Algorithms and Fundamental Limits for Unlabeled Detection using Types

Stefano Marano, and Peter Willett, *Fellow, IEEE*

***Abstract*—We deal with the classical problem of testing two simple statistical hypotheses but, as a new element, it is assumed that the data vector is observed after an unknown permutation of its entries. What is the fundamental limit for the detection performance in this case? How much information for detection is contained in the entry values and how much in their positions? In the first part of the paper we answer these questions. In the second part we focus on practical algorithms. A low-complexity detector solves the detection problem without attempting to estimate the permutation. A modified version of the auction algorithm is then considered, and two greedy algorithms with affordable worst-case complexity are presented. The detection operational characteristics of these detectors are investigated by computer experiments. The problem we address is referred to as *unlabeled detection* and is motivated by large sensor network applications, but applications are also foreseen in different fields, including image processing, social sensing, genome research, molecular communication.**

***Index Terms*—Unlabeled detection, unordered data, unknown permutation, fundamental limits of hypothesis testing, error exponents, types, assignment problem.**

I. Introduction

Mostly motivated by emerging applications of sensor networks, recent years have seen the birth of a field that can be referred to as *signal processing with unlabeled data*. This terminology refers to the bulk of classical algorithms and methods of signal processing, revisited under the new paradigm of a central unit that must process a vector of data received from some peripheral units, but must do so — or choose do so — without access to the data *labels*, namely without knowing the original position of each datum inside the vector. The meaning here given to "labeling" is that of *provenance* or *identity*, and is not to be conflated with the labeling obtained by data classification, as typical, for instance, of machine learning applications. To avoid misunderstanding, this aspect should be stressed: following [1]–[6], in this article an unlabeled vector of observations refers to the set of values, with no information about the position of the individual entries, nor from which peripheral units they may have arisen. Note that there is a debate about the terminology, and the adjective "unordered" is sometimes preferred to "unlabeled". Note also that in this work we are interested in the first case mentioned before, that the processing must proceed without labels by necessity; a related idea is that of random finite set (RFS), and good entry points are [7], [8].

As a notional example, consider a binary hypothesis testing using two sensors, and suppose that under the null hypothesis the two sensors' observations are independent and identically distributed (*iid*) unit-normal, whereas under the alternative their means are shifted, respectively by $+1.2$ and -1.2. The central decision-maker receives the set $\{-1.3, +1.3\}$, and is specifically told that it should make no assumption about which observation came from which sensor. Intuition suggests that the first sensor's observation is $+1.3$ and the second sensor saw -1.3; and hence that there is a fairly decent fit with the alternative hypothesis. How much decision-making performance has been lost by *label-agnostic* decision-making with respect to *label-aware* in this case? That is, how much information is in knowing who said what, as opposed simply to knowing what was said? And how about the case that the two mean shifts were respectively 1.1 and 1.3: clearly the quality of the match is much lower; but equally clearly the impact of making a labeling error is far lower.

A. Contribution

For the problem of testing two statistical hypotheses, upon observing an n-vector $\mathbf{x}^n$ made of independent entries, one goal of this article is to develop the fundamental limiting ($n \to \infty$) detection rate when data are unlabeled. That is, we pose the question: what is the optimal theoretical detection performance in situations where only an unordered version of $\mathbf{x}^n$ is observed, namely, when we know the values of the entries of $\mathbf{x}^n$ but not their ordering? How much information for detection is contained in the entry *labels* and hence is lost, and how much in the entry *values*, and hence retained by the unlabeled version of $\mathbf{x}^n$? The aforementioned notional example suggests that even the unlabeled version of $\mathbf{x}^n$ carries some information for detection. However, little more than this naïve notion is known, and we aim toward filling this gap.

After answering the above theoretical questions we make a step further. Characterizing the ultimate detection performance does not tell very much about the possibility of solving the unlabeled detection problem with *practical* detectors. This motivates us to investigate if there exist detection algorithms with affordable computational complexity and acceptable performance for finite values of n. A solution which avoids any attempt to estimate the labels is firstly proposed. Then, we show that the unlabeled detection problem with discrete data can be recast in the form of a classical assignment problem, for which optimal algorithms are known, but which can be highly inefficient for our problem. Thus, we present two algorithms which require lower computational complexity.

S. Marano is with DIEM, University of Salerno, I-84084, Fisciano (SA), Italy (e-mail: marano@unisa.it) P. Willett is with ECE Dept., University of Connecticut, Storrs, CT, USA (e-mail: peter.willett@uconn.edu).

P. Willett was supported by NPS via ONR contract N00244-16-1-0017 and by AFOSR under contract FA9500-18-1-0463.

Part of Sec. IV of this work has been presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018) 15-20 April 2018, Calgary, Alberta, Canada [1]. Theorem 2 offers an improved expression that should be considered a replacement for that in [1].

Detection performance and computational burden of these detectors are investigated by computer simulations.

B. Motivations, Related Work & Organization

Modern networks are vulnerable to malicious attacks. For instance, the civilian global positioning system (GPS) is particularly exposed to spoofing attacks [9], which can impair wireless ad-hoc communication systems [10], or alter the timing information in smart grids, by carrying out a so-called timing synchronization attack (TSA) [11], [12]. As a consequence of TSA, the timestamp information of the system may be altered to the point that the data arriving to a central decision unit can be considered unlabeled.

Even in absence of an attack, modern sensor networks and other networked inference/communication systems are similarly vulnerable, especially when faced with big-data applications. Assignment algorithms are used in multi-target tracking procedures as an efficient way to associate observations to targets [13]. Networked control systems with packetized messages can be subject to various timing errors due to uncontrollable packet delays [14]. In [15] the lack of a precise timestamp of data is considered in connection with the usage of automatic identification system (AIS) in real-world maritime surveillance problems. The issue common to all these examples is that data must be processed with partial information about their relative time/space ordering, which is often related to their provenance from a peripheral unit of the network.

Data labeling can be also completely missing. Modern applications of large wireless sensor networks frequently impose severe constraints on the messages delivered to a fusion center, due to limited sensors' resources, e.g., energy, bandwidth, etc. In these applications, including the identities of the reporting sensors in the delivered messages might constitute an excessive burden [16] and, for the same reason, the delivered data are usually constrained to belong to a finite alphabet with small cardinality. On the other hand, the characteristics of the sensors can be known, in the sense that the data distributions at each sensor, under two alternative hypotheses, are available — but with the identity of the transmitting sensor unknown to the fusion center, how can this be used? This scenario represents the main motivation behind our work. But the problem studied here has applications in different fields. Recovering of data labeling arises in image processing, see e.g., [17], [18]. Dealing with partial of fully unordered data is typical of modern genome research [19], as well as archeology [20]. Among the most promising application fields for the unlabeled detection problem studied in this article, we mention the vibrant area of molecular communication, where molecules emitted by human cells or nonomachines undergo the effect of a timing channel and appear completely unordered at the receiver [21], [22]. Analogous permutation channels were also considered in [23].

A problem similar to that addressed here is studied in [24]. The author of [24] considered the problem of matching a set of observed sequences to a set of candidate distributions (or to a set of candidate training sequences), with application to the so-called de-anonymization attack, where one attempts to recovering the identity of individuals from anonymized datasets. The approach of [24] is based on finding the best matching between sequences and candidate source statistics, by exploiting the KL divergence as metric. This approach becomes effective when the length of each sequence is large, so that its empirical distribution matches well the distribution of the source that produced the data. In the present work we have instead a single datum for each source, and the number of sources is large. As a consequence, our approach and results are very different from those in [24].

Similar studies have been addressed in [25]–[28], where the goal is to profile individuals, rather than recovering their identities as done in [24]. In particular, in the nonparametric setting considered in [28] the *type* of the data plays a central role, as in the present paper. Aside from that, our approach is very different also from those pursued in [25]–[28].

A systematic study of the lack-of-identity issue in the form addressed here, which is nowadays referred to as the *unlabeled data* paradigm, has been prompted by [2], [3]. The authors of [2] and [3] consider a signal recovery problem from a set of unlabeled linear projections. They also compare their unlabeled sensing formulation with the setting of compressed sensing (see, e.g., [29], [30]), and highlight connections with a classical problem in robotics, which is known as simultaneous location and mapping (SLAM). With SLAM, measurements are made in an unknown environment, which makes it necessary to estimate relative permutations between measurements [31]. Very recent studies with a similar data-reconstruction focus can be found in [32]–[35].

