim to explore if  $\beta^{(n)}(\epsilon, \boldsymbol{\nu})$  decays exponentially fast as  $n \rightarrow \infty$ , and characterize the corresponding error exponent. For notational convenience, we define upper and lower bounds on the exponent:

$$\begin{aligned} \overline{E}^*(\epsilon, \boldsymbol{\alpha}) &\triangleq \limsup_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log_2 \beta^{(n)}(\epsilon, \boldsymbol{\nu}) \right\}, \\ \underline{E}^*(\epsilon, \boldsymbol{\alpha}) &\triangleq \liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log_2 \beta^{(n)}(\epsilon, \boldsymbol{\nu}) \right\}, \end{aligned}$$

where in taking the limits, we assume that  $\lim_{n \rightarrow \infty} \frac{n_k}{n} = \alpha_k$ , for all  $k = 1, \dots, K$ . If the upper and lower bound match, we simply denote it as  $E^*(\epsilon, \boldsymbol{\alpha})$ .

**Remark 2.1.** *The original distributed detection problem [1], [2], [6] involves local decision functions at the sensors to address the limited communication between each sensor and the fusion center. In order to focus on the impact of anonymity, we first absorb them into the distributions  $\{P_{\theta;k} : k = 1, \dots, K\}$  because they are symbol-by-symbol maps. Later, we will discuss how to find the best local decision functions according to the characterized error exponent.*

## B. Notations

Let us introduce notations that will be used throughout this paper.

- $n$  denotes the total number of observations, and  $K$  denotes the number of groups of sensors.
- $\boldsymbol{\nu} \triangleq [n_1 \dots n_K]^\top$  denotes the number of sensors in the  $K$  groups. That is,  $n_k \geq 0$ ,  $n_k \in \mathbb{Z}$ , and  $\sum_{k=1}^K n_k = n$ .
- $\boldsymbol{\alpha} \triangleq [\alpha_1 \dots \alpha_K]^\top$  denotes the fraction of each group of sensors in all sensors in the asymptotic regime. That is,  $\alpha_k \geq 0$ , and  $\sum_{k=1}^K \alpha_k = 1$ .
- $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$  is the labeling function which assigns the index of each sensor to a group. We also denote the collection of indices of sensors in group  $k$  as

$$\mathcal{I}_k = \sigma^{-1}(k) \triangleq \{i \mid \sigma(i) = k\}. \quad (2)$$

- Let  $\mathcal{S}_{n, \boldsymbol{\nu}}$  be the collection of all  $\sigma$  satisfying (2). We also use  $\mathcal{S}_n$  to denote the collection of length- $n$  permutations:

$$\mathcal{S}_n \triangleq \left\{ \tau : \{1, 2, \dots, n\} \xrightarrow{1 \rightarrow 1} \{1, 2, \dots, n\} \right\}.$$

Note that the cardinalities of the two sets are

$$|\mathcal{S}_{n, \boldsymbol{\nu}}| = \binom{n}{n_1, n_2, \dots, n_K}, \quad |\mathcal{S}_n| = n!.$$

- We usually write  $\mathbf{P}_\theta$  as the vector of  $\{P_{\theta;k}\}$ :

$$\mathbf{P}_\theta \triangleq \begin{bmatrix} P_{\theta;1} \\ P_{\theta;2} \\ \vdots \\ P_{\theta;K} \end{bmatrix}.$$

### C. Method of Types

For a sequence  $x^n \in \mathcal{X}^n$ , where  $\mathcal{X} = \{a_1, a_2, \dots, a_d\}$ , its type (empirical distribution) is defined as

$$\Pi_{x^n} = [\pi(a_1|x^n), \pi(a_2|x^n), \dots, \pi(a_d|x^n)],$$

where  $\pi(a_i|x^n)$  is the frequency of  $a_i$  in the sequence  $x^n$ , that is,

$$\pi(a_i|x^n) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{x_j=a_i\}}.$$

For a given length  $n$ , we use  $\mathcal{P}_n$  to denote the collection of possible types of length- $n$  sequences. In other words,

$$\mathcal{P}_n \triangleq \left\{ \left[ \frac{i_1}{n}, \frac{i_2}{n}, \dots, \frac{i_d}{n} \right] \mid \forall i_1, \dots, i_d \in \mathbb{N} \cup \{0\}, i_1 + i_2 + \dots + i_d = n \right\}.$$

Let  $U \in \mathcal{P}_n$  be an  $n$ -type. The type class  $T_n(U)$  is the set of all length- $n$  sequences with type  $U$ ,

$$T_n(U) \triangleq \{x^n \in \mathcal{X}^n \mid \Pi_{x^n} = U\}.$$

Let us introduce some useful lemmas about type.

**Lemma 2.1** (Cardinality Bound of  $\mathcal{P}_n$ ).

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}.$$

In words,  $|\mathcal{P}_n|$  grows polynomial in  $n$ .

**Lemma 2.2** (Probability of Type Class). *Let  $P \in \mathcal{P}_n, Q \in \mathcal{P}_{\mathcal{X}}$ . Then*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q\|P)} \leq Q^{\otimes n}(T_n(P)) \leq 2^{-nD(Q\|P)}.$$

For finite  $\mathcal{X}$ ,  $\mathcal{P}_{\mathcal{X}}$  can be viewed as a subspace in  $\mathbb{R}^d$  endowed with Euclidean metric and standard topology. The following theorem, developed by Sanov, depicts the probability of a large deviation event.

**Lemma 2.3** (Sanov's Theorem). *Let  $\Gamma \subseteq \mathcal{P}_{\mathcal{X}}$ . Then we have*

$$-\inf_{T \in \text{int } \Gamma} D(T\|Q) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} \leq -\inf_{T \in \text{cl } \Gamma} D(T\|Q), \quad (3)$$

where  $\text{int } \Gamma$  and  $\text{cl } \Gamma$  respectively denote the interior and the closure of  $\Gamma$ , with respect to the standard topology on  $\mathbb{R}^d$ . In particular, if the infimum on the right-hand side is equal to the infimum on the left-hand side in (3), we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} = -\inf_{T \in \Gamma} D(T\|Q).$$

Proofs of the lemmas mentioned above can be found in standard information theory textbooks, Chapter 11 in [16] for example. Alternatively, a more rigorous proof of Sanov's theorem Lemma 2.3 can be found in [17].

### III. MAIN RESULTS

As mentioned in Section II, the observations come from the measurable space  $(\mathcal{X}^n, \mathcal{F}^{\otimes n})$ . Throughout the rest of the paper, we assume that  $\mathcal{X}$  is a totally ordered set, and  $\mathcal{F}^{\otimes n}$  satisfies the following two assumptions:

1)  $\mathcal{F}^{\otimes n}$  contains the following set:

$$\tilde{\mathcal{X}}^n \triangleq \{(x_1, x_2, \dots, x_n) \mid x_1 \geq x_2 \geq \dots \geq x_n\}. \quad (4)$$

2)  $\mathcal{F}^{\otimes n}$  is closed under permutation. That is, if  $\mathcal{A} \in \mathcal{F}^{\otimes n}$ , for any length- $n$  permutation  $\tau : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ ,

$$\mathcal{A}_\tau \triangleq \{(x_{\tau(1)}, \dots, x_{\tau(n)}) \mid (x_1, \dots, x_n) \in \mathcal{A}\} \in \mathcal{F}. \quad (5)$$

**Remark 3.1.** We assume that  $\mathcal{X}$  is a totally ordered set in order to set the condition such that  $\tilde{\mathcal{X}}$  is measurable. The purpose to require  $\tilde{\mathcal{X}}$  to be measurable is to preserve the measurability of the ordering map  $\Pi(\cdot)$ , as later defined in Definition 4.1. In general, if  $\mathcal{X}$  is not totally ordered, we can still require the collection of representatives in the equivalent classes induced by  $\Pi^{-1}$  to be measurable. However, the regularity assumptions on  $\mathcal{F}^{\otimes n}$  need to be carefully concerned in that case.

**Remark 3.2.** The second assumption always holds for tensor  $\sigma$ -fields. The first assumption typically holds too. For example, if  $\mathcal{X}$  is finite, we can simply choose  $\mathcal{F}$  as the power set  $2^{\mathcal{X}}$ , and if  $\mathcal{X} \subseteq \mathbb{R}$ , we can choose  $\mathcal{F}$  as the Borel  $\sigma$ -field. In particular, for  $\mathcal{X}$  being a finite set, it is straightforward to define a total order over it, and hence it is a totally ordered set. Moreover, the above two assumptions are automatically satisfied.

#### A. Main Contributions

Our first contribution is the characterization of the optimal test:

**Theorem 3.1** (Optimal Test). Define the mixture likelihood ratio  $\ell(x^n)$ :

$$\ell(x^n) \triangleq \frac{\sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma}(x^n)}{\sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma}(x^n)}. \quad (6)$$

Suppose  $\mathcal{F}^{\otimes n}$  satisfies the two assumptions (4), (5). Then an optimal tests  $\phi^*(x^n)$  takes the following form:

$$\phi^*(x^n) = \begin{cases} 1, & \text{if } \ell(x^n) > \tau \\ \gamma, & \text{if } \ell(x^n) = \tau \\ 0, & \text{if } \ell(x^n) < \tau. \end{cases} \quad (7)$$

That is, for any test  $\phi$ , we have

$$\mathbb{P}_F(\phi) \leq \mathbb{P}_F(\phi^*) \Rightarrow \mathbb{P}_M(\phi) \geq \mathbb{P}_M(\phi^*).$$

**Remark 3.3.** We see that the optimal test, MLRT, is the likelihood ratio test between two uniform mixture distributions

$$\frac{1}{|\mathcal{S}_{n,\nu}|} \sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{\theta;\sigma}, \theta \in \{0, 1\}.$$



Interestingly, the optimality of MLRT indicates that the widely used decision rule, generalized likelihood ratio test (GLRT), which is defined as the randomized thresholded test according to the following likelihood ratio

$$\ell_{GLRT}(x^n) \triangleq \frac{\sup_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma}(x^n)}{\sup_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma}(x^n)},$$

is strictly sub-optimal in the anonymous hypothesis testing problem.

*Sketch of proof:* The proof consists of two steps. In the first step, we introduce *symmetric tests* (as later defined in Definition 4.2), which do not depend on the order of the observations. Then, we show that among all symmetric tests, (7) is optimal. The key is to reduce the original composite hypothesis testing problem into a simple one through the ordering map  $\Pi(x^n)$  in Definition 4.1, and then apply Neyman-Pearson lemma.

In the second step, we prove that for any test  $\psi$ , one can always *symmetrize* it and construct a symmetric one  $\phi$  which is as good as  $\psi$ , so (7) is optimal among all tests. However,  $\psi$  is constructed by assigning values on each equivalence classes introduced by the ordering map  $\Pi(\cdot)$ , so the measurability of  $\psi$  need to be carefully examined. For the detailed proof, please refer to Section IV. ■

Our second result specifies the exponent of type-II error in Neyman-Pearson formulation, which does not depend on the type-I error probability  $\epsilon$ :

**Theorem 3.2** (Asymptotic Behavior). *Let us consider the case  $|\mathcal{X}| < \infty$ , The exponent of type-II error probability is characterized as follows.*

$$E^*(\epsilon, \alpha) = \min_{U \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{1;k}) \quad (8)$$

subject to  $\alpha^\top U = \alpha^\top P_0$ .

**Remark 3.4.** *A standard way to derive the exponent of type-II error probability is to identify the acceptance region (of  $\mathcal{H}_0$ ) of the optimal test (7) as an large-deviation event under  $\mathcal{H}_1$ , and further apply a strong converse lemma to obtain a bound. However, notice that the mixture measure,  $\sum_{\sigma} \mathbb{P}_{\theta;\sigma}$ ,  $\theta \in \{0, 1\}$ , cannot be factorized into a product form, which makes it hard to single-letterize. Instead, if we add an additional assumption that  $\mathcal{X}$  is finite, then we can utilize method of types, such as Sanov's theorem, to circumvent the difficulties.*

*Sketch of proof:* For the achievability part, we propose a sub-optimal test based on Hoeffding's result [8], in which we accept observations  $x^n$  satisfying  $D(\Pi_{x^n} \| M_0(\alpha)) \leq \epsilon$  for some threshold  $\epsilon$ . We apply tools in method of types to bound the type-I and type-II error probabilities, showing that (8) is achievable.