In contrast to data reconstruction, our focus is on *inference* by unlabeled data, which has been addressed in the last few years by [4]–[6], [36]. The analysis in [36] is limited to the detection of a signal embedded in Gaussian noise, where the signal values are sequentially sampled from a known sequence, which is a form of unlabeling different from that used here. Closer to our formulation are the works in [4]–[6]. In particular, we elaborate on a detection problem similar to that addressed in [6]. However, in [6] a Gaussian (hence continuous) shift-in-mean model is considered, with label uncertainty only under the alternative hypothesis, while here we deal with general distributions over finite alphabets, and data unlabeling under both hypotheses.

The remainder of this paper is organized as follows. The next section introduces the classical setup of detection with labeled data. Section III formalizes the unlabeled detection problem and presents the main theoretical results. Practical algorithms for unlabeled detection are considered in Sec. IV, while the results of computer experiments are presented in Sec. V. Section VI concludes the paper. Some technical material is postponed to Appendices A-D.

II. Classical Detection with Labeled Data

Let $\mathbf{X}^n = (X_1, \ldots, X_n)$ be a vector whose entries are n random variables defined over a common finite alphabet $\mathcal{X}$, and let $\mathbf{x}^n = (x_1, \cdots, x_n)$ be the correspondent realization. We focus on the asymptotic scenario of $n \to \infty$, and is therefore appropriate to add a superscript ^{n} to specify the size of the vectors. Also, let $\mathcal{P}(\mathcal{X})$ denote the set of all probability mass functions (PMFs) on $\mathcal{X}$. As usual, $\mathcal{X}^n$ denotes the n-th extension of the alphabet $\mathcal{X}$ — the concatenation of n letters from $\mathcal{X}$ — and $\mathcal{P}(\mathcal{X}^n)$ denotes the set of PMFs over $\mathcal{X}^n$.

The binary hypothesis test we consider is as follows. Under hypothesis $\mathcal{H}_0$ the joint probability $q_{1:n}(\mathbf{x}^n)$ of vector $\mathbf{X}^n$ is the product of possibly non-identical marginal PMFs $q_{1:n}(\mathbf{x}^n) = \prod_{i=1}^n q_i(x_i)$, where $q_i \in \mathcal{P}(\mathcal{X})$. Thus, $q_i(x)$ refers to the i-th entry of the sequence of distributions $q_1, q_2, \dots, q_n$, computed at the alphabet element $x \in \mathcal{X}$. Likewise, under $\mathcal{H}_1$ the joint probability $p_{1:n}(\mathbf{x}^n)$ is the product of possibly non-identical marginal PMFs $p_{1:n}(\mathbf{x}^n) = \prod_{i=1}^n p_i(x_i)$, with $p_i \in \mathcal{P}(\mathcal{X})$. This means that, under both hypotheses, data are independent but not necessarily identically distributed. Formally, we have

$$\mathbf{X}^n \sim r_{1:n}(\mathbf{x}^n) = \prod_{i=1}^n r_i(x_i), \quad \left\{ \begin{array}{l} \mathcal{H}_1 : r_i(x_i) = p_i(x_i), \\ \mathcal{H}_0 : r_i(x_i) = q_i(x_i), \end{array} \right. \tag{1}$$

for $n = 1, 2, \dots$. It is assumed throughout that $q_i(x) > 0$ and $p_i(x) > 0$, for all $i = 1, 2, \dots, n$, and all $x \in \mathcal{X}$. This simplifies some results and excludes the singular cases in which the test can be solved without error for $n \to \infty$. Note also that the sequences $q_{1:n}$ and $p_{1:n}$ are assumed known.

The Kullback-Leibler divergence from $q_i(x)$ to $p_i(x)$ is defined as [37]: $D(q_i \| p_i) \triangleq \sum_{x \in \mathcal{X}} q_i(x) \log \frac{q_i(x)}{p_i(x)}$, and the assumption of strictly positive PMFs implies that $D(q_i \| p_i)$ exists and is finite for all i. All logarithms are to base e.

The error probabilities of test (1) are

$$\mathbb{P}_0(\mathbf{X}^n \notin A_n), \qquad \text{type I error,} \tag{2}$$

$$\mathbb{P}_1(\mathbf{X}^n \in A_n), \qquad \text{type II error,} \tag{3}$$

where $A_n \subseteq \mathcal{X}^n$ is some decision region in favor of $\mathcal{H}_0$, and $\mathbb{P}_h$ is the probability operator under $\mathcal{H}_h$, $h = 0, 1$.

For two sequences of distributions[1] $q_{1:\infty}, p_{1:\infty} \in \mathcal{P}(\mathcal{X}^\infty)$, let us define the *divergence rate*

$$\bar{D}(q_{1:\infty} \| p_{1:\infty}) \triangleq \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n D(q_i \| p_i). \tag{4}$$

We assume that all divergence rates encountered exist, are finite, and are continuous and convex functions of their arguments. These are very mild requirements that rule out pathological choices of the sequences $p_{1:\infty}, q_{1:\infty}$ that are of no practical interest. Let us introduce now the error exponent function, and then state two classical results about the asymptotic error exponents of the hypothesis test.

DEFINITION (Error Exponent for Labeled Data)*: For $0 < \alpha < \infty$, let us define*

$$\Omega_{\text{lab}}(\alpha) \triangleq \inf_{\omega_{1:\infty} \in \mathcal{P}(\mathcal{X}^\infty) :\, \bar{D}(\omega_{1:\infty} \| q_{1:\infty}) < \alpha} \bar{D}(\omega_{1:\infty} \| p_{1:\infty}). \tag{5}$$

It is useful to bear in mind that $\Omega_{\text{lab}}(\alpha)$ depends on the sequences $p_{1:\infty}$ and $q_{1:\infty}$. When needed, we use the more precise notation $\Omega_{\text{lab}}(\alpha; p_{1:\infty}, q_{1:\infty})$.

PROPOSITION 1 (Labeled Detection [38]) *Consider the hypothesis test (1). Let $0 < \alpha < \infty$.*

a) *Let $A_n \subseteq \mathcal{X}^n$ be any sequence of acceptance regions for $\mathcal{H}_0$. Then:*

$$\liminf_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_0(\mathbf{X}^n \notin A_n) \geq \alpha$$
$$\Rightarrow \limsup_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_1(\mathbf{X}^n \in A_n) \leq \Omega_{\text{lab}}(\alpha). \tag{6}$$

b) *There exists a sequence $A_n^* \subseteq \mathcal{X}^n$ of acceptance regions for $\mathcal{H}_0$ such that*

$$\liminf_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_0(\mathbf{X}^n \notin A_n^*) \geq \alpha, \tag{7a}$$

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_1(\mathbf{X}^n \in A_n^*) = \Omega_{\text{lab}}(\alpha). \tag{7b}$$

Proof: This is a standard result and the proof is sketched in Appendix A •

Part a) of the proposition states that whatever the sequence of decision regions is, if the type I error tends exponentially to zero at rate not smaller than α, then the type II error tends to zero exponentially at rate not larger than $\Omega_{\text{lab}}(\alpha)$. Part b) states that the above limits are tight in the sense that there exists a sequence A_n^* of decision regions such that the best rate $\Omega_{\text{lab}}(\alpha)$ for type II error is achieved.

For two sequences a_n and b_n, the symbol $a_n \doteq b_n$ means equality to the first order in the exponent, namely $\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Informally, we can summarize the content of Proposition 1 by saying that there exist tests with type I error $\doteq e^{-n\alpha}$ and type II error $\doteq e^{-n\Omega_{\text{lab}}(\alpha)}$, but no stronger pairs of asymptotic expressions can be found.

Note from (5) that $\lim_{\alpha \to 0} \Omega_{\text{lab}}(\alpha) = \bar{D}(q_{1:\infty} \| p_{1:\infty})$. The following standard result emphasizes the operational meaning of this divergence rate. Let VAR_0 be the variance under $\mathcal{H}_0$.