For the converse part, given an arbitrary test, we define its acceptance region as  $\mathcal{A}$  (if the given test is randomized, we can round the test by 1/2 and make it deterministic, that is, we accept  $\mathcal{H}_1$  if  $\phi(x^n) > 1/2$ ) and consider another high-probability set  $\mathcal{B}$ . We analyze the probability of  $\mathbb{P}_{1;\sigma} \{\mathcal{A} \cap \mathcal{B}\}$ , and show that the exponent cannot be greater than (8), which concludes the converse part. For the detailed proof, please refer to Section V. ■

Finally, we give a structural result of the error exponent.

**Proposition 3.1.** *For the case  $|\mathcal{X}| < \infty$ , the type-II error exponent  $E^*(\epsilon, \alpha)$  as characterized in Theorem 3.2 only depends on  $\alpha$ . Moreover, it is a convex function of  $\alpha$ .*

*Proof:* See Appendix A. ■

### B. Numerical Evaluations

To quantify the price of anonymity, note that when the sensors are not anonymous (termed the ‘‘informed’’ setting), it becomes a simple hypothesis testing problem, and the error exponent of the type-II probability of error in the Neyman-Pearson setting is straightforward to derive:

$$E_{\text{Informed}}^*(\epsilon, \boldsymbol{\alpha}) = \sum_{k=1}^K \alpha_k D(P_{0;k} \| P_{1;k}).$$

For ease of illustration, in the following we restrict to the special case of binary alphabet, that is,  $|\mathcal{X}| = 2$ , and  $K = 2$  groups. Let  $P_{\theta;1} = \text{Ber}(p_\theta)$  and  $P_{\theta;2} = \text{Ber}(q_\theta)$ , for  $\theta = 0, 1$ , where  $\text{Ber}(p)$  is the Bernoulli distribution with parameter  $p$ . Since there are only two groups, we set  $\boldsymbol{\alpha} \equiv [1 - \alpha \quad \alpha]^\top$ . Numerical examples are given in Figure 1 to illustrate the price of anonymity versus the mixing parameter  $\alpha$ . In general, anonymity may cause significant performance loss. In certain regimes, the type-II error exponent can even be pushed to zero.

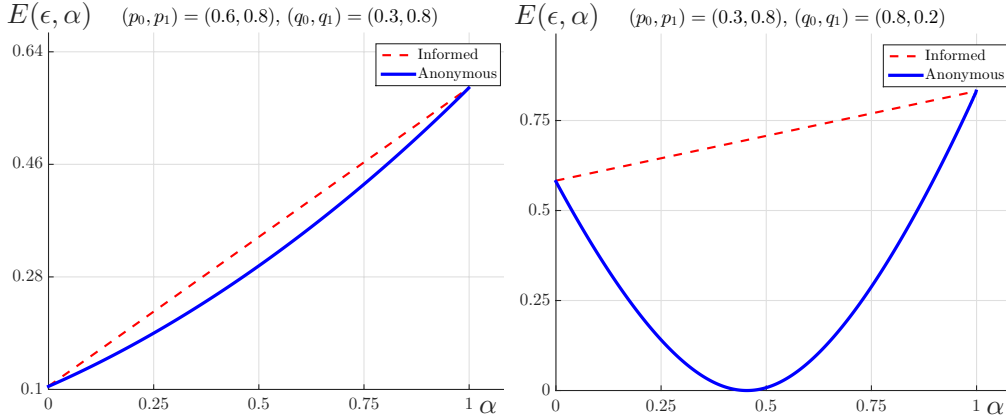


Fig. 1: Price of anonymity

### C. Distributed Detection with Byzantine Attacks

Let us apply the results to distributed detection under Byzantine attacks, where the sensors are partitioned into two groups. One group consists of  $n(1 - \alpha)$  *honest* sensors reporting true i.i.d. observations, while the other consists of  $n\alpha$  *Byzantine* sensors reporting fake i.i.d. observations. Here we again neglect the local decision function and assume that each sensor can report its observation to the fusion center. The true observations follow  $P_\theta$  i.i.d. across honest sensors, while the compromised ones follow  $Q_\theta$  i.i.d. across Byzantine sensors, for  $\theta = 0, 1$ . In general,  $Q_\theta$  is unknown to the fusion center, but in terms of error exponent, one can find the least favorable pair  $Q_0, Q_1$  which minimize the error exponent. Hence, our results can be applied here and arrive the worst-case type-II error exponent as follows:

$$\begin{aligned} & \min_{Q_0, Q_1, U, V \in \mathcal{P}_X} (1 - \alpha)D(U \| P_1) + \alpha D(V \| Q_1) \\ & \text{subject to } (1 - \alpha)U + \alpha V = (1 - \alpha)P_0 + \alpha Q_0. \end{aligned} \tag{9}$$

In [3], it assumes that each sensor can be compromised with probability  $\alpha$ , and hence it becomes a homogeneous distributed detection problem, where the observation of each sensor follows a mixture distribution  $(1 - \alpha)P_\theta + \alpha Q_\theta$  under hypothesis  $\theta$ , i.i.d. across all sensors. The worst-case exponent of type-II error probability, as derived in [3], is hence

$$\min_{Q_0, Q_1 \in \mathcal{P}_X} D((1 - \alpha)P_0 + \alpha Q_0 \parallel (1 - \alpha)P_1 + \alpha Q_1). \quad (10)$$

We see that the achievable type-II error exponent (9) in our setting is always greater than that in the i.i.d. scenario (10) (and is *strictly* larger for some  $\alpha$ ) due to the convexity of KL divergence. This implies the i.i.d. mixture model [3] might be too pessimistic. Figure 2 shows a numerical evaluation.

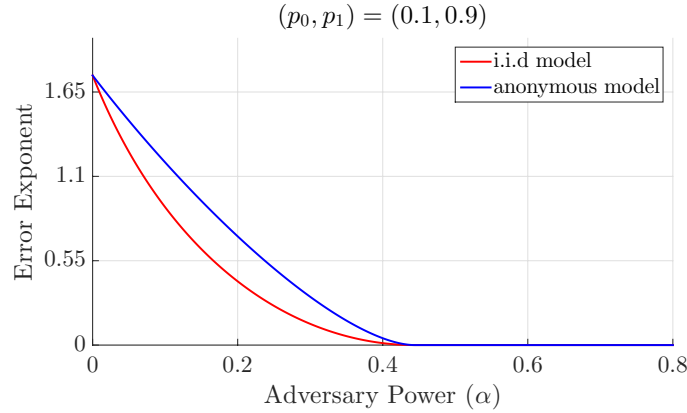


Fig. 2: Comparison between i.i.d. and our setting

#### IV. PROOF OF THEOREM 3.1

Before proving Theorem 3.1, let us introduce some definitions that help the exposition.

**Definition 4.1** (Ordering Map). *The ordering map  $\Pi(\cdot) : (\mathcal{X}^n, \mathcal{F}^{\otimes n}) \rightarrow (\tilde{\mathcal{X}}^n, \tilde{\mathcal{F}})$ , where  $\tilde{\mathcal{X}}^n$  is from (4) and  $\tilde{\mathcal{F}} \triangleq \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n$ , is defined as follows:*

$$\Pi(x^n) \triangleq (x_{i_1}, x_{i_2}, \dots, x_{i_n}), \text{ such that } x_{i_1} \geq x_{i_2} \geq \dots \geq x_{i_n}.$$

*The measurability of  $\Pi$  is easy to check.*

**Remark 4.1.** *If  $|\mathcal{X}| < \infty$ , the mapping  $\Pi$  maps a sample  $x^n$  to its type, and the space  $\tilde{\mathcal{X}}^n$  is equivalent to  $\mathcal{P}_X$ .*

**Remark 4.2.** *We will use  $\Pi^{-1}$  to denote the pre-image of  $\Pi$ . That is, for all  $\tilde{\mathcal{E}} \subseteq \tilde{\mathcal{X}}^n$ ,*

$$\Pi^{-1}(\tilde{\mathcal{E}}) \triangleq \{x^n \in \mathcal{X}^n \mid \Pi(x^n) \in \tilde{\mathcal{E}}\}.$$

*Notice that the measurability of  $\Pi$  implies for any  $\tilde{\mathcal{E}} \in \tilde{\mathcal{F}}$ , we have  $\Pi^{-1}(\tilde{\mathcal{E}}) \in \mathcal{F}^{\otimes n}$ .*

**Definition 4.2** (Symmetric Test). *We say a test  $\phi(x^n)$  is symmetric, if it is  $\sigma(\Pi(X^n))$ -measurable, that is, it can be represented as a composition*

$$\phi(x^n) = \tilde{\phi} \circ \Pi(x^n),$$

for some measurable function  $\tilde{\phi} : \tilde{\mathcal{X}}^n \rightarrow [0, 1]$ . This implies the test  $\phi$  maps a sequence of observations  $x^n$  and all its permutations to the same value.

**Lemma 4.1.** *Among all symmetric test,  $\phi^*(x^n)$ , as defined in (7), is optimal.*

*proof of Lemma 4.1:* To show the optimality of  $\phi^*$ , we first transform the original composite hypothesis testing problem to another one in the auxiliary space  $\tilde{\mathcal{X}}^n$  through the ordering mapping  $\Pi(\cdot)$ , which turns out to be a simple hypothesis testing problem. Hence, applying Neyman-Pearson lemma, we obtain the optimal test. See Figure 3 for illustration of the relation between the original space and the auxiliary space.

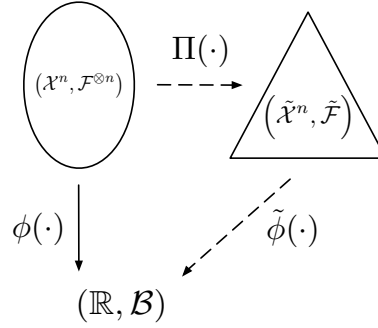


Fig. 3: Illustration of the auxiliary space

*Part 1.* First, we claim that for all  $\sigma \in \mathcal{S}_{n,\nu}$ , the probability measure  $\mathbb{P}_{0;\sigma} \circ \Pi^{-1}$ , defined on  $(\tilde{\mathcal{X}}^n, \tilde{\mathcal{F}})$ , does not depend on  $\sigma$  anymore. Thus we can define the probability measure  $\tilde{\mathbb{P}}_0 \triangleq \mathbb{P}_{0;\sigma} \circ \Pi^{-1}$ , such that for all  $\sigma$ ,

$$(\mathbb{P}_{0;\sigma}, \mathcal{F}^{\otimes n}, \mathcal{X}^n) \xrightarrow{\Pi(\cdot)} (\tilde{\mathbb{P}}_0, \tilde{\mathcal{F}}, \tilde{\mathcal{X}}^n).$$

This claim is quite intuitive, since the labeling  $\sigma$  corresponds to the order of observations, and the ordering map removes the order.

To show this claim, we first observe that for all  $\mathcal{E} \in \tilde{\mathcal{F}}$ , its pre-image

$$\Pi^{-1}(\mathcal{E}) = \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau, \quad (11)$$

where  $\mathcal{E}_\tau \triangleq \{(x_{\tau(1)}, \dots, x_{\tau(n)}) \mid (x_1, \dots, x_n) \in \mathcal{E}\}$ . Therefore, for any two  $\sigma, \sigma' \in \mathcal{S}_{n,\nu}$ , we can write  $\sigma' = \pi \circ \sigma$  for some  $\pi \in \mathcal{S}_n$ , and thus have

$$\begin{aligned} \mathbb{P}_{0;\sigma} \circ \Pi^{-1} \{\mathcal{E}\} &= \mathbb{P}_{0;\sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \stackrel{(a)}{=} \mathbb{P}_{0;\sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_{\tau \circ \pi} \right\} \\ &= \mathbb{P}_{0;\pi \circ \sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} = \mathbb{P}_{0;\sigma'} \circ \Pi^{-1} \{\mathcal{E}\}, \end{aligned}$$

where the equality (a) holds due to the following fact:

$$\forall \pi \in \mathcal{S}_n, \mathcal{S}_n \circ \pi \triangleq \{\tau \circ \pi \mid \tau \in \mathcal{S}_n\} = \mathcal{S}_n.$$

Following the same argument,  $\tilde{\mathbb{P}}_1 \triangleq \mathbb{P}_{1;\sigma} \circ \Pi^{-1}$  does not depend on  $\sigma$  either.

*Part 2.* Second, let us we consider an auxiliary hypothesis testing problem on  $\tilde{\mathcal{X}}^n$ :

$$\begin{cases} \tilde{\mathcal{H}}_0 : Z \sim \tilde{\mathbb{P}}_0 \\ \tilde{\mathcal{H}}_1 : Z \sim \tilde{\mathbb{P}}_1, \end{cases} \quad (12)$$

and let  $\tilde{\phi} : \tilde{\mathcal{X}}^n \rightarrow [0, 1]$  be a test with type-I and type-II error probabilities as follows:

$$\begin{cases} P_F(\tilde{\phi}) \triangleq \mathbb{E}_{\tilde{\mathbb{P}}_0} [\tilde{\phi}(Z)] \\ P_M(\tilde{\phi}) \triangleq \mathbb{E}_{\tilde{\mathbb{P}}_1} [1 - \tilde{\phi}(Z)]. \end{cases}$$

We claim that for any symmetric test  $\phi(x^n) = \tilde{\phi}(\Pi(x^n))$  as defined in Definition 4.2, the following holds:

$$\begin{cases} P_F(\tilde{\phi}) = P_F(\phi) \\ P_M(\tilde{\phi}) = P_M(\phi). \end{cases}$$

To show this, note that a direct calculation gives

$$\begin{aligned} P_F(\phi) &= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\phi(X^n)] \\ &= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\tilde{\phi}(\Pi(X^n))] \\ &= \max_{\sigma} \int \tilde{\phi}(\Pi(x^n)) \mathbb{P}_{0;\sigma}(dx^n) \\ &= \max_{\sigma} \int \tilde{\phi}(z) \mathbb{P}_{0;\sigma}(\Pi^{-1}(dz)) \\ &= \mathbb{E}_{\tilde{\mathbb{P}}_0} [\tilde{\phi}(Z)] = P_F(\tilde{\phi}). \end{aligned}$$

For the same reason,  $P_M(\phi) = P_M(\tilde{\phi})$ . Therefore, for any symmetric test on  $\mathcal{X}^n$ , the corresponding  $\tilde{\phi}$  has exactly the same type-I and type-II error probability. Notice that the auxiliary hypothesis testing problem (12) is simple, so by Neyman-Pearson lemma, we have readily seen that the optimal symmetric test on the original problem should be

$$\phi^*(x^n) = \begin{cases} 1, & \text{if } \ell'(x^n) > \tau \\ \gamma, & \text{if } \ell'(x^n) = \tau \\ 0, & \text{if } \ell'(x^n) < \tau, \end{cases}$$

where  $\ell'(x^n)$  is defined as

$$\ell'(x^n) = \frac{\tilde{\mathbb{P}}_1(\Pi(x^n))}{\tilde{\mathbb{P}}_0(\Pi(x^n))} = \frac{\mathbb{P}_{1;\sigma} \{\Pi^{-1}(\Pi(x^n))\}}{\mathbb{P}_{0;\sigma} \{\Pi^{-1}(\Pi(x^n))\}}.$$

*Part 3.* Finally, we show that  $\ell'(x^n)$  is indeed the mixture likelihood ratio  $\ell(x^n)$ , as defined in (6). With a slight abuse of notation, let  $\Pi_{x^n} \triangleq \Pi^{-1}(\Pi(x^n)) = \{x_{\tau(1)}, \dots, x_{\tau(n)} \mid \tau \in \mathcal{S}_n\}$ . In words,  $\Pi_{x^n}$  is the collection of  $x^n$  and all its permutations. We observe that

$$\begin{aligned} \mathbb{P}_{1;\sigma} \{\Pi^{-1}(\Pi(x^n))\} &= \sum_{y^n \in \Pi_{x^n}} \mathbb{P}_{1;\sigma}(y^n) \\ &\stackrel{(a)}{=} \left( \sum_{\tau \in \mathcal{S}_n} \mathbb{P}_{1;\sigma}(\tau(x^n)) \right) c_1(x^n) \end{aligned}$$

$$\stackrel{(b)}{=} \left( \sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma).$$

The constant  $c_1(x^n)$  in (a) is due to the fact that  $x^n = (x_1, \dots, x_n)$  might not be all distinct, so summing over the set  $\{\tau(x^n) \mid \tau \in \mathcal{S}_n\}$  may count an element  $y^n \in \Pi_{x^n}$  multiple times. Note that if  $x^n$  are all distinct, then  $c_1(x^n) = 1$ . (b) holds because  $\mathbb{P}_{1;\sigma}(\tau(x^n)) = \mathbb{P}_{1;\sigma \circ \tau}(x^n)$  and  $\mathcal{S}_{n,\nu} = \mathcal{S}_{n,\nu} \circ \tau \triangleq \{\sigma \circ \tau \mid \sigma \in \mathcal{S}_{n,\nu}\}$ . Again, the summation counts  $\sigma$  repeatedly, so we normalize by the constant  $c_2(\sigma)$ . Following the same reason,

$$\mathbb{P}_{0;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \} = \left( \sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma).$$

Hence,

$$\begin{aligned} \ell'(x^n) &= \frac{\mathbb{P}_{1;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}}{\mathbb{P}_{0;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}} \\ &= \frac{\left( \sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma)}{\left( \sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma)} \\ &= \frac{\sum_{\sigma} \mathbb{P}_{1;\sigma}(x^n)}{\sum_{\sigma} \mathbb{P}_{0;\sigma}(x^n)} = \ell(x^n), \end{aligned}$$

which establishes the claim. ■

**Lemma 4.2.** *For any general (measurable) test  $\psi(x^n) : \mathcal{X}^n \rightarrow [0, 1]$ , there exists a symmetric test  $\phi(x^n)$  whose performance is not worse than  $\psi$ . That is,*

$$\begin{cases} \mathbb{P}_F(\phi) \leq \mathbb{P}_F(\psi) \\ \mathbb{P}_M(\phi) \leq \mathbb{P}_M(\psi). \end{cases} \quad (13)$$

*proof of Lemma 4.2:* With a slight abuse of notation, let  $\tau(x^n)$  denote the coordinate-permutation function with respect to  $\tau \in \mathcal{S}_n$ , i.e.  $\tau(x^n) = (x_{\tau(1)}, \dots, x_{\tau(n)})$ . Then we construct  $\phi(x^n)$  as follows:

$$\phi(x^n) \triangleq \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(x^n).$$

We claim the following two facts:

- 1)  $\phi(x^n)$  is symmetric, and thus can be written as  $\tilde{\phi} \circ \Pi(x^n)$  for some  $\tilde{\mathcal{F}}$ -measurable  $\tilde{\phi}$ .
- 2) (13) holds for the constructed  $\phi$ .

*Part 1.* To see that  $\phi(x^n) = \tilde{\phi} \circ \Pi(x^n)$ , we observe that for any  $y^n, z^n \in \Pi^{-1}(\tilde{x}^n)$ , there exists a permutation  $\pi \in \mathcal{S}_n$  such that  $y^n = \pi(z^n)$ . Hence it suffices to verify that for all  $\pi \in \mathcal{S}_n$ ,  $\phi(x^n) = \phi(\pi(x^n))$ .

$$\begin{aligned} \phi(\pi(x^n)) &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(\pi(x^n)) \\ &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau \circ \pi(x^n) \\ &\stackrel{(a)}{=} \frac{1}{n!} \sum_{\tau' \in \mathcal{S}_n} \psi \circ \tau'(x^n) = \phi(x^n). \end{aligned}$$

The equality (a) holds due to the fact that

$$\mathcal{S}_n \circ \pi \triangleq \{\tau \circ \pi \mid \tau \in \mathcal{S}_n\} = \mathcal{S}_n.$$

Therefore,  $\phi(x^n)$  can be decomposed into  $\tilde{\phi} \circ \Pi(x^n)$ .

Next, we check the measurability of  $\tilde{\phi}$ . Notice that  $\phi$  is  $\mathcal{F}^{\otimes}$ -measurable, since both  $\psi$  and  $\tau$  are measurable. The measurability of  $\tau$  follows from the  $\tau$ -permuted closedness assumption of  $\mathcal{F}^{\otimes n}$ :

$$\forall \mathcal{A} \in \mathcal{F}^{\otimes n}, \mathcal{A}_\tau \triangleq \{\tau(x^n) \mid x^n \in \mathcal{A}\} \in \mathcal{F}^{\otimes n}.$$

Observe that for all Borel-measurable set  $\mathcal{B}$ , we have

$$\phi^{-1}\{\mathcal{B}\} = \Pi^{-1}\{\tilde{\phi}^{-1}\{\mathcal{B}\}\} \in \mathcal{F}^{\otimes n} \Leftrightarrow \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n},$$

where we use  $\mathcal{E}$  to denote event  $\tilde{\phi}^{-1}\{\mathcal{B}\}$ , and  $\mathcal{E}_\tau$  to denote the  $\tau$ -permuted event of  $\mathcal{E}$ , as defined in (5). Notice here we use the fact given by (11). Therefore it suffices to check

$$\forall \mathcal{E} \subseteq \tilde{\mathcal{X}}^n, \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n} \Rightarrow \mathcal{E} \in \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n = \tilde{\mathcal{F}}.$$

We claim that indeed,

$$\left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n = \mathcal{E},$$

for every  $\mathcal{E} \subseteq \tilde{\mathcal{X}}^n$ . This is because

- 1) Since  $\mathcal{E} \subseteq \tilde{\mathcal{X}}^n$ , we have  $\mathcal{E} = \mathcal{E} \cap \tilde{\mathcal{X}}^n \subseteq \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n$ .
- 2) For any  $\tau$  and for any  $x^n \in \mathcal{E}_\tau \cap \tilde{\mathcal{X}}^n$ ,  $x^n \in \mathcal{E}$ . Hence,  $\forall \tau \in \mathcal{S}_n$ ,  $\mathcal{E}_\tau \cap \tilde{\mathcal{X}}^n \subseteq \mathcal{E}$ , that is,  $\left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n \subseteq \mathcal{E}$ .

Hence,

$$\bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n} \Rightarrow \mathcal{E} = \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n \in \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n = \tilde{\mathcal{F}},$$

showing that  $\tilde{\phi}$  is  $\tilde{\mathcal{F}}$ -measurable.

*Part 2.* We show that  $\phi(x^n)$  cannot be worse than  $\psi(x^n)$ . Observe that for all  $\tau \in \mathcal{S}_n$ , we have

$$\mathbb{P}_F(\psi \circ \tau) = \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0,\sigma}}[\psi(\tau(X^n))] = \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0,\sigma \circ \tau^{-1}}}[\psi(X^n)] = \max_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0,\sigma'}}[\psi(X^n)] = \mathbb{P}_F(\psi).$$

Again, the third equality holds due to the fact

$$\mathcal{S}_{n,\nu} \circ \tau^{-1} \triangleq \{\sigma \circ \tau^{-1} \mid \sigma \in \mathcal{S}_{n,\nu}\} = \mathcal{S}_{n,\nu}.$$

Therefore, we have

$$\begin{aligned} \mathbb{P}_F(\phi) &= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0,\sigma}} \left[ \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(X^n) \right] \\ &\leq \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0,\sigma}}[\psi \circ \tau(X^n)] \\ &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \mathbb{P}_F(\psi \circ \tau) = \mathbb{P}_F(\psi). \end{aligned}$$

Following the same argument, we obtain  $P_M(\phi) \leq P_M(\psi)$ , and the proof completes. ■

Finally, the proof of Theorem 3.1 directly follows from Lemma 4.1 and Lemma 4.2.

*Proof of Theorem 3.1:* From Lemma 4.2, we only need to consider symmetric tests. From Lemma 4.1, we see that the optimal test among all symmetric tests is the mixture likelihood test, as defined in (7). This establishes Theorem 3.1. ■

**Remark 4.3.** Notice that in the above proof, we do not make use of assumptions on the distribution of  $X^n$ , such as independence. Indeed, the proof indicates that for the anonymous composite hypothesis testing problem, under the minimax criterion (i.e. to minimize the worst case error), we should always design tests based on the empirical distribution of  $X^n$  (i.e. as a function of  $\Pi(x^n)$ ). This principle also holds for other statistical inference problems, such as  $M$ -ary hypothesis testing.