PROPOSITION 2 (Chernoff-Stein's Lemma [37]) *Suppose that* $\text{VAR}_0[\log(q_i(X_i)/p_i(X_i))] \leq \sigma^2 < \infty$, *for all $i = 1, 2, \dots$, and let* $P^*_{n,\theta} = \min_{A_n \subseteq \mathcal{X}^n :\, \mathbb{P}_0(\mathbf{X}^n \notin A_n) \leq \theta} \mathbb{P}_1(\mathbf{X}^n \in A_n)$, *where $0 < \theta < 1/2$. Then*

$$\lim_{n \to \infty} -\frac{1}{n} \log P^*_{n,\theta} = \bar{D}(q_{1:\infty} \| p_{1:\infty}). \tag{8}$$

Proof: See Appendix A for a sketch of proof. •

In words: for "arbitrarily" constrained type I error, the exponent of type II error can be made equal to $\bar{D}(q_{1:\infty} \| p_{1:\infty})$, but not larger.

III. DETECTION WITH UNLABELED DATA

Consider now the case of *unlabeled* data. Suppose that, instead of (1), we are faced with a binary hypothesis test in which we observe the unlabeled vector $\mathbf{X}_u^n \triangleq \mathcal{M}^{(\pi)} \mathbf{X}^n$, where $\mathcal{M}^{(\pi)}$ is a permutation matrix, indexed by an unknown $\pi \in \{1, \dots, n!\}$. Namely, let us consider the following test:

$$\mathbf{X}_u^n = \mathcal{M}^{(\pi)} \mathbf{X}^n \text{ with } \mathbf{X}^n \sim r_{1:n}(\mathbf{x}^n) = \textstyle\prod_{i=1}^n r_i(x_i),$$
$$\text{where } \left\{ \begin{array}{ll} \mathcal{H}_1 : & r_i(x_i) = p_i(x_i), \\ \mathcal{H}_0 : & r_i(x_i) = q_i(x_i), \end{array} \right. \tag{9}$$

for $n = 1, 2, \dots$. In (9) the permutation matrix applied to the data vector $\mathbf{X}^n$ is unknown.

[1] We often simplify the notation by omitting the argument $\mathbf{x}^n$: we simply write $q_{1:n}$, $p_{1:n}$, for $q_{1:n}(\mathbf{x}^n)$, $p_{1:n}(\mathbf{x}^n)$, and similar.

We know that the n observations are drawn from the n PMFs $\{p_i\}_{i=1}^n$ under $\mathcal{H}_1$ and from the n PMFs $\{q_i\}_{i=1}^n$ under $\mathcal{H}_0$, but we cannot make the association between observations and PMFs. In other words, under $\mathcal{H}_1$, for each X_j, $j = 1, \ldots, n$, we do not know from which, among the n PMFs $\{p_i\}_{i=1}^n$, it has been drawn, and the same is true under $\mathcal{H}_0$, with $\{p_i\}_{i=1}^n$ replaced by $\{q_i\}_{i=1}^n$.

Given a constraint on type I error, what is the best asymptotic performance in terms of exponential rate for type II error, when one has only access to the unlabeled vector $\mathbf{X}_u^n$? Does there exist an equivalent of Proposition 1 for unlabeled data? The answers are based on the following obvious but important lemma, stating a one-to-one correspondence between unlabeled vectors and types. Let $\mathbb{I}(A)$ be the indicator of event A.

LEMMA (Unlabeled Vectors and Types)*: For independent random variables drawn from a common finite alphabet $\mathcal{X}$, knowledge of the unlabeled version $\mathbf{X}_u^n$ of vector $\mathbf{X}^n$ is equivalent, for the detection purposes at hand, to knowledge of the* type *(or empirical PMF) of $\mathbf{X}^n$, which is*

$$t_{\mathbf{X}^n}(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = x), \quad x \in \mathcal{X}. \tag{10}$$

Thus, a detection problem where the observation is the unlabeled vector $\mathbf{x}_u^n$, is equivalent to a detection problem in which one observes $t_{\mathbf{x}^n} \in \mathcal{P}_n$, where $\mathcal{P}_n$ denotes the class of n-types. For later use, we note that $t_{\mathbf{X}^n}$ converges, for $n \to \infty$, as shown in Appendix D.

Detection with unlabeled data can be also regarded in the framework of invariance theory [39, Chap. 6]. Under $\mathcal{H}_1$ we have a class of possible distributions because we only know that one of the $n!$ permutation matrices $\mathcal{M}^{(\pi)}$ has been applied to the unobserved $\mathbf{X}^n$, but we do not know which. The permutation plays the role of a nuisance parameter. Then, we can consider the class of invariant tests under the group of the $n!$ permutations of the data, which are the tests that depend on the data only through the type vector $t_{\mathbf{X}^n}$, see [39, Th. 6.2.1]. A UMP (uniformly most powerful) invariant test can be found as shown in [39, Th. 6.3.1]. This test, however, is not feasible for any reasonable problem size, nor it is used in the following analysis.

We now describe our next steps. First, we define the error exponent function $\Omega(\alpha)$ in (14), then we present its main properties in Theorem 1, and finally we show — which is our main result given in Theorem 2 — that $\Omega(\alpha)$ is the ultimate performance limit for detection in the unlabeled case, just as $\Omega_{\text{lab}}(\alpha)$ represents the ultimate performance limit in the classical case of labeled observations. The starting point of the analysis is function $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ in (12), because $\Omega(\alpha)$ is defined in terms of its Legendre transform $\Psi_{\mathcal{H}_h}(\omega; r_{1:\infty})$. Note that we shall use vector λ as the "original" independent variable, and vector ω as the corresponding variable in the Legendre domain. In turn, function $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ is the arithmetic average of the functions defined in (11). Thus, let us begin by introducing function $\varphi_{\mathcal{H}_h} : \Re^{|\mathcal{X}|-1} \mapsto (0, \infty)$ with $h = 0, 1$. Consider the reduced alphabet $\mathcal{X}' \triangleq \mathcal{X} \setminus \{x'\}$ in which an arbitrarily selected entry, say $x' \in \mathcal{X}$, is excluded. Recall from (9) that r_i denotes the distribution of the i-th (unpermuted) observation. We let

$$\varphi_{\mathcal{H}_h}(\lambda; r_i) \triangleq \log \sum_{x\in\mathcal{X}} r_i(x) e^{\lambda(x)}, \tag{11}$$

where vector $\lambda \in \Re^{|\mathcal{X}|-1}$ has entries $\lambda(x)$, $x \in \mathcal{X}'$, and we add the *dummy* entry $\lambda(x') = 0$. Clearly, $\varphi_{\mathcal{H}_h}(\lambda; r_i) < \infty$ for all $\lambda \in \Re^{|\mathcal{X}|-1}$ and $\varphi_{\mathcal{H}_h}(0; r_i) = 0$. In Appendix B it is shown that $\varphi_{\mathcal{H}_h}(\lambda; r_i)$ is strictly convex and twice continuously differentiable throughout $\lambda \in \Re^{|\mathcal{X}|-1}$. It is also shown that the gradient $\nabla\varphi_{\mathcal{H}_h}$ is a mapping from $\lambda \in \Re^{|\mathcal{X}|-1}$ to the set of $|\mathcal{X}|-1$ positive values $0 < \omega(x) < 1$, $x \in \mathcal{X}'$, which, with the addition of the entry $\omega(x') = 1 - \sum_{x\in\mathcal{X}'} \omega(x)$, becomes the set of probability distributions $\omega \in \mathcal{P}(\mathcal{X})$ having strictly positive entries. Henceforth, we assume that vector $\lambda(x)$, $x \in \mathcal{X}'$, is enlarged by the addition of $\lambda(x') = 0$ and, likewise, vector $\omega(x)$, $x \in \mathcal{X}'$, is enlarged by the addition of $\omega(x')$. This way, a point of the domain or range of the gradient mapping is specified by $|\mathcal{X}| - 1$ coordinates. Using this formalism, Appendix B also shows that the gradient of (11) evaluated at the origin is $\nabla\varphi_{\mathcal{H}_h}(0; r_i) = r_i$.

Let us introduce the arithmetic average of the $\varphi_{\mathcal{H}_h}(\lambda; r_i)$'s:

$$\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty}) \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathcal{H}_h}(\lambda; r_i), \tag{12}$$

and let $\bar{r} \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} r_i$. The assumption of the theorems to be presented shortly is that the aforementioned properties of $\varphi_{\mathcal{H}_h}(\lambda; r_i)$ carry over to $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ after taking the arithmetic average:

ASSUMPTION $\mathcal{A}$. *For $h = 0, 1$, function $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ is finite, strictly convex and twice continuously differentiable throughout $\Re^{|\mathcal{X}|-1}$, with $\psi_{\mathcal{H}_h}(0; r_{1:\infty}) = 0$. Its gradient defines a mapping $\nabla\psi_{\mathcal{H}_h} : \Re^{|\mathcal{X}|-1} \mapsto \mathcal{P}(\mathcal{X})$, with $\nabla\psi_{\mathcal{H}_h}(0; r_{1:\infty}) = \bar{r}$.*

One important example in which Assumption $\mathcal{A}$ is easily verified is when the infinite sequence of probability distributions $r_{1:\infty}$ contains only a finite number of different elements, in which case the arithmetic average in (12) reduces to a finite sum. Note also that strict convexity of $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ always follows by the analogous property of $\varphi_{\mathcal{H}_h}(\lambda; r_i)$ because infinite positively-weighted sums of strictly convex functions preserve strict convexity [40].

The Legendre transform of $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty})$ is [41]:

$$\Psi_{\mathcal{H}_h}(\omega; r_{1:\infty}) = \sup_{\lambda\in\Re^{|\mathcal{X}|-1}} \left\{ \sum_{x\in\mathcal{X}'} \lambda(x)\omega(x) - \psi_{\mathcal{H}_h}(\lambda; r_{1:\infty}) \right\}, \tag{13}$$

where $\omega \in \mathcal{P}(\mathcal{X})$. In the next definition we use the notation $\Omega(\alpha)$ as an abbreviation for $\Omega(\alpha; p_{1:\infty}, q_{1:\infty})$.

DEFINITION (Error Exponent for Unlabeled Data)*: For $0 < \alpha < \infty$, let:*

$$\Omega(\alpha) \triangleq \inf_{\omega\in\mathcal{P}(\mathcal{X}):\, \Psi_{\mathcal{H}_0}(\omega; q_{1:\infty}) < \alpha} \Psi_{\mathcal{H}_1}(\omega; p_{1:\infty}). \tag{14}$$

THEOREM 1 *(Properties of $\Omega(\alpha)$) Suppose that Assumption $\mathcal{A}$ is verified. The error exponent $\Omega(\alpha)$ for unlabeled detection is continuous and convex for $\alpha > 0$, takes the value*

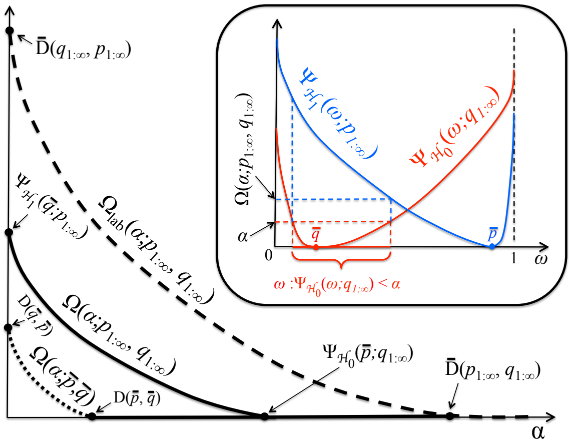
$\bar{\mathrm{D}}(q_{1:\infty}, p_{1:\infty})$
$\Psi_{\mathcal{H}_1}(\bar q; p_{1:\infty})$
$\Omega_{\text{lab}}(\alpha; p_{1:\infty}, q_{1:\infty})$
$\mathrm{D}(\bar q, \bar p)$
$\Omega(\alpha; p_{1:\infty}, q_{1:\infty})$
$\Omega(\alpha; \bar p, \bar q)$
$\mathrm{D}(\bar p, \bar q)$
$\Psi_{\mathcal{H}_0}(\bar p; q_{1:\infty})$
$\bar{\mathrm{D}}(p_{1:\infty}, q_{1:\infty})$
α
$\Psi_{\mathcal{H}_1}(\omega; p_{1:\infty})$
$\Psi_{\mathcal{H}_0}(\omega; q_{1:\infty})$
$\Omega(\alpha; p_{1:\infty}, q_{1:\infty})$
α
0
$\bar q$
$\bar p$
1
ω
$\omega : \Psi_{\mathcal{H}_0}(\omega; q_{1:\infty}) < \alpha$

Fig. 1. Error exponent for unlabeled detection $\Omega(\alpha; p_{1:\infty}, q_{1:\infty})$, see (14). Also shown are the upper bound $\Omega_{\text{lab}}(\alpha; p_{1:\infty}, q_{1:\infty})$ in (5), and the lower bound $\Omega(\alpha; \bar p, \bar q)$. The inset shows how $\Omega(\alpha; p_{1:\infty}, q_{1:\infty})$ is computed from $\Psi_{\mathcal{H}_1}(\omega; p_{1:\infty})$ and $\Psi_{\mathcal{H}_0}(\omega; q_{1:\infty})$, for the case $|\mathcal{X}| = 2$ with ω scalar.

$\Omega(0) = \Psi_{\mathcal{H}_1}(\bar q; p_{1:\infty})$ at the origin, is strictly decreasing over the interval $0 < \alpha < \Psi_{\mathcal{H}_0}(\bar p; q_{1:\infty})$, and is identically zero for $\alpha \geq \Psi_{\mathcal{H}_0}(\bar p; q_{1:\infty})$. In addition, for all $\alpha > 0$,

$$\Omega(\alpha; p_{1:\infty}, q_{1:\infty}) \quad \begin{cases} \leq \Omega_{\text{lab}}(\alpha; p_{1:\infty}, q_{1:\infty}), \\ \geq \Omega(\alpha; \bar p, q_{1:\infty}), \\ \geq \Omega(\alpha; p_{1:\infty}, \bar q), \\ \geq \Omega(\alpha; \bar p, \bar q). \end{cases} \tag{15}$$

When $r_{1:\infty}$ is the constant sequence $(\bar r, \bar r, \dots)$, we have $\Psi_{\mathcal{H}_h}(\omega; \bar r) = D(\omega||\bar r)$, $h = 0, 1$, and the quantities in (15) simplify accordingly. For instance: $\Omega(\alpha; \bar p, \bar q) = \inf_{\omega \in \mathcal{P}(\mathcal{X}) :\, D(\omega||\bar q) < \alpha} D(\omega||\bar p)$.

Proof: The proof is given in Appendix B. •

Our main theoretical result is contained in the following theorem, which provides the operational meaning of $\Omega(\alpha)$ and extends Proposition 1 to the case of unlabeled detection.

THEOREM 2 (Unlabeled Detection) *Consider the hypothesis test with unlabeled data formalized in (9). Suppose that Assumption $\mathcal{A}$ is verified, and let $0 < \alpha < \infty$.*

a) For any closed acceptance region $E \subseteq \mathcal{P}(\mathcal{X})$ for $\mathcal{H}_0$:

$$\begin{aligned} &\liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_0(t_{\mathbf{X}^n} \notin E) \geq \alpha \\ &\Rightarrow \limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1(t_{\mathbf{X}^n} \in E) \leq \Omega(\alpha). \end{aligned} \tag{16}$$

b) Setting $E^* = \{\omega \in \mathcal{P}(\mathcal{X}) : \Psi_{\mathcal{H}_0}(\omega; q_{1:\infty}) \leq \alpha\}$, we get

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_0(t_{\mathbf{X}^n} \notin E^*) \geq \alpha, \tag{17a}$$

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1(t_{\mathbf{X}^n} \in E^*) = \Omega(\alpha). \tag{17b}$$

Proof: The proof is given in Appendix C. •

Note that the asymptotically optimal region of part b) does not require knowledge of the sequence $p_{1:\infty}$. This suggests the possibility of considering universal unlabeled detection problems, in which the distributions under $\mathcal{H}_1$ are not known. These extensions are out of the scope of the present article.