## V. PROOF OF THEOREM 3.2

For the case  $|\mathcal{X}| < \infty$ , the auxiliary space  $\tilde{\mathcal{X}}$  is equivalent to the space of all probability measures on  $\mathcal{X}$ , that is,  $\mathcal{P}_{\mathcal{X}}$ , and the mapping  $\Pi(x^n)$  maps a sequence of samples to its type  $\Pi_{x^n}$ . According to Lemma 4.2, the optimal test is symmetric, which implies that we only need to consider tests depending on the type. For tests depending only on the empirical distribution, it is natural to view their acceptance region as a collection of empirical distribution, that is, a (measurable) subset of  $\mathcal{P}_{\mathcal{X}}$ . This motivates us to apply Sanov's theorem. We begin with the following generalization of Sanov's result:

**Lemma 5.1** (Generalized Sanov Theorem). *Let  $|\mathcal{X}| < \infty$ , and  $\Gamma \subseteq \mathcal{P}_{\mathcal{X}}$  be a collection of distributions on  $\mathcal{X}$ . Then for all  $\sigma \in \mathcal{S}_{n,\nu}$  and  $\theta \in \{0, 1\}$ , we have*

$$- \inf_{\substack{[U_1 \dots U_K]^T \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^T U \in \text{int } \Gamma}} \sum_{k=1}^K \alpha_k D(U_k \| P_{\theta;k}) \quad (14)$$

$$\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} \quad (15)$$

$$\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} \quad (16)$$

$$\leq - \inf_{\substack{[U_1 \dots U_K]^T \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^T U \in \text{cl } \Gamma}} \sum_{k=1}^K \alpha_k D(U_k \| P_{\theta;k}), \quad (17)$$

where in taking the limits, we assume that  $\lim_{n \rightarrow \infty} \frac{n_k}{n} = \alpha_k$ , for all  $k = 1, \dots, K$ . In particular, if the infimum in the right-hand side is equal to the infimum in the left-hand side, then we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} = - \inf_{\substack{[U_1 \dots U_K]^T \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^T U \in \text{cl } \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta;k}).$$

The proof is a direct extension of Lemma 2.3, except that we replace the i.i.d. measure with the product of independent non-identical ones,  $\mathbb{P}_{\theta;\sigma}$ . For the detailed proof, please refer to Appendix B.



Motivated by the generalized Sanov Theorem, we further define the following generalized divergence to measure how far from one set of distributions  $\mathbf{Q} \triangleq [Q_1 \dots Q_K]^\top$  to another set of distributions  $\mathbf{P} \triangleq [P_1 \dots P_K]^\top$ :

**Definition 5.1.** Let  $\mathbf{P} = [P_1 \dots P_K]^\top$  and  $\mathbf{Q} = [Q_1 \dots Q_K]^\top$  are both in  $(\mathcal{P}_X)^K$ . Let  $\boldsymbol{\alpha} = [\alpha_1 \dots \alpha_K]^\top$  be a  $K$ -tuple probability vector. Define

$$D_{\boldsymbol{\alpha}}(\mathbf{P}; \mathbf{Q}) \triangleq \inf_{\mathbf{U} \in (\mathcal{P}_X)^K} \sum_{k=1}^K \alpha_k D(U_k \| Q_k) \quad (18)$$

subject to  $\boldsymbol{\alpha}^\top \mathbf{U} = \boldsymbol{\alpha}^\top \mathbf{P}$

Thus (14) in Lemma 5.1 can be rewritten as

$$- \inf_{\boldsymbol{\alpha}^\top \mathbf{U} \in \text{int } \Gamma} D_{\boldsymbol{\alpha}}(\mathbf{U}; \mathbf{P}_\theta) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} \leq - \inf_{\boldsymbol{\alpha}^\top \mathbf{U} \in \text{cl } \Gamma} D_{\boldsymbol{\alpha}}(\mathbf{U}; \mathbf{P}_\theta).$$

Also, the result of Theorem 3.2, (8), is equivalent to the following statement:

$$E^*(\epsilon, \boldsymbol{\alpha}) = D_{\boldsymbol{\alpha}}(\mathbf{P}_0; \mathbf{P}_1).$$

**Remark 5.1.** Intuitively,  $D_{\boldsymbol{\alpha}}(\mathbf{P}; \mathbf{Q})$  measures how far between  $\mathbf{P}$  and  $\mathbf{Q}$ . However,  $D_{\boldsymbol{\alpha}}(\cdot; \cdot)$  is not a divergence, since  $D_{\boldsymbol{\alpha}}(\mathbf{P}; \mathbf{Q}) = 0$  does not always imply  $\mathbf{P} = \mathbf{Q}$ .

Notice that for any fixed  $\mathbf{Q} \in (\mathcal{P}_X)^K$ ,  $D_{\boldsymbol{\alpha}}(\mathbf{P}; \mathbf{Q})$  can be regarded as a function of  $\mathbf{P}$ . Moreover, this function depends only on the mixture of  $\mathbf{P}$ , say,  $\boldsymbol{\alpha}^\top \mathbf{P}$ . Therefore, for notional convenience, let us use  $f_{\mathbf{Q}}(\cdot) : \mathcal{P}_X \rightarrow \mathbb{R} \cup \{+\infty\}$  to denote this function:

$$f_{\mathbf{Q}}(T) \triangleq \inf_{\mathbf{U} \in (\mathcal{P}_X)^K} \sum_{k=1}^K \alpha_k D(U_k \| Q_k)$$

subject to  $\boldsymbol{\alpha}^\top \mathbf{U} = T$

In other words,

$$f_{\mathbf{Q}}(\boldsymbol{\alpha}^\top \mathbf{P}) = D_{\boldsymbol{\alpha}}(\mathbf{P}; \mathbf{Q}).$$

Before entering the main proof of Theorem 3.2, let us introduce some properties of  $f_{\mathbf{Q}}(\cdot)$ .

**Lemma 5.2.** Let  $\mathbf{Q} \in (\mathcal{P}_X)^K$  and  $f_{\mathbf{Q}}(\cdot) : \mathcal{P}_X \rightarrow \mathbb{R} \cup \{+\infty\}$  be defined as Definition 5.1 and above. Then,

- 1)  $f_{\mathbf{Q}}(\boldsymbol{\alpha}^\top \mathbf{Q}) = 0$
- 2) The collection of all  $T \in \mathcal{P}_X$  such that  $f_{\mathbf{Q}}(T) < \infty$ , denoted as

$$\mathcal{C}_{\mathbf{Q}} \triangleq \{T \in \mathcal{P}_X : f_{\mathbf{Q}}(T) < \infty\},$$

is a compact, convex subset of  $\mathcal{P}_X$ .

- 3)  $f_{\mathbf{Q}}(T)$  is a convex, continuous function of  $T$  on  $\mathcal{C}_{\mathbf{Q}}$  (and by the compactness of  $\mathcal{C}_{\mathbf{Q}}$ ,  $f_{\mathbf{Q}}(T)$  is also uniformly continuous).

Proof of Lemma 5.2 can be found in Appendix C.

*proof of Theorem 3.2:*

*Part 1 (Achievability).* Let  $\delta > 0$  and consider the test :

$$\phi(x^n) \triangleq \mathbb{1}_{\{x^n : D(\Pi_{x^n} \| M_0(\boldsymbol{\alpha})) > \delta\}}.$$

Denote the acceptance region of  $\phi$  as  $\Gamma \triangleq \{T \in \mathcal{P}_{\mathcal{X}} : D(T \| M_0(\boldsymbol{\alpha})) > \delta\}$ . Then the exponent of type-I error probability  $P_F(\phi)$  can be bounded by

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\phi(X^n)] \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{0;\sigma} \{\Pi_{x^n} \in \Gamma\} \\ &\stackrel{(a)}{\geq} \inf_{T \in \text{cl } \Gamma} f_{\mathcal{P}_0}(T) \\ &\stackrel{(b)}{\geq} \delta, \end{aligned}$$

where (a) holds by Lemma 5.1, and (b) holds due to the convexity of KL divergence:

$$\begin{aligned} D(T \| M_0(\boldsymbol{\alpha})) &\leq \min_{\mathbf{U} \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{0;k}) = f_{\mathcal{P}_0}(T) \\ &\text{subject to } \boldsymbol{\alpha}^\top \mathbf{U} = T \end{aligned}$$

Notice that for any  $\delta > 0$ , as  $n$  large enough, we must have

$$P_F(\phi) < \epsilon.$$

On the other hand, the exponent of type-II error probability  $\underline{E}^*(\epsilon, \boldsymbol{\alpha})$  can be bounded by

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{1;\sigma}} [\phi(X^n)] \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{1;\sigma} \{X^n : D(\Pi_{X^n} \| M_0(\boldsymbol{\alpha})) \leq \delta\} \\ &\geq \inf_{T \in \text{cl } \Gamma^c} f_{\mathcal{P}_1}(T), \end{aligned} \tag{19}$$

By Pinsker's inequality (Theorem 6.5 in [7]), we have

$$\text{cl}(\Gamma^c) = \{T \in \mathcal{P}_{\mathcal{X}} : D(T \| M_0(\boldsymbol{\alpha})) \leq \delta\} \subseteq \left\{T \in \mathcal{P}_{\mathcal{X}} : \|T - M_0(\boldsymbol{\alpha})\|_1 \leq \sqrt{2\delta}\right\} \triangleq B_{\sqrt{2\delta}}(M_0(\boldsymbol{\alpha})),$$

so (19) can be further lower bounded by

$$\inf_{T \in \text{cl } \Gamma^c} f_{\mathcal{P}_1}(T) \geq \inf_{T \in B_{\sqrt{2\delta}}(M_0(\boldsymbol{\alpha}))} f_{\mathcal{P}_1}(T).$$

Also, by the continuity (Lemma 5.2) of  $f_{\mathcal{P}_1}(\cdot)$ ,

$$\inf_{T \in B_{\sqrt{2\delta}}(M_0(\boldsymbol{\alpha}))} f_{\mathcal{P}_1}(T) = f_{\mathcal{P}_1}(M_0(\boldsymbol{\alpha})) + \Delta(\delta),$$

with

$$\lim_{\delta \rightarrow 0} \Delta(\delta) = 0.$$

Finally, since  $\delta$  can be chosen arbitrarily small, we have

$$\underline{E}^*(\epsilon, \boldsymbol{\alpha}) \geq f_{\mathcal{P}_1}(M_0(\boldsymbol{\alpha})) = D_{\boldsymbol{\alpha}}(\mathcal{P}_0; \mathcal{P}_1). \tag{20}$$

*Part 2 (Converse).* We have shown that symmetric test is optimal in Lemma 4.2. Hence, in the following, it suffices to consider symmetric tests.

For an arbitrary symmetric test  $\psi : \mathcal{P}_n \rightarrow [0, 1]$  such that its type-I error probability  $P_F(\psi) < \epsilon$ , we shall lower bound its type-II error probability as follows. Let  $\mathcal{A}^{(n)} \triangleq \{T \in \mathcal{P}_n : \psi(T) \leq 1/2\}$ , and recall that

$$\tilde{\mathbb{P}}_0 \triangleq \mathbb{P}_{0;\sigma} \circ \Pi^{-1}$$

is a probability measure independent of  $\sigma$ . Then, we have

$$\begin{aligned} \epsilon > \mathbb{E}_{\tilde{\mathbb{P}}_0} [\psi(T)] &= \sum_{T \in \mathcal{P}_n} \tilde{\mathbb{P}}_0(T) \psi(T) \geq \sum_{T \in (\mathcal{A}^{(n)})^c} \tilde{\mathbb{P}}_0(T) \psi(T) \\ &\stackrel{(a)}{>} \frac{1}{2} \sum_{T \in (\mathcal{A}^{(n)})^c} \tilde{\mathbb{P}}_0(T) = \frac{1}{2} \left(1 - \tilde{\mathbb{P}}_0 \left\{ \mathcal{A}^{(n)} \right\}\right), \end{aligned}$$

(a) holds since for all  $T \notin \mathcal{A}^{(n)}$ ,  $\psi(T) > 1/2$ . In other words, we have

$$\tilde{\mathbb{P}}_0 \left\{ \mathcal{A}^{(n)} \right\} > 1 - 2\epsilon.$$

On the other hand, let  $\mathcal{B}^{(n)} \triangleq \{T \in \mathcal{P}_n \mid D(T \| M_0(\alpha)) \leq \delta\}$ . Then, according to the analysis in type-I error probability in the achievability part, we have

$$\tilde{\mathbb{P}}_0 \left\{ \mathcal{B}^{(n)} \right\} > 1 - \epsilon.$$

Applying union bound, we see that

$$\tilde{\mathbb{P}}_0 \left\{ \mathcal{A}^{(n)} \cap \mathcal{B}^{(n)} \right\} > 1 - 3\epsilon,$$

and hence for  $\epsilon < \frac{1}{3}$ ,  $\mathcal{A}^{(n)} \cap \mathcal{B}^{(n)}$  is non-empty.