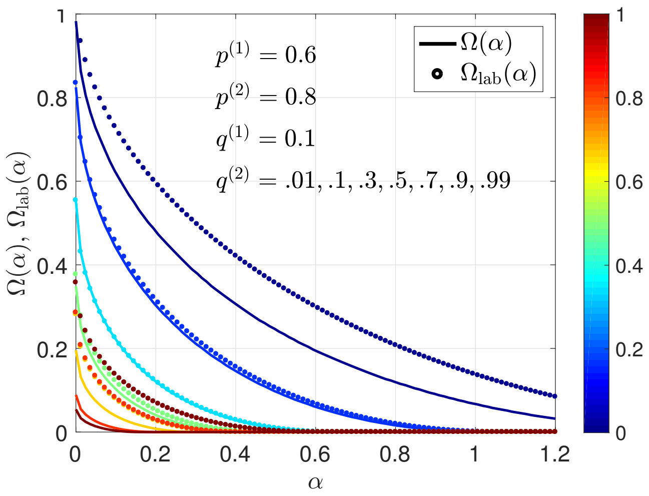

Fig. 2. Error exponents for labeled and unlabeled detection in the case of half-and-half binary observations described in the main text. We know that $\Omega_{\text{lab}}(\alpha) \geq \Omega(\alpha)$ and the difference between the two error exponents represents the information for detection contained in the data labels. There exist cases (e.g., $q^{(2)} = 0.1$, light blue) in which the difference is minimal, as well as cases ($q^{(2)} = 0.3$, cyan) in which there is no difference at all.

The interpretation of Theorem 2 is similar to the interpretation of Proposition 1: With unlabeled data there exist tests with type I error $\doteq e^{-n\alpha}$ and type II error $\doteq e^{-n\Omega(\alpha)}$, but no stronger pairs of asymptotic expressions can be simultaneously achieved. Figure 1 depicts the typical behavior of the error exponent $\Omega(\alpha)$, and highlights its main properties.

Note by Theorem 1 that $\Omega(\alpha)$ is upper bounded by the error exponent for labeled data, and lower bounded by the exponent obtained when data under either (or both) hypotheses are drawn *iid* according to the average distributions $\bar p$ or $\bar q$. The upper and lower bounds in $\Omega(\alpha; \bar p, \bar q) \leq \Omega(\alpha; p_{1:\infty}, q_{1:\infty}) \leq \Omega_{\text{lab}}(\alpha; p_{1:\infty}, q_{1:\infty})$ coincide when data are *iid* under both hypotheses, as it must be. Note also that $\Psi_{\mathcal{H}_1}(\bar q, p_{1:\infty})$ is the equivalent for unlabeled detection of the exponent $\bar D(q_{1:\infty}||p_{1:\infty})$ appearing in Proposition 2, see Fig. 1.

As an example of application of the theorem, let us consider the binary case $|\mathcal{X}| = 2$, and suppose that under $\mathcal{H}_1$ half the observations are drawn from distribution $(p^{(1)}, 1 - p^{(1)})^T$ and half from $(p^{(2)}, 1 - p^{(2)})^T$, where ^{T} denotes vector transposition. Likewise, under $\mathcal{H}_0$ half the observations are drawn from distribution $(q^{(1)}, 1 - q^{(1)})^T$ and half from $(q^{(2)}, 1 - q^{(2)})^T$. In this case the divergence rates appearing in definition (5) reduce to the balanced sum of only two divergences, and the infimum in (5) is computed over the set $\mathcal{P}(\mathcal{X}^2)$. The error exponents $\Omega(\alpha)$ and $\Omega_{\text{lab}}(\alpha)$ for this detection problem are depicted in Fig. 2, where different values of $q^{(2)}$ are shown with the colors indicated by the color bar. Note that there exist combinations of the parameters for which $\Omega(\alpha)$ and $\Omega_{\text{lab}}(\alpha)$ are very close to each other, and there exist combinations for which the information contained in the labels is very relevant, making $\Omega(\alpha)$ substantially smaller than $\Omega_{\text{lab}}(\alpha)$. The extreme case $\Omega(\alpha) = \Omega_{\text{lab}}(\alpha)$ is also possible. Aside from the obvious *iid* case $p^{(1)} = p^{(2)}$ and $q^{(1)} = q^{(2)}$, this happens, for instance, when $p^{(1)} + q^{(2)} = p^{(2)} + q^{(1)} = 1$, which can be explained by noting that the log-likelihood ratio becomes a function of the *type* of the observed vector, and therefore the optimal unlabeled detector performs as the optimal labeled one.

IV. Practical Algorithms for Unlabeled Detection

Part b) of Theorem 2 gives an explicit expression for the acceptance region of the optimal test. However, this leaves open many practical questions. First, the optimality of the test shown in Theorem 2 is only asymptotic and little can be said on the performance for finite — possibly "small" — values of n. Second and more important, no attention has been paid to the computational complexity required to implement the test. Third, in some applications it is desirable to recover the lost labels. These practical aspects are now addressed by considering specific detectors.

A first detector is introduced by making an analogy with the following detection problem with *labeled* data: $\mathcal{H}_1$: $\widetilde{\mathbf{X}}^n \sim \bar{p}_{1:\infty} = (\bar{p}, \bar{p}, \dots)$ versus $\mathcal{H}_0$: $\widetilde{\mathbf{X}}^n \sim \bar{q}_{1:\infty} = (\bar{q}, \bar{q}, \dots)$, where the entries of $\widetilde{\mathbf{X}}^n$ are now *iid* under both hypotheses. The optimal decision statistic for this test is the log-likelihood ratio $\sum_{x\in\mathcal{X}} n t_{\widetilde{\mathbf{x}}^n}(x) \log \frac{\bar{p}(x)}{\bar{q}(x)}$. For large n, $t_{\widetilde{\mathbf{x}}^n} \approx t_{\mathbf{x}^n}$, in the sense shown in Appendix D. We then propose the following detection statistic[2] for unlabeled data:

$$\sum_{x\in\mathcal{X}} t_{\mathbf{x}^n}(x) \log \frac{\bar{p}(x)}{\bar{q}(x)}, \tag{18}$$

which is referred to as the statistic of the unlabeled log-likelihood ratio (ULR) detector. Were $t_{\widetilde{\mathbf{x}}^n}$ equal to $t_{\mathbf{x}^n}$ the error exponent of the test would be $\Omega(\alpha; \bar{p}, \bar{q})$, which is a lower bound to the optimal performance of unlabeled detection, as shown by Theorem 2. However, closeness of $t_{\mathbf{x}^n}$ to $t_{\widetilde{\mathbf{x}}^n}$ tells nothing about the rate of convergence to zero of the detection errors, and nothing can be anticipated as to the performance of this detector. Its main advantage is the low computational complexity: With the type vector $t_{\mathbf{x}^n}$ available, its implementation only requires $|\mathcal{X}|$ multiplications and $|\mathcal{X}|-1$ additions, independently of n: the complexity is $\mathcal{O}(1)$.

The ULR detector makes no attempts to estimate the labels, and we next present detection algorithms for which, instead, the labels matter. To elaborate, let $\mathcal{X} = \{1, 2, \dots, m\}$ be the observation alphabet, which entails no loss of generality, and let us start from the case in which the detector observes the *labeled* vector $\mathbf{x}^n = (x_1, \dots, x_n)$, see (1). Let $\log p_i(k) - \log q_i(k) \triangleq u_{ki} - v_{ki}$ be the marginal log-likelihood ratio of the i-th observed sample x_i, when $x_i = k$, $k = 1, \dots, m$. Organizing these values in m-by-n matrix form, we have:

$$\begin{pmatrix}
u_{11}-v_{11} & \boldsymbol{u_{12}-v_{12}} & u_{13}-v_{13} & \dots & u_{1n}-v_{1n} \\
u_{21}-v_{21} & u_{22}-v_{22} & u_{23}-v_{23} & \dots & u_{2n}-v_{2n} \\
\boldsymbol{u_{31}-v_{31}} & u_{32}-v_{32} & u_{33}-v_{33} & \dots & \boldsymbol{u_{3n}-v_{3n}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
u_{m1}-v_{m1} & u_{m2}-v_{m2} & \boldsymbol{u_{m3}-v_{m3}} & \dots & u_{mn}-v_{mn}
\end{pmatrix}. \tag{19}$$

The optimal log-likelihood ratio statistic for test (1) is $\mathcal{U} - \mathcal{V}$, where $\mathcal{U} = \sum_{i=1}^n u_{k_i i}$ and $\mathcal{V} = \sum_{i=1}^n v_{k_i i}$, with k_i denoting the value taken by the i-th observation x_i. The statistic $\mathcal{U} - \mathcal{V}$ involves n entries of matrix (19). Precisely, one entry over each column and $n\,t_{\mathbf{x}^n}(k)$ entries over the k-th row. In other words, regarding the above matrix as a trellis (left to right), the optimal log-likelihood statistic for test (1) is obtained by summing the entries belonging to a specific *path* over the trellis (19). For instance, if the observed vector is $\mathbf{x}^n = (3, 1, m, \dots, 3)$, the optimal path contains the entries shown in (19) by boldface symbols.