Let  $V_n^* \in \mathcal{A}^{(n)} \cap \mathcal{B}^{(n)}$  and define  $\tilde{\mathbb{P}}_1 \triangleq \mathbb{P}_{1;\sigma} \circ \Pi^{-1}$  (which is also independent of  $\sigma$ ). Again we have

$$\begin{aligned} P_F(\psi) &= \mathbb{E}_{\tilde{\mathbb{P}}_1} [1 - \psi(T)] \\ &\geq \sum_{T \in \mathcal{A}^{(n)}} (1 - \psi(T)) \tilde{\mathbb{P}}_1 \{T\} \\ &\geq \frac{1}{2} \tilde{\mathbb{P}}_1 \{V_n^*\}. \end{aligned}$$

We further estimate  $\tilde{\mathbb{P}}_1 \{V_n^*\}$  by

$$\begin{aligned} \tilde{\mathbb{P}}_1 \{V_n^*\} &= \mathbb{P}_{1;\sigma} \{T_n(V_n^*)\} \\ &= \sum_{\substack{U_k \in \mathcal{P}_{n_k} : \\ \sum_k \alpha_k U_k = V_n^*}} \prod_{k=1}^K P_{1;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &= \sum_{\substack{U_k \in \mathcal{P}_{n_k} : \\ \sum_k \alpha_k U_k = V_n^*}} 2^{-\sum_k n_k D(U_k \| P_{1;k})} \\ &\geq \max_{\substack{U_k \in \mathcal{P}_{n_k} : \\ \sum_k \alpha_k U_k = V_n^*}} 2^{-\sum_k n_k D(U_k \| P_{1;k})} \\ &= 2^{-n \tilde{D}_n}, \end{aligned}$$

where

$$\tilde{D}_n \triangleq \min_{\substack{U_k \in \mathcal{P}_{n_k}: \\ \sum_k \alpha_k U_k = V_n^*}} \left( \sum_k \frac{n_k}{n} D(U_k \| P_{1;k}) \right).$$

Notice that since  $V_n^* \in \mathcal{B}^{(n)}$ , so we have

$$D(V_n^* \| M_0(\boldsymbol{\alpha})) \leq \delta.$$

Since  $\delta$  can be chosen arbitrarily small, as  $\delta \rightarrow 0$  and  $n \rightarrow \infty$  (with  $\frac{n_k}{n} \rightarrow \alpha_k$ ), we have

$$\begin{aligned} \bar{E}^*(\epsilon, \boldsymbol{\alpha}) &\leq \lim_{n \rightarrow \infty} \tilde{D}_n \\ &= \min_{\substack{U_k \in \mathcal{P}_{\mathcal{X}}: \\ \sum_k \alpha_k U_k = M_0(\boldsymbol{\alpha})}} \left( \sum_k \alpha_k D(U_k \| P_{1;k}) \right) \\ &= f_{\mathcal{P}_1}(M_0(\boldsymbol{\alpha})) \\ &= D_{\boldsymbol{\alpha}}(\mathbf{P}_0; \mathbf{P}_1), \end{aligned}$$

which completes the proof. ■

## VI. A GEOMETRICAL PERSPECTIVE IN CHERNOFF'S REGIME

So far, for asymptotic regime, we have been focusing on Neyman-Pearson's formulation, in which we minimize the worst-case type-II error probability, subject to the worst-case type-I error probability not being larger than a constant  $\epsilon$ . It is natural to extend the result from Section III to Chernoff's regime, where we aim to minimize the average probability of error:

$$P_e^{(n)}(\phi) \triangleq \pi_0 P_F^{(n)} + \pi_1 P_M^{(n)}.$$

Note that  $\pi_0$  and  $\pi_1$  are the prior distributions of  $\mathcal{H}_0$  and  $\mathcal{H}_1$  and do not scale with  $n$ . As suggested by Theorem 3.1, the optimal test is the mixture likelihood ratio test, so we only need to specify the corresponding threshold  $\tau$ . However, the mixture likelihood ratio involves summation over  $\mathcal{S}_{n,\nu}$ , making the computation complexity extremely high. Even for the case  $|\mathcal{X}| < \infty$ , the computation still takes  $\Theta(n^{|\mathcal{X}|})$  operations and thus is difficult to implement. To break the computational barrier, we propose an asymptotically optimal test, based on information projection, which achieves the optimal exponent of the average probability of error. Moreover, the result can be generalized to determine the achievable exponent region  $\mathcal{R}$ , the collection of all achievable pairs of exponents:

$$\mathcal{R} \triangleq \left\{ (E_0, E_1) \mid \text{there exists a test } \phi, \text{ such that } P_F^{(n)}(\phi) \preceq 2^{-nE_0}, P_M^{(n)}(\phi) \preceq 2^{-nE_1} \right\},$$

where a sequence  $a_n \preceq 2^{-nE_0}$  means  $a_n$  decays to zero at the rate faster than  $E_0$ , that is,

$$-\liminf_{n \rightarrow \infty} \frac{1}{n} \log a_n \geq E_0.$$

### A. Asymptotically Optimal Test in Chernoff's Regime

**Theorem 6.1** (Efficient Test). Recall the function  $f_{\mathbf{P}}(T) : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R} \cup \{+\infty\}$  defined in Definition 5.1. Consider the following test based on the function  $f_{\mathbf{P}_0}(\cdot)$  and  $f_{\mathbf{P}_1}(\cdot)$ :

$$\phi_{\text{eff}}(x^n) \triangleq \begin{cases} 0, & \text{if } f_{\mathbf{P}_1}(\Pi_{x^n}) > f_{\mathbf{P}_0}(\Pi_{x^n}) \\ 1, & \text{else } f_{\mathbf{P}_1}(\Pi_{x^n}) \leq f_{\mathbf{P}_0}(\Pi_{x^n}). \end{cases} \quad (21)$$

Then  $\phi_{\text{eff}}$  is asymptotically optimal in Chernoff's regime. That is, for all priors  $\pi_0, \pi_1$ , for all tests  $\phi$ , and for all  $n$  large enough,

$$-\frac{1}{n} \log(P_e(\phi)) \leq -\frac{1}{n} \log(P_e(\phi_{\text{eff}})).$$

**Remark 6.1.** From the convexity of KL-divergence and the space  $\mathcal{P}_{\mathcal{X}}$ , the function  $f_{\mathbf{P}}(\cdot)$  is indeed the minimization of a convex function. Hence the proposed test in Theorem 6.1 can be computed efficiently.

*Proof:* Let us set some notations. For each  $\mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K$ , we use  $B_r(\mathbf{P}) \subseteq \mathcal{P}_{\mathcal{X}}$  to denote the  $r$ -ball centered at  $T$  with respect to  $f_{\mathbf{P}}(\cdot)$ :

$$B_r(\mathbf{P}) \triangleq \{T \in \mathcal{P}_{\mathcal{X}} \mid f_{\mathbf{P}}(T) < r\}.$$

By the continuity of  $f_{\mathbf{P}}(\cdot)$  (from Lemma 5.2),  $B_r(\mathbf{P})$  is an open set. Then, define the largest packing radius between  $\mathbf{P}_0, \mathbf{P}_1$  as follows:

$$r^* \triangleq \sup_r \{B_r(\mathbf{P}_0) \cap B_r(\mathbf{P}_1) = \emptyset\}.$$

See Figure 4 for illustration.

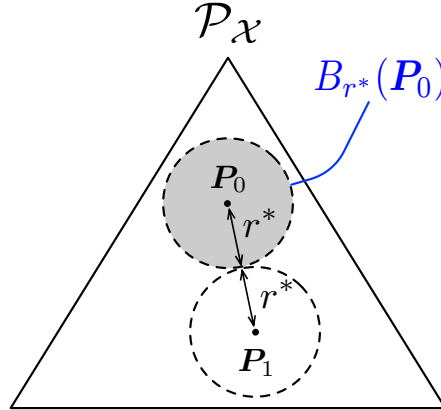


Fig. 4: Illustration of  $B_r(\cdot)$  and  $r^*$

The rest of the proof will be organized as follows: we first show that  $\phi_{\text{eff}}$  has error exponent at least  $r^*$  (the achievability part):

$$-\lim_{n \rightarrow \infty} \frac{1}{n} \log(P_e(\phi_{\text{eff}})) \geq r^*.$$

Then, we will prove that for all tests, the error exponent will be at most  $r^*$  (the converse part).

Part 1 (Achievability). Define

$$\mathcal{A} \triangleq \{T \in \mathcal{P}_{\mathcal{X}} \mid f_{\mathbf{P}_1}(T) \leq f_{\mathbf{P}_0}(T)\},$$

and notice that

$$\begin{cases} \mathbb{P}_{\mathbf{F}}^{(n)}(\phi_{\text{eff}}) = \mathbb{P}_{0;\sigma} \{\Pi_{x^n} \in \mathcal{A}\} \\ \mathbb{P}_{\mathbf{M}}^{(n)}(\phi_{\text{eff}}) = \mathbb{P}_{1;\sigma} \{\Pi_{x^n} \in \mathcal{A}^c\}, \end{cases}$$

for any arbitrary  $\sigma$  (recall that  $\phi_{\text{eff}}$  depends only on the empirical distribution and therefore is symmetrical, so the error is independent of the choice of a specific  $\sigma$ ).

By the generalized Sanov's theorem (Lemma 5.1), we see that the exponent of  $\mathbb{P}_{\mathbf{F}}^{(n)}(\phi_{\text{eff}})$  is lower bounded by  $\inf_{T \in \text{cl } \mathcal{A}} f_{\mathbf{P}_0}(T)$ . Similarly, the exponent of  $\mathbb{P}_{\mathbf{M}}^{(n)}(\phi_{\text{eff}})$  is lower bounded by  $\inf_{T \in \text{cl } \mathcal{A}^c} f_{\mathbf{P}_1}(T)$ . It is not hard to see that indeed,

$$\inf_{T \in \text{cl } \mathcal{A}} f_{\mathbf{P}_0}(T) = \inf_{T \in \mathcal{A}} f_{\mathbf{P}_0}(T), \quad (22)$$

and

$$\inf_{T \in \text{cl } \mathcal{A}^c} f_{\mathbf{P}_1}(T) = \inf_{T \in \mathcal{A}^c} f_{\mathbf{P}_1}(T). \quad (23)$$

Equation (22) holds since  $\mathcal{A}$  is a closed set (it is a pre-image of a continuous function from a closed set), so  $\text{cl } \mathcal{A} = \mathcal{A}$ . For the equation (23), we notice that  $\mathcal{A}^c$  is open, and hence the infimum of a continuous function on  $\mathcal{A}^c$  is actually equal to the infimum on  $\text{cl } \mathcal{A}^c$ .