The point with unlabeled detection is that we do not observe $\mathbf{x}^n$ but only its type $t_{\mathbf{x}^n}$, and the optimal path across the trellis in unknown. Note that the "optimal" test (Bayesian, assuming that all permutations are equally likely) is the ratio of two averaged likelihoods. One is sum of the likelihoods for $\mathcal{H}_1$ over all possible permutations of labels, divided by the number of permutations, and the other is the analogous average of the likelihoods for $\mathcal{H}_0$. That this Bayesian test is infeasible, for any reasonable size of the problem, is self evident.

One possible approach to circumvent the lack of precise knowledge of the optimal path is to resort to the generalized likelihood ratio test (GLRT). The GLRT consists of replacing the unknown labeling by its maximum likelihood estimate under each hypothesis, and then constructing the ratio of the resulting likelihoods [42]. The GLRT is not an optimal test but in many instances may lead to nicely-performing tests amenable to simple implementation, and gives us as by-product an estimate of the permutation under $\mathcal{H}_1$ and under $\mathcal{H}_0$. Thus, after the decision about the hypothesis is made, the pertinent estimate of the labeling is also available.

To see how the GLRT works, consider first the log-likelihood for $\mathcal{H}_1$. This log-likelihood, expressed in matrix/trellis form, is given by matrix (19) with all the v_{ik}'s set to zero. Among all the possible paths across such trellis, the GLRT selects the one yielding the largest sum among all paths compatible with the observed $t_{\mathbf{x}^n}$. The compatible paths are those with one entry per column, and $n\,t_{\mathbf{x}^n}(k)$ entries over the k-row. A convenient way to visualize these paths is to introduce an *augmented* version of the trellis, where the k-th row of (19) is copied $n\,t_{\mathbf{x}^n}(k)$ times, for $k = 1, \dots, m$. This yields the following n by n trellis:

$$\begin{array}{rl}
\begin{array}{r}
n\,t_{\mathbf{x}^n}(1)\text{ copies}\left\{\begin{array}{c}\\ \\ \\ \end{array}\right. \\
n\,t_{\mathbf{x}^n}(2)\text{ copies}\left\{\begin{array}{c}\\ \\ \\ \end{array}\right. \\
\vdots \\
n\,t_{\mathbf{x}^n}(m)\text{ copies}\left\{\begin{array}{c}\\ \\ \\ \end{array}\right.
\end{array}
&
\left(\begin{array}{ccccc}
u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\
u_{21} & u_{22} & u_{23} & \cdots & u_{2n} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
u_{21} & u_{22} & u_{23} & \dots & u_{2n} \\
\vdots & & \vdots & \ddots & \vdots \\
u_{m1} & u_{m2} & u_{m3} & \cdots & u_{mn} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
u_{m1} & u_{m2} & u_{m3} & \dots & u_{mn}
\end{array}\right)
\end{array}. \tag{20}$$

Finding the "GLRT path" across the augmented trellis (20) amounts to select one entry over each row and one entry over each column, with the goal of maximizing the sum of the n selected entries. Let us denote by $\mathcal{U}_{\text{GLRT}}$ this maximum sum. Likewise, for $\mathcal{H}_0$, we consider the trellis similar to (20) with the u_{ki}'s replaced by the v_{ki}'s. The best path over this new trellis must be found[3], and the sum of the corresponding

[2] By detection (or decision) statistic we mean a quantity that, compared to a threshold, leads to a decision in favor of $\mathcal{H}_1$ if the threshold strictly crossed, and in favor of $\mathcal{H}_0$ otherwise.

[3] Of course, the orderings may (and likely will) be completely different under the two hypotheses.

entries is denoted by $\mathcal{V}_{\rm GLRT}$. The GLRT statistic is $\mathcal{U}_{\rm GLRT}-\mathcal{V}_{\rm GLRT}$, and requires to find two optimal paths, which give the estimate of the labels under the two hypotheses.

Finding the best path over these trellises is not a combinatorial problem, because exhaustive search is not necessary. Indeed, the search of the GLRT path across a trellis like that in (20) is an instance of the transportation problem — a special case of the assignment problem — for which efficient algorithms have been developed [43]. In the jargon of the assignment problem, each row of (20) represents a "person", each column represents an "object", and the (k,i)-th entry is the benefit for person k if obtains object i. The problem is to assign one distinct object to each person providing the maximum global benefit.

The Hungarian (a.k.a. Munkres or Munkres-Kuhn) algorithm solves exactly the assignment problem in $\mathcal{O}(n^3)$ operations [44], [45]. The auction method usually has lower complexity and is amenable to parallel implementation. A nice overview of the auction procedure and its applications to data association can be found in [46]. Among the many variants of the auction method, the ϵ-scaled implementation achieves a solution of the assignment problem $n\epsilon$-close to the actual maximum [47]. The computational complexity of the auction algorithm depends on the data structure and when the assignment problem involves *similar persons* (i.e., equal rows, as in our case) it can be highly inefficient [48]. A variation of the auction algorithm specifically tailored to address assignment problems with similar persons and similar objects has been proposed in [48], [49]. The auction algorithm used in this paper is ϵ-scaled and is a special form of that proposed in [48], accounting for the presence of similar persons but not of similar objects. This algorithm is here referred to as "auction-sp".

Aside from the auction-sp algorithm, we also discuss approximate solutions to the assignment problem. While it is clear that many such solutions can be conceived, in the following we only consider two very simple greedy algorithms. Recall the trellis representing the log-likelihood under $\mathcal{H}_1$:

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ u_{21} & u_{22} & u_{23} & \cdots & u_{2n} \\ u_{31} & u_{32} & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & u_{m3} & \cdots & u_{mn} \end{pmatrix}. \tag{21}$$

1) Detector A: Algorithm A processes sequentially the observations encoded in $n\,t_{\mathbf{x}^n}$. Consider the value taken by the first observation. This value identifies a row on the trellis (21), and the column with largest value on that row is selected. Such column is then deleted from the trellis, and so excluded from the successive steps. The second observation is processed in the same way, and so forth up to process all the n observed values. The first contribution $\mathcal{U}_{\rm A}$ to the decision statistic for Detector A is the sum of the entries of (21) identified by this algorithm. By repeating this path search over trellis (21) with the u_{ki}'s replaced by the v_{ki}'s, we obtain the second contribution $\mathcal{V}_{\rm A}$, and the decision statistic for Detector A is given by the difference $\mathcal{U}_{\rm A}-\mathcal{V}_{\rm A}$. Note that the order in which the observations are processed is arbitrary and, indeed, observations are unordered. For instance, one can process the observations in increasing value order. Of course different ordering leads, in general, to different paths.

The computational cost of Algorithm A can be approximately evaluated by noting that the k-th iteration amounts to computing the maximum of a $(n-k+1)$-sized vector of reals, and there are $n-1\approx n$ such iterations. If we assume that computing the maximum over ℓ numbers requires a number of elementary operations proportional to ℓ, an approximate value for the computational cost is proportional to $\sum_{k=1}^{n}(n-k+1) = n(n+1)/2$, namely, the computational complexity of Algorithm A is $\mathcal{O}(n^2)$.

2) Detector B: Algorithm B works as follows. First, regardless of the observations, we select the best path on trellis (21), in the sense of achieving the largest sum of n entries, choosing one entry per column. Then, we proceed to make modifications to such path, up to obtain a path compatible with the observed $t_{\mathbf{x}^n}$. For instance, suppose that row ℓ has been selected too many times and row k too few times. The necessary modification amounts to choosing one of the columns selected on row ℓ, and assigning instead that column to row k. In doing so, among all possible columns currently selected on row ℓ, we choose column i for which $u_{\ell,i}-u_{k,i}$ is minimum, in order to maintain the sum of the selected entries of (21) as large as possible. Column i is now deleted from the trellis and, if needed, a new modification step begins. This procedure eventually leads to a compatible path on the trellis according to Algorithm B, yielding $\mathcal{U}_{\rm B}$. Note that the ordering of the modifications is arbitrary and leads, in general, to different final paths. Likewise, running Algorithm B over the trellis with entries $\{v_{ki}\}$ gives the contribution $\mathcal{V}_{\rm B}$, and the final decision statistic for Detector B is $\mathcal{U}_{\rm B}-\mathcal{V}_{\rm B}$. Algorithm B is shown below in Matlab©-style code.

Algorithm B

```
Input: $L$, $x$
    $L$: $m$-by-$n$ matrix of log-likelihood values
      ($m$ = alphabet size, $n$ = No. of samples)
    $x$: 1-by-$n$ vector with entries $\in \{1,\dots,m\}$
Output: $p$    path over trellis $L$

function $p$ = algorithmB($L, x$)
$[m,n]$ = size($L$);
$[d,p]$ = max($L$, [ ], 1);
$sx$ = sort($x$);
$sp$ = sort($p$);
$g$ = not($sp == sx$);
$ch$ = [$sp(g); sx(g)$];
$[d,j]$ = size($ch$);
$bl$ = false(1, $n$);
for  $i = 1 : j$
    $g$ = find(($p == ch(1,i)$) & not($bl$));
    $[d,k]$ = min($L(ch(1,i)$ * ones(size($g$)) + $m$ * ($g$ − 1)) − ...
            $L(ch(2,i)$ * ones(size($g$)) + $m$ * ($g$ − 1)), [ ], 2);
    $p(g(k)) = ch(2,i)$;
    $bl(g(k))$ = true;
end
```

The computational complexity of Algorithm B can be estimated by considering that the "for" cycle is the part of

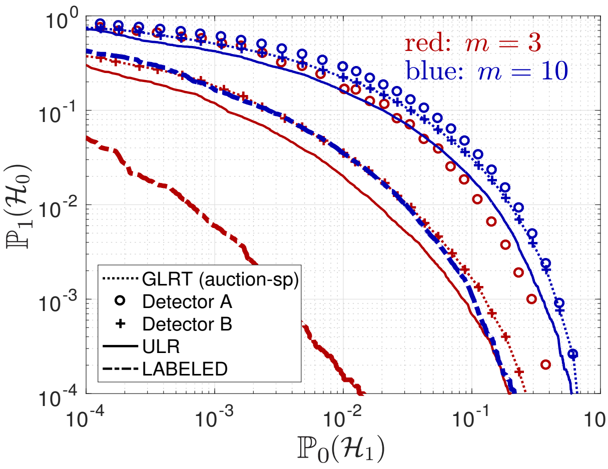

Fig. 3. First computer experiment. Type II error probability versus type I error probability, for $n = 100$ and $m = 3, 10$.

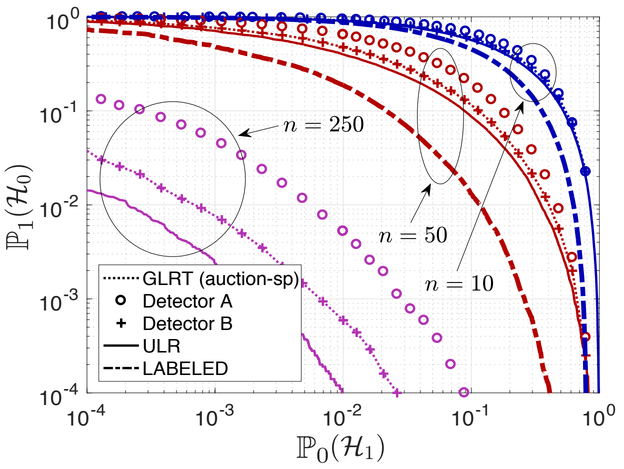

Fig. 4. First computer experiment. Type II error probability versus type I error probability for $m = 5$ and $n = 10, 50, 250$.

the routine that essentially determines the computational cost. In this cycle the minimum over a decreasing-size vector is computed. In the worst case where all the n entries of the initial path must be changed, such vector has size n, and the same argument used for Algorithm A leads to the conclusion that the computational complexity of Algorithm B is $\mathcal{O}(n^2)$. However, the actual number of modifications required is less (and possibly much less) than n, and depends on the realization of $t_{\mathbf{x}}$, on the order in which its entries are processed, and on the trellis values. This implies that the computational complexity of Algorithm B is only upper bounded by $\mathcal{O}(n^2)$, but can be substantially less.

V. Computer Experiments

Let us begin by assuming that data are *iid* under $\mathcal{H}_0$, so that the path search must be performed on a single trellis. In the first computer experiment we assume that under $\mathcal{H}_0$ data are uniformly distributed, namely, using column vector notation for the PMFs, $q_i = (1/m \dots 1/m)^T$, for all $i = 1, \dots, n$. Under $\mathcal{H}_1$ the n PMFs of size m, written as columns of an m-by-n matrix, are as follows:

$$\begin{pmatrix} 0 & \frac{1/m}{n-1} & 2\frac{1/m}{n-1} & & \frac{1}{m} \\ \kappa & \kappa+\frac{1/m-\kappa}{n-1} & \kappa+2\frac{1/m-\kappa}{n-1} & \cdots & \frac{1}{m} \\ 2\kappa & 2\kappa+\frac{1/m-2\kappa}{n-1} & 2\kappa+2\frac{1/m-2\kappa}{n-1} & \cdots & \frac{1}{m} \\ 3\kappa & 3\kappa+\frac{1/m-3\kappa}{n-1} & 3\kappa+2\frac{1/m-3\kappa}{n-1} & \cdots & \frac{1}{m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (m-1)\kappa & (m-1)\kappa+\frac{1/m-(m-1)\kappa}{n-1} & (m-1)\kappa+2\frac{1/m-(m-1)\kappa}{n-1} & \cdots & \frac{1}{m} \end{pmatrix}, \tag{22}$$

where $\kappa = \frac{2}{m(m-1)}$. Thus, the first PMF (leftmost column) is[4] $p_1 = \left(0\, \frac{2}{m(m-1)}\, \frac{4}{m(m-1)} \cdots \frac{2}{m}\right)^T$, the n-th PMF (rightmost) $p_n = \left(\frac{1}{m}\, \frac{1}{m} \cdots \frac{1}{m}\right)^T$ is uniform, and all other columns of (22) are such that the entries on each row vary linearly from the leftmost to the rightmost value (i.e., increase or decrease linearly). Straightforward calculation shows that the entries of the n-averaged PMF $\frac{1}{n}\sum_{i=1}^n p_i$ are

$$\left(\frac{1}{2m}\ \frac{m+1}{2m(m-1)}\ \frac{m+3}{2m(m-1)}\ \frac{m+5}{2m(m-1)} \cdots \frac{m+2(m-1)-1)}{2m(m-1)}\right)^T. \tag{23}$$

Since these values do not depend on n, we have that $\bar{p} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n p_i$ is given by (23).

For this case study, we now investigate the performance of the four detectors presented in Sect. IV: ULR, auction-sp, detector A, and detector B. For the auction-sp algorithm, after trials and errors we found that $\epsilon = 10^{-3}/m$ practically achieves the same total benefit as the Hungarian algorithm, and this value of ϵ is therefore selected in all numerical experiments. In the figures, the results of the auction-sp algorithm are labeled as "GLRT (auction-sp)".

Let $\mathbb{P}_i(\mathcal{H}_h)$ denote the probability of deciding for $\mathcal{H}_h$, under hypothesis $\mathcal{H}_i$, $h, i = 0, 1$. In Figs. 3 and 4 we show the ROC[5] (Receiver Operational Characteristic), namely the type II error $\mathbb{P}_1(\mathcal{H}_0)$ versus the type I error $\mathbb{P}_0(\mathcal{H}_1)$, obtained by 10^5 Monte Carlo runs. Clearly, the lower is the ROC curve, the better is the detection performance. In Fig. 3 we set $n = 100$, and consider two values of the alphabet size $m = 3, 10$. For $m = 3$ we see that detector B outperforms detector A, performs exactly as the GLRT and close to the ULR, which achieves the best performance. For the sake of comparison, we also report the ROC curve for the "labeled" detector, namely, for the case in which the association between data and generating PMFs is perfectly known (no data permutation takes place). As it must be, the labeled detector performs much better. Next, looking at the case $m = 10$ in Fig. 3, we see that the performances of the detectors worsen, and their relative ordering is as for $m = 3$, with a minor gain of detector B over detector A, and a minor loss of the unlabeled detectors over the labeled one.