Hence, it suffices to show that

$$\inf_{T \in \mathcal{A}} f_{\mathbf{P}_0}(T) \geq r^*, \quad \inf_{T \in \mathcal{A}^c} f_{\mathbf{P}_1}(T) \geq r^*.$$

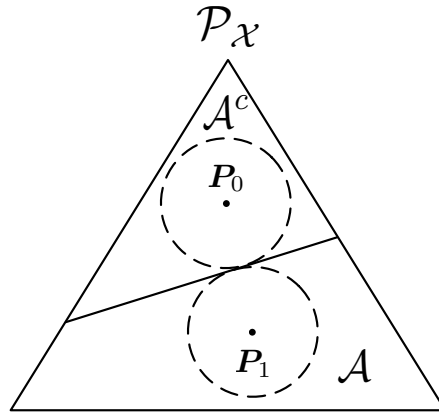


Fig. 5: Relation between  $\mathcal{A}$ ,  $\mathcal{A}^c$  and  $B_{r^*}(\mathbf{P}_0)$ ,  $B_{r^*}(\mathbf{P}_1)$

It is straightforward to see that  $\mathcal{A}^c$  contains  $B_{r^*}(\mathbf{P}_0)$  and  $\mathcal{A}$  contains  $B_{r^*}(\mathbf{P}_1)$ , since we must have

- 1)  $\forall T \in B_{r^*}(\mathbf{P}_0), f_{\mathbf{P}_0}(T) < f_{\mathbf{P}_1}(T),$
- 2)  $\forall T \in B_{r^*}(\mathbf{P}_1), f_{\mathbf{P}_0}(T) > f_{\mathbf{P}_1}(T).$

Otherwise  $B_{r^*}(\mathbf{P}_0)$  intersects  $B_{r^*}(\mathbf{P}_1)$ , violating our assumption on  $r^*$ . Also notice that  $\mathcal{A}, \mathcal{A}^c$  are disjoint, so

$$\mathcal{A}^c \cap B_{r^*}(\mathbf{P}_0) = \mathcal{A} \cap B_{r^*}(\mathbf{P}_1) = \emptyset,$$

implying that

$$\mathcal{A}^c \subseteq B_{r^*}(\mathbf{P}_1)^c, \mathcal{A} \subseteq B_{r^*}(\mathbf{P}_0)^c.$$

Therefore, we have

$$\begin{cases} \inf_{T \in \mathcal{A}} f_{\mathbf{P}_0}(T) \geq \inf_{T \in B_{r^*}(\mathbf{P}_0)^c} f_{\mathbf{P}_0}(T) \geq r^* \\ \inf_{T \in \mathcal{A}^c} f_{\mathbf{P}_1}(T) \geq \inf_{T \in B_{r^*}(\mathbf{P}_1)^c} f_{\mathbf{P}_1}(T) \geq r^*, \end{cases}$$

proving the achievability part.

*Part 2 (Converse).* We show that for any test  $\phi^{(n)}$ , the exponent of the average probability of error greater than  $r^*$  leads to contradiction. Suppose the type-I and type-II error exponents of  $\phi^{(n)}$  are  $r_1, r_2$  respectively, and  $r_1 > r^*, r_2 > r^*$ . By Lemma 4.2, we only need to consider symmetric tests, that is, tests depend only on the type. Therefore, we can write the acceptance region of  $\mathcal{H}_0, \mathcal{H}_1$  as

$$\begin{cases} \mathcal{B}_1^{(n)} = \{\Pi_{x^n} : \phi^{(n)}(x^n) = 1\} \\ \mathcal{B}_0^{(n)} = \{\Pi_{x^n} : \phi^{(n)}(x^n) = 0\}. \end{cases}$$

The exponents of type-I and type-II errors thus are greater than  $r_1, r_2$  respectively, we have

$$\begin{cases} \liminf_{n \rightarrow \infty} \left\{ \min_{T \in \mathcal{B}_1^{(n)}} f_{\mathbf{P}_0}(T) \right\} = r_1 > r^* \\ \liminf_{n \rightarrow \infty} \left\{ \min_{T \in \mathcal{B}_0^{(n)}} f_{\mathbf{P}_1}(T) \right\} = r_2 > r^*. \end{cases} \quad (24)$$

Define  $\min\{r_1, r_2\} = \tilde{r}$ , and  $\delta \triangleq (\tilde{r} - r^*)/2 > 0$ . By (24), there exists  $M$  large enough, such that for all  $n > M$ ,

$$\begin{cases} \min_{T \in \mathcal{B}_1^{(n)}} f_{\mathbf{P}_0}(T) > \tilde{r} - \delta > r^* \\ \min_{T \in \mathcal{B}_0^{(n)}} f_{\mathbf{P}_1}(T) > \tilde{r} - \delta > r^*. \end{cases}$$

We further define

$$\begin{cases} \mathcal{B}_1 = \bigcup_{n > M} \mathcal{B}_1^{(n)} \\ \mathcal{B}_0 = \bigcup_{n > M} \mathcal{B}_0^{(n)}. \end{cases}$$

We see that

1)  $\mathcal{B}_0 \cup \mathcal{B}_1$  are dense in  $\mathcal{P}_{\mathcal{X}}$ , since

$$\mathcal{B}_0^{(n)} \cup \mathcal{B}_1^{(n)} = \mathcal{P}_n,$$

and  $\bigcup_{n > M} \mathcal{P}_n$  is dense in  $\mathcal{P}_{\mathcal{X}}$ . So we have

$$(\text{cl } \mathcal{B}_0 \cup \text{cl } \mathcal{B}_1)^c = (\text{cl } \mathcal{B}_0)^c \cap (\text{cl } \mathcal{B}_1)^c = \emptyset. \quad (25)$$

2) By construction,

$$\begin{cases} \inf_{T \in \mathcal{B}_1} f_{\mathbf{P}_0}(T) = \min_{T \in \text{cl } \mathcal{B}_1} f_{\mathbf{P}_0}(T) > \tilde{r} - \delta > r^* \\ \inf_{T \in \mathcal{B}_0} f_{\mathbf{P}_1}(T) = \min_{T \in \text{cl } \mathcal{B}_0} f_{\mathbf{P}_1}(T) > \tilde{r} - \delta > r^*. \end{cases} \quad (26)$$

From (26), we have

$$\begin{cases} B_{(\tilde{r}-\delta)}(\mathbf{P}_0) \subseteq (\text{cl } \mathcal{B}_1)^c \\ B_{(\tilde{r}-\delta)}(\mathbf{P}_1) \subseteq (\text{cl } \mathcal{B}_0)^c, \end{cases}$$

and by (25)  $B_{(\tilde{r}-\delta)}(\mathbf{P}_0) \cap B_{(\tilde{r}-\delta)}(\mathbf{P}_1) = \emptyset$ . However, this violates our assumption that  $r^*$  is the supreme of radius such that the two sets do not overlap. This proves the converse part. ■

**Remark 6.2.** In Theorem 6.1, we provide an asymptotically optimal test based on an information-geometric perspective. However, we do not specify the exact error exponent. As stated in the proof, the optimal exponent of average probability of error can be obtain by solving the information projection problem:

$$\min_{T \in \mathcal{A}} f_{\mathbf{P}_0}(T),$$

where  $\mathcal{A}$  is the acceptance region of  $\phi_{\text{eff}}$ . The optimization problem, though convex, is hard to obtain a closed-form expression, but we can still evaluate it numerically.

### B. Characterization of Achievable Exponent Region $\mathcal{R}$

One can generalize the result from Theorem 6.1. Define the following test:

$$\phi_\lambda(x^n) \triangleq \mathbb{1}_{\{f_{\mathbf{P}_0}(\Pi_{x^n}) - f_{\mathbf{P}_1}(\Pi_{x^n}) \geq \lambda\}},$$

where  $\lambda \in [-f_{\mathbf{P}_1}(M_0(\boldsymbol{\alpha})), f_{\mathbf{P}_0}(M_1(\boldsymbol{\alpha}))]$ . Following a similar idea in the proof of Theorem 6.1, one can show that  $\phi_\lambda$  is optimal in a sense that for any test  $\phi$  and  $\forall \lambda$ ,

$$E_0(\phi) \geq E_0(\phi_\lambda) \Rightarrow E_1(\phi) \leq E_1(\phi_\lambda),$$

and

$$E_1(\phi) \geq E_1(\phi_\lambda) \Rightarrow E_0(\phi) \leq E_0(\phi_\lambda),$$

where  $(E_0(\phi), E_1(\phi))$  are the error exponents with respect to test  $\phi$  :

$$\begin{cases} E_0(\phi) \triangleq \liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log \mathbb{P}_{\mathbf{F}}^{(n)}(\phi) \right\} \\ E_1(\phi) \triangleq \liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log \mathbb{P}_{\mathbf{M}}^{(n)}(\phi) \right\}. \end{cases}$$

To obtain a parametrization of the boundary of  $\mathcal{R}$ , it suffices to solve the following information projection problem:

$$\begin{cases} E_0(\lambda) \triangleq \inf_{T \in \mathcal{A}_\lambda} f_{\mathbf{P}_0}(T) \\ E_1(\lambda) \triangleq \inf_{T \in (\mathcal{A}_\lambda)^c} f_{\mathbf{P}_1}(T), \end{cases}$$



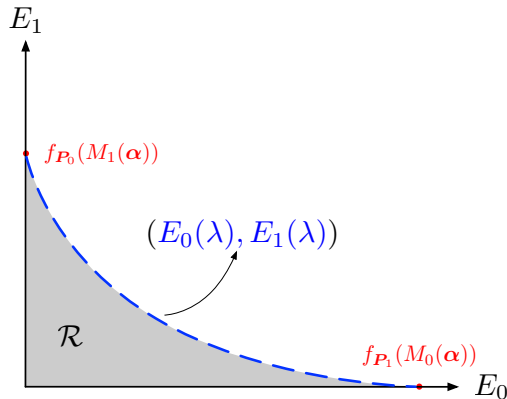


Fig. 6: Illustration of  $(E_0(\lambda), E_1(\lambda))$

where  $\mathcal{A}_\lambda \triangleq \{f_{\mathcal{P}_0}(\Pi_{x^n}) - f_{\mathcal{P}_1}(\Pi_{x^n}) \geq \lambda\}$  is the acceptance region of  $\phi_\lambda$ . Therefore,  $(E_0(\lambda), E_1(\lambda))$  parametrizes the boundary of  $\mathcal{R}$ , for  $\lambda \in [-f_{\mathcal{P}_1}(M_0(\alpha)), f_{\mathcal{P}_0}(M_1(\alpha))]$ .

In particular, we see that for the corners  $\lambda = f_{\mathcal{P}_0}(M_1(\alpha))$  and  $\lambda = -f_{\mathcal{P}_1}(M_0(\alpha))$ , we obtain the same results as in Neyman-Pearson regime (Theorem 3.2). Note that although the information-projection problem is a convex optimization problem, the closed-form expression remains unknown.

## VII. DISCUSSION

### A. Extension to Polish $\mathcal{X}$

Theorem 3.1 characterizes the optimal test in the anonymous detection problem, where only a few conditions on the  $\sigma$ -field  $\mathcal{F}$  are required. In Theorem 3.2, we further assume the alphabet  $\mathcal{X}$  is finite, in order to apply large deviation tools based on the method of types (see Remark 3.4 for discussion). However, the optimal exponent of the type-II error probability, given by the result of Theorem 3.2, depends only on the possible distributions under  $\mathcal{H}_\theta$ , and hence it is interesting to see if one can remove the assumption that  $\mathcal{X}$  being finite. Recall that in the proof, the main tool we employed is the generalized version of Sanov's theorem (see Lemma 5.1), and thus the question turns out to be whether it is possible to prove Lemma 5.1 without using method of types. Surprisingly, the answer is yes if  $\mathcal{X}$  is a Polish space (a completely separable metrizable topological space). If  $\mathcal{X}$  is Polish, the space of all probability measures on  $\mathcal{X}$  ( $\mathcal{P}_\mathcal{X}$ ) is also Polish, equipped with weak-topology induced by weak convergence. One can choose, for example, Levy-Prokhorov metric on  $\mathcal{P}_\mathcal{X}$ . The proof of standard Sanov's Theorem on Polish  $\mathcal{X}$ , however, is far more complicated than the case of finite  $\mathcal{X}$ , see [18], [19] for detailed proof. Lemma 5.1 for Polish  $\mathcal{X}$  can be proved with similar techniques. Nevertheless, in order not to digress further from the subject, we only present a proof for finite  $\mathcal{X}$  in this paper.

### B. The Benefit of Partial Information about the Group Assignment

From Figure 1, we see that in some cases, the type-II error exponent can be pushed to zero, making reliable detection no longer possible. If each sensor is allowed to transmit a few bits of information to *partially reveal* their

groups, how such partial information can improve the type-II error exponent? Formally speaking, we assume that the total number of groups is  $K$ , and each sensor can transmit  $L$  bits (with  $L < \log K$ ) through a noiseless channel to the fusion center, providing partial information about the group that it belongs to.

Unsurprisingly, the optimal strategy is the *cluster-and-detect* approach, that is, we first *cluster* the  $K$  groups into  $2^L$  super-groups, and each sensor sends  $L$  bits to indicate which *super-groups* it belongs to. Inside each super-group, we adopt the optimal anonymous hypothesis testing, and between super-groups, the problem boils down to the equivalent informed hypothesis testing, and hence standard likelihood ratio test can be applied there.

However, the difficulty lies in the clustering step: even the fusion center knows the distribution of each group, the optimal clustering algorithm is indeed a discrete optimization problem and thus NP-hard. When the group number  $K$  is large enough, it is intractable to find the optimal clustering. Nevertheless, some suboptimal algorithms suggested by heuristic do demonstrate that this partial information can significantly ameliorate the performance loss caused by anonymity. Below is a numerical example, showing the benefit of partial information.

In the example, we assume there are totally  $K = 1024$  ( $2^{10}$ ) groups, and each group accounts for  $1/K$  proportion of total sensors, that is,  $\alpha = [\frac{1}{K}, \dots, \frac{1}{K}]^\top$ . For the sensors in the  $k$ -th group, their observations follow i.i.d. distribution  $\text{Ber}(\theta_k)$  under  $\mathcal{H}_0$ , and follow i.i.d.  $\text{Ber}(1 - \theta_k)$  under  $\mathcal{H}_1$ , with  $\theta_k = \frac{k}{K}$ ,  $k = 1, \dots, K$ . Suppose there are  $L$  bits available for each sensor to partially inform the fusion center the group it belongs to, then as the clustering-detection algorithm suggests, we first cluster the  $K$  groups into  $2^L$  super-groups and then apply anonymous hypothesis testing inside each super-group. As the numerical evaluation in Figure 7 illustrates, even with few bits, say,  $L = 1$  or 2, type-II error exponents are significantly improved.

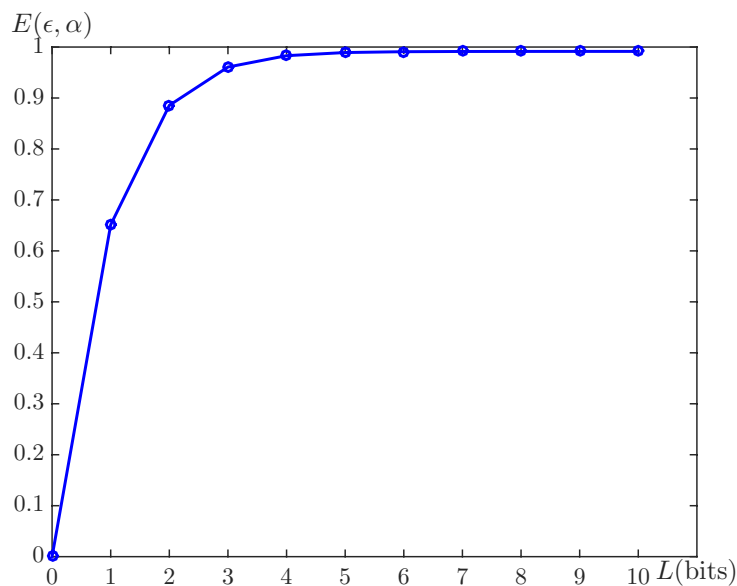


Fig. 7: Exponents with Partial Information

## VIII. CONCLUSION

In this paper, we explore the heterogeneous distributed detection problem with sensor anonymity. To address sensor anonymity, a composite hypothesis testing approach is taken. Focusing on the Neyman-Pearson setting, we provide an optimal test, and characterize the exponent of type-II error probability for the case that  $\mathcal{X}$  is finite. Unlike the settings considered in robust hypothesis testing literatures [4]–[6], since the hypothesis classes considered in our framework are discrete, the least favorable distribution might not exist. To circumvent the difficulty, we map the original problem into an auxiliary space by employing the symmetric property of the hypothesis classes, in which the original composite hypothesis testing problem becomes a simple hypothesis testing problem. Therefore, Neyman-Pearson lemma can be applied to obtain an optimal test, which is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under  $\mathcal{H}_0$  to the uniform mixture of those under  $\mathcal{H}_1$ . For the asymptotic regime, we analyze the type-II error exponent using method of types and show that the optimal exponent is the minimization of linear combination of KL-divergences, with the  $k$ -th term being  $D(U_k \| P_{1;k})$  and  $\alpha_k$  being the coefficient, for  $k = 1, \dots, K$ . The minimization is over all possible distributions  $U_1, \dots, U_K$  such that  $\sum_{k=1}^K \alpha_k U_k = \sum_{k=1}^K \alpha_k P_{0;k}$ . We further extend our result to Chernoff's regime, and indicate that the exponent region can be obtained by solving a convex optimization problem.

There are still many open problems in anonymous heterogeneous hypothesis testing. For example, the closed-form expression for the exponents in asymptotic regime, even in Neyman-Pearson formulation, are still unknown. Besides, the solution of information projection is conjectured to have similar form like tilted-distributions, as the classical results in simple hypothesis testing suggested. In addition to hypothesis testing, it is also interesting to investigate other problems such as regression, estimation, or pattern recognition under the anonymous setting.

## REFERENCES

- [1] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, H. V. Poor and J. B. Thomas, Eds. JAI Press Inc., 1990, vol. 2.
- [2] —, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals and Systems*, vol. 1, no. 2, pp. 167–182, 1988.
- [3] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, January 2009.
- [4] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [5] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.
- [6] V. V. Veeravalli, T. Başar, and H. V. Poor, "Minimax robust decentralized detection," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 35–40, January 1994.
- [7] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," August 2017. [Online]. Available: [http://people.lids.mit.edu/yp/homepage/data/itlectures\\_v5.pdf](http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf)
- [8] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, 1965.
- [9] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, 1981.
- [10] C. George and R. L. Berger, *Statistical inference*. Duxbury, 2002.
- [11] O. Zeitouni and M. Gutman, "On universal hypothesis testing via large deviations," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 285–290, March 1991.

- [12] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.
- [13] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *AMC Transactions on Programming Languages and Systems*, vol. 4, July 1982.
- [14] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Asymptotic analysis of distributed Bayesian detection with Byzantine data," *IEEE Signal Processing Letters*, vol. 22, 2015.
- [15] W.-N. Chen, H.-C. Chen, and I.-H. Wang, "On the fundamental limits of heterogeneous distributed detection: Price of anonymity," *IEEE International Symposium on Information Theory (ISIT)*, June 2018.
- [16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006, no. 0471241954.
- [17] I. Csiszár, "A simple proof of Sanov's theorem," *Bull. Braz. Math. Soc. (N.S.)*, 2006.
- [18] F. den Hollander, *Large Deviations*, ser. Fields Institute Monographs. American Mathematical Society, 2000, no. 14.
- [19] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, ser. 38. Springer-Verlag, 2010, vol. Stochastic Modelling and Applied Probability.
- [20] H. Royden and P. Fitzpatrick, *Real Analysis*. Pearson, 2010.

## APPENDIX A

### PROOF OF PROPOSITION 3.1

*proof of Proposition 3.1:* Since the optimal type-II exponent does not depend on  $\epsilon$ , we denote it as  $E^*(\boldsymbol{\alpha})$  and for simplicity. It suffices to show

$$E^*(\lambda\boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2) \leq \lambda E^*(\boldsymbol{\alpha}_1) + (1-\lambda)E^*(\boldsymbol{\alpha}_2), \forall \lambda \in [0, 1].$$

First, let

$$\begin{aligned} E^*(\boldsymbol{\alpha}_1) &= \sum_{k=1}^K \alpha_{1k} D(U_{1k}^* \| P_{1;k}) \\ E^*(\boldsymbol{\alpha}_2) &= \sum_{k=1}^K \alpha_{2k} D(U_{2k}^* \| P_{1;k}) \end{aligned}$$

where  $\boldsymbol{\alpha}_1 = [\alpha_{11}, \dots, \alpha_{1K}]^\top$ ,  $\boldsymbol{\alpha}_2 = [\alpha_{21}, \dots, \alpha_{2K}]^\top$ , and  $\mathbf{U}_1^* \triangleq [U_{11}^*, \dots, U_{1K}^*]$ ,  $\mathbf{U}_2^* \triangleq [U_{21}^*, \dots, U_{2K}^*]$  are the minimizers of (8). Then, by the convexity of KL divergence, we have

$$\begin{aligned} \lambda E^*(\boldsymbol{\alpha}_1) + (1-\lambda)E^*(\boldsymbol{\alpha}_2) &= \sum_{k=1}^K \lambda \alpha_{1k} D(U_{1k}^* \| P_{1;k}) + (1-\lambda) \alpha_{2k} D(U_{2k}^* \| P_{1;k}) \\ &\geq \sum_{k=1}^K (\lambda \alpha_{1k} + (1-\lambda) \alpha_{2k}) D\left(\frac{\lambda \alpha_{1k} U_{1k}^* + (1-\lambda) \alpha_{2k} U_{2k}^*}{\lambda \alpha_{1k} + (1-\lambda) \alpha_{2k}} \middle\| P_{1;k}\right) \end{aligned} \quad (27)$$

Now we claim that  $\tilde{\mathbf{U}} \triangleq \left(\frac{\lambda \alpha_{1k} U_{1k}^* + (1-\lambda) \alpha_{2k} U_{2k}^*}{\lambda \alpha_{1k} + (1-\lambda) \alpha_{2k}}\right)_{k=1, \dots, K}$  satisfies

$$(\lambda \boldsymbol{\alpha}_1 + (1-\lambda) \boldsymbol{\alpha}_2)^\top \tilde{\mathbf{U}} = (\lambda \boldsymbol{\alpha}_1 + (1-\lambda) \boldsymbol{\alpha}_2)^\top \mathbf{P}_0, \quad (28)$$

and thus

$$\begin{aligned} (27) &= \sum_{k=1}^K (\lambda \alpha_{1k} + (1-\lambda) \alpha_{2k}) D\left(\tilde{U}_k \middle\| P_{1;k}\right) \\ &\geq \min_{\mathbf{U} \in (\mathcal{P}_X)^K} \sum_{k=1}^K (\lambda \alpha_{1k} + (1-\lambda) \alpha_{2k}) D(U_k \| P_{1;k}) \\ &\quad (\lambda \boldsymbol{\alpha}_1 + (1-\lambda) \boldsymbol{\alpha}_2)^\top \mathbf{U} = (\lambda \boldsymbol{\alpha}_1 + (1-\lambda) \boldsymbol{\alpha}_2)^\top \mathbf{P}_0 \\ &= E^*(\lambda \boldsymbol{\alpha}_1 + (1-\lambda) \boldsymbol{\alpha}_2). \end{aligned}$$

To show (28), we notice that  $U_1^*, U_2^*$  satisfy the constraints

$$\alpha_1^\top U_1^* = \alpha_1^\top P_0, \alpha_2^\top U_2^* = \alpha_2^\top P_0. \quad (29)$$

Then we have

$$\begin{aligned} & (\lambda \alpha_1 + (1 - \lambda) \alpha_2)^\top \tilde{U} \\ &= \sum_{k=1}^K (\lambda \alpha_{1k} + (1 - \lambda) \alpha_{2k}) \left( \frac{\lambda \alpha_{1k} U_{1k}^* + (1 - \lambda) \alpha_{2k} U_{2k}^*}{\lambda \alpha_{1k} + (1 - \lambda) \alpha_{2k}} \right) \\ &= \sum_{k=1}^K \lambda \alpha_{1k} U_{1k}^* + (1 - \lambda) \alpha_{2k} U_{2k}^* \\ &= \lambda \alpha_1^\top U_1^* + (1 - \lambda) \alpha_2^\top U_2^* \\ &= (\lambda \alpha_1 + (1 - \lambda) \alpha_2)^\top P_0, \end{aligned}$$

which completes the proof. ■

## APPENDIX B

### PROOF OF LEMMA 5.1

*proof of Lemma 5.1:* First, observe that since  $\text{int } \Gamma$  is open, the set

$$\tilde{\Gamma} \triangleq \{(U_1, \dots, U_K) \mid \alpha^\top U \in \text{int } \Gamma\} \subset (\mathcal{P}_{\mathcal{X}})^K$$

is open too. This is because the mapping  $g(U) = \alpha^\top U$  is continuous, so the pre-image preserves the openness (under standard topology). Therefore, we can find a sequence

$$\{U^{(n)} \in (\mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K}) \cap \tilde{\Gamma}\},$$

such that

$$\sum_k \alpha_k D(U_k^{(n)} \parallel P_{\theta;k}) \rightarrow - \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top U \in \text{int } \Gamma}} \sum_k \alpha_k D(U_k \parallel P_{\theta;k}),$$

where the limit is taken such that  $\frac{n_k}{n} \rightarrow \alpha_k$ . So we have

$$\begin{aligned} \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} &= \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\geq \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\geq \max_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k^{(n)})\} \\ &\stackrel{(a)}{\geq} \max_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \left( \frac{1}{(n_k + 1)^{|\mathcal{X}|}} \right) 2^{\sum_{k=1}^K n_k D(U_k^{(n)} \parallel P_{\theta;k})}, \end{aligned}$$

where inequality (a) holds by Lemma 2.2. Thus we have

$$\frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} \geq - \min_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top \mathbf{U} \in \text{int } \Gamma}} \left( \sum_{k=1}^K \frac{n_k}{n} D(U_k^{(n)} \| P_{\theta; k}) + o(1) \right).$$

As  $n \rightarrow \infty$  such that  $\frac{n_k}{n} \rightarrow \alpha_k$ , we see that

$$- \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top \mathbf{U} \in \text{int } \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta; k}) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\}.$$

On the other hand, for the upper bound, consider

$$\begin{aligned} \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} &= \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top \mathbf{U} \in \Gamma}} \prod_{k=1}^K P_{\theta; k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\stackrel{(a)}{\leq} \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top \mathbf{U} \in \Gamma}} 2^{\sum_{k=1}^K D(U_k^{(n)} \| P_{\theta; k})} \\ &\leq \left( \prod_k |\mathcal{P}_{n^k}| \right) 2^{\sum_{k=1}^K n_k D(U_k^{(n)} \| P_{\theta; k})} \\ &\stackrel{(b)}{=} 2^{\left( \sum_{k=1}^K n_k D(U_k^{(n)} \| P_{\theta; k}) + o(1) \right)}, \end{aligned}$$

where where inequality (a) holds by Lemma 2.2, and (b) holds due to the cardinality bound Lemma 2.1.

As  $n \rightarrow \infty$  and  $\frac{n_k}{n} \rightarrow \alpha_k$ , we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} \leq - \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top \mathbf{U} \in \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta; k}).$$

Notice that for the case  $\mathcal{X}$  finite, the infimum takes over  $\Gamma$  is equal to that one takes in the closure of  $\Gamma$ , since we can use standard topology to find a sequence approaching to the limit point. Thus the proof is complete.  $\blacksquare$

## APPENDIX C

### PROOF OF LEMMA 5.2

*proof of Lemma 5.2:* Let  $\mathbf{Q} \in (\mathcal{P}_{\mathcal{X}})^K$  be a K-tuple of probability measure on  $\mathcal{X}$ . We first show that

$$\mathcal{C}_{\mathbf{Q}} \triangleq \{T \in \mathcal{P}_{\mathcal{X}} : f_{\mathbf{Q}}(T) < \infty\}$$

is a compact set.

*Part 1 (Compactness).* Observe that  $f_{\mathbf{Q}}(T) < \infty$  if and only if there exists a  $\mathbf{P} = (P_1, \dots, P_K) \in (\mathcal{P}_{\mathcal{X}})^K$ , such that

- 1)  $\alpha^\top \mathbf{P} = T$
- 2) for all  $i = 1, \dots, K$ ,  $P_i \ll Q_i$ .

Therefore, let us denote

$$\mathcal{M}_{\mathbf{Q}} \triangleq \left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \ll Q_i, \forall i = 1, \dots, K \right\} \subseteq (\mathcal{P}_{\mathcal{X}})^K.$$

We claim that  $\mathcal{M}_Q$  is a compact set, and thus

$$\mathcal{C}_Q = \{\alpha^\top \mathbf{P} \mid \mathbf{P} \in \mathcal{M}_Q\}$$

is also compact, since  $\alpha^\top \mathbf{P}$  is a linear mapping from  $(\mathcal{P}_\mathcal{X})^K$  to  $\mathcal{P}_\mathcal{X}$  so compactness is preserved. To prove the claim, it suffices to show that  $\mathcal{M}_Q$  is a closed set, because the boundness is directly followed by the boundness of  $(\mathcal{P}_\mathcal{X})^K$ . It is equivalent to show

$$\mathcal{M}_Q^C = \left\{ \mathbf{P} \in (\mathcal{P}_\mathcal{X})^K : P_i \not\ll Q_i, \text{ for some } i \right\}$$

is open. Notice that

$$\left\{ \mathbf{P} \in (\mathcal{P}_\mathcal{X})^K : P_i \not\ll Q_i, \text{ for some } i \right\} = \bigcup_{i=1}^K \left\{ \mathbf{P} \in (\mathcal{P}_\mathcal{X})^K : P_i \not\ll Q_i \right\},$$

so it suffices to show  $\left\{ \mathbf{P} \in (\mathcal{P}_\mathcal{X})^K : P_i \not\ll Q_i \right\}$  is open for all  $i$ . Assume  $P_i \not\ll Q_i$ . Then there must exist some measurable event  $\mathcal{E} \subset \mathcal{X}$ , such that  $Q_i(\mathcal{E}) = 0$ , and  $P_i(\mathcal{E}) = \epsilon > 0$ . Therefore, if  $\mathcal{X}$  is finite and thus  $\mathcal{P}_\mathcal{X}$  equipped with total-variation distance (i.e. one norm), then obviously for any  $\tilde{Q}$  such that  $\|\tilde{Q} - P_i\| < \frac{\epsilon}{2}$ ,  $\tilde{Q} \not\ll Q_i$ . Hence  $\mathcal{M}_Q^C$  is open, proving the claim.

*Remark 3.1.* If  $\mathcal{X}$  is Polish, then  $\mathcal{P}_\mathcal{X}$  is equipped with Prokhorov's metric, and one can use similar argument to show that  $\mathcal{M}_Q^C$  is open.

Next, we show that  $f_Q(\cdot)$  is a convex function, so the convexity of  $\mathcal{C}_Q$  follows: for all  $T_1, T_2 \in \mathcal{C}_Q$ ,

$$f_Q(\lambda T_1 + (1 - \lambda)T_2) \leq \lambda f_Q(T_1) + (1 - \lambda)f_Q(T_2) < \infty, \quad (30)$$

implying  $\lambda T_1 + (1 - \lambda)T_2 \in \mathcal{C}_Q$ .

*Part 2 (Convexity).* To show (30), we observe

$$\begin{aligned} & \lambda f_Q(T_1) + (1 - \lambda)f_Q(T_2) \\ &= \inf_{\mathbf{U}: \alpha^\top \mathbf{U} = T_1} \lambda \sum_k \alpha_k D(U_k \| P_k) + \inf_{\mathbf{V}: \alpha^\top \mathbf{V} = T_2} (1 - \lambda) \sum_k \alpha_k D(V_k \| P_k) \\ &\stackrel{(a)}{\geq} \inf_{\mathbf{U}, \mathbf{V}: \alpha^\top \mathbf{U} = T_1, \alpha^\top \mathbf{V} = T_2} \sum_k \alpha_k D(\lambda U_k + (1 - \lambda)V_k \| P_k) \\ &\stackrel{(b)}{\geq} \inf_{\mathbf{P}: \alpha^\top \mathbf{P} = \lambda T_1 + (1 - \lambda)T_2} \sum_k \alpha_k D(P_k \| P_k) \\ &= f_Q(\lambda T_1 + (1 - \lambda)T_2), \end{aligned}$$

where (a) is due to the convexity of KL-divergence, and (b) is because

$$\alpha^\top \mathbf{U} = T_1, \alpha^\top \mathbf{V} = T_2 \Rightarrow \alpha^\top (\lambda \mathbf{U} + (1 - \lambda)\mathbf{V}) = \lambda T_1 + (1 - \lambda)T_2.$$

Therefore, we conclude that  $f_Q(\cdot)$  is a convex function and  $\mathcal{C}_Q$  is a convex set.

At the final step, we show  $f_Q(\cdot)$  is a continuous function on  $\mathcal{C}_Q$ . Notice that the convexity of  $f_Q(\cdot)$  only guarantees the continuity on the interior of  $\mathcal{C}_Q$ , and thus we need to additionally check the boundary points.

**Remark 3.2.** Note that in general, the interior of  $\mathcal{C}_Q$  may be an empty set since it may lie in a subspace of  $\mathcal{P}_X$ . Alternatively, we can define a point  $P$  being interior, if it can be written as

$$\lambda U + (1 - \lambda)V, \text{ for some } \lambda \in (0, 1), \text{ and some } V, U \in \mathcal{C}_Q.$$

*Part 3 (Continuity).* First, if the interior of  $\mathcal{C}_Q$  is empty, then by the convexity, either  $\mathcal{C}_Q$  is a empty set, or it is a singleton. For both cases, the continuity holds obviously. Hence without losing of generality, we assume that the interior of  $\mathcal{C}_Q$  is non-empty, and  $T_0$  is an interior point.

Then for any  $T \in \mathcal{C}_Q$ , we can construct a sequence  $T_n \in \mathcal{C}_Q$ ,  $T_n \rightarrow T$ . For example, one can let  $T_n = \lambda_n T_0 + (1 - \lambda_n)T$ , with  $\lambda_n \rightarrow 0$ . Let  $\mathbf{U}^{(n)} = (U_1^{(n)}, \dots, U_K^{(n)}) \in (\mathcal{P}_X)^K$  be a sequence such that

- 1)  $\alpha^\top \mathbf{U}^{(n)} = T_n$
- 2)  $\mathbf{U}^{(n)}$  achieves the infimum of  $f_Q(T_n)$  :

$$\sum_{k=1}^K \alpha_k D(U_k^{(n)} \| P_k) = \inf_{\mathbf{V}: \alpha^\top \mathbf{V} = T_n} \sum_{k=1}^K \alpha_k D(V_k \| P_k) = f_Q(T_n).$$

Notice that the infimum can always be achieved since  $g(\mathbf{V}) \triangleq \sum_{k=1}^K \alpha_k D(V_k \| P_k)$  is a continuous function over the compact set  $\mathcal{M}_Q$ .

By construction,  $\mathbf{U}^{(n)}$  is a sequence in a compact set  $\mathcal{M}_Q$ , and hence by Bolzano-Weierstrass theorem (see Chapter 1 in [20], for example), there exists a convergent subsequence  $\mathbf{U}^{(n_i)}$ , and let us denote the convergent point

$$\lim_{i \rightarrow \infty} \mathbf{U}^{(n_i)} = \mathbf{U}.$$

Since  $\alpha^\top \mathbf{U}^{(n_i)} = T_{n_i}$ , and  $T_{n_i} \rightarrow T$ , we have

$$\alpha^\top \mathbf{U} = T.$$

Notice that the function  $f(\mathbf{V}) \triangleq \sum_{k=1}^K \alpha_k D(V_k \| P_k)$  is a continuous function over the compact set  $\mathcal{M}_Q$ , we must have

$$\lim_{n \rightarrow \infty} f_Q(T_n) = \lim_{i \rightarrow \infty} f_Q(T_{n_i}) = \sum_{k=1}^K \alpha_k D(U_k \| P_k),$$

and therefore

$$f_Q(T) = \inf_{\mathbf{U}: \alpha^\top \mathbf{U} = T} \sum_{k=1}^K \alpha_k D(U_k \| P_k) \leq \sum_{k=1}^K \alpha_k D(U_k \| P_k) = \lim_{n \rightarrow \infty} f_Q(T_n).$$

On the other hand, by the convexity of  $f_Q(\cdot)$ , we must have

$$f_Q(T) \geq f_Q(T_n), \text{ for all } n \text{ large enough.}$$

Otherwise

$$f_Q(\lambda T_0 + (1 - \lambda)T) > \lambda f_Q(T_0) + (1 - \lambda)f_Q(T),$$

for some  $\lambda$  small enough, which violates the fact that  $f_Q(\cdot)$  is a convex function. ■