[4] Note that we have always assumed strictly positive PMFs. Thus, for the sake of rigor, we could replace the zero in (22) with a sufficiently small positive value, and then normalize to unit the first column. This removes the zero and leaves essentially unchanged the arguments and the results that follow.

[5] Actually, the "ROC" curve is the complement of type II error, $1-\mathbb{P}_1(\mathcal{H}_0)$, in function of type I error $\mathbb{P}_0(\mathcal{H}_1)$.

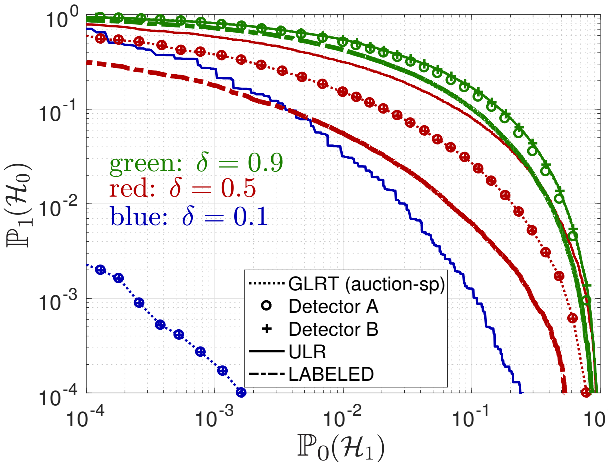

Fig. 5. Second computer experiment. Type II error probability versus type I error probability for $m = 5$, $n = 20$, and three values of δ.

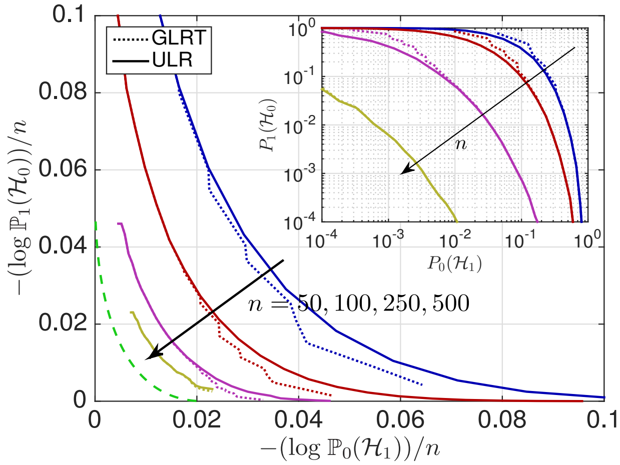

Fig. 6. Error probabilities in the third experiment, with binary observations, $m = 2$, and $n = 50, 100, 250, 500$. Also shown (green, dashed) is the curve $\Omega(\alpha)$ versus α, see (14).

A similar analysis is carried out in Fig. 4, where $m = 5$ and three values of n are considered. We see that by increasing n the detection performance improves. The performance of the labeled detector for $n = 250$ is not shown because the values of $\mathbb{P}_0(\mathcal{H}_1)$ and $\mathbb{P}_1(\mathcal{H}_0)$ fall out of the axis range.

We now consider the computational complexity of the detection algorithms for the addressed example. The precise evaluation of the program execution time is, of course, highly dependent on a number of factors related to the specific hardware and software, and therefore would be of limited interest. We instead report in Table I the ratio between the execution times of the different detectors, which is expected to be less machine/software-dependent. The data in Table I have been obtained by averaging the results of 10^3 Monte Carlo computer runs. As expected, the ULR detector is by far the most efficient. Among the others, it is seen that detector-B is uniformly the less demanding, and the GLRT implemented by auction-sp is the most expensive.

TABLE I
EXECUTION TIME NORMALIZED TO THAT OF THE ULR DETECTOR

GLRT (auction-sp)/ULR

	$\mathcal{H}_0$		$\mathcal{H}_1$	
	$n = 10$	$n = 10^2$	$n = 10$	$n = 10^2$
$m = 5$	1666	13163	1486	11607
$m = 20$	15855	64099	14509	63426

detector-A/ULR

	$\mathcal{H}_0$		$\mathcal{H}_1$	
	$n = 10$	$n = 10^2$	$n = 10$	$n = 10^2$
$m = 5$	65	1052	67	1000
$m = 20$	97	800	111	816

detector-B/ULR

	$\mathcal{H}_0$		$\mathcal{H}_1$	
	$n = 10$	$n = 10^2$	$n = 10$	$n = 10^2$
$m = 5$	32	431	31	434
$m = 20$	71	417	76	433

The relative ordering of the detectors' performance depends on the specific detection problem, and the performance assessment requires a case-by-case analysis. For instance, the superiority of the ULR detector shown in Figs. 3 and 4 is not a general rule, as it is obvious if one considers examples in which $\bar{p} = \bar{q}$. In this case the ULR detector no longer makes sense, as seen in (18). We direct the reader to [50] for more discussion on this issue. A less extreme situation is the second case study, which we now describe. Suppose that under $\mathcal{H}_0$ data are *iid* with common PMF $q_i = (\Delta, 2\Delta, \dots, m\Delta)^T$, where $\Delta = 2/(m(m+1))$, and under $\mathcal{H}_1$ the PMFs are as follows: for some $0 < \delta < 1$, the first $n/2$ distributions have mass $(1-\delta)$ at the first entry, and the remaining $n/2$ have mass $(1-\delta)$ at the last entry. All other entries have mass $\delta/(m-1)$. Figure 5 reports the results of 10^5 Monte Carlo runs, for $m = 5$, $n = 20$, and $\delta = 0.1, 0.5, 0.9$. When $\delta = 0.9$ all detectors perform similarly, but for smaller values of δ the ULR detector is substantially outperformed by the other three detectors. For $\delta = 0.1$ the ULR is extremely poor with respect to its competitors. For $\delta = 0.5, 0.9$ the figure also shows, as benchmark, the ROC of the labeled detector.

As final example, let us consider the case of binary observations, $m = 2$, and let us consider now non-identically distributed data under both hypotheses. Under $\mathcal{H}_0$ we assume that the first $n/2$ data are drawn from distribution $(.5, .5)^T$, and the remaining $n/2$ from $(.3, .7)^T$. Likewise, under $\mathcal{H}_1$, the first half of the data come from the distribution $(.1, .9)^T$, and the other half from $(.9, .1)^T$. The case of binary alphabets has some special features, which are investigated in [50]. In particular, for binary alphabets, GLRT, detector A and detector B, are exactly the same, and we accordingly report a single curve. The inset of Fig. 6 shows that the ULR may be worse than the GLRT for small n, but is essentially equivalent when n grows. In the main plot, the same data of the inset are used but we depict $-\frac{1}{n}\log \mathbb{P}_1(\mathcal{H}_0)$ versus $-\frac{1}{n}\log \mathbb{P}_0(\mathcal{H}_1)$. Note that, in the main plot, for a fixed n, the ordering of the curves is reversed with respect to the inset. On the same axes, for comparison, we also show (lowermost curve, dashed green) the curve $\Omega(\alpha)$, obtained by resolving numerically the convex optimization (14). The theoretical results of this article tell us that the optimal performance converges to $\Omega(\alpha)$.

Benchmark curves are not shown in Fig. 6, because the error probabilities of the labeled benchmark case are very small, and the number of Monte Carlo runs needed to estimating them reliably is much larger than 10^5, which is used the figure.

VI. Conclusions & Future Work

We consider a canonical binary hypothesis test with independent data under both hypotheses. Motivated by modern applications of sensor networks engaged in big data analysis, we assume that the observation vector $\mathbf{X}^n = (X_1, \dots, X_n)$ collected by the peripheral units is delivered to the fusion center in the form of a random *set* $\mathbf{X}_u^n = \{X_1, \dots, X_n\}$, rather than a random vector. Namely, the values of the entries $\{X_i\}_{i=1}^n$ are known to the fusion center, but the positions (*labels*) that these values had in the original vector $\mathbf{X}^n$ are not. The set $\mathbf{X}_u^n$ is also known as the unlabeled version of $\mathbf{X}^n$ and the problem is becoming known as *detection by unlabeled data*.

The theoretical question addressed is how much information for detection is carried by $\mathbf{X}_u^n$. We provide the asymptotic ($n \to \infty$) characterization of the performance of the optimal test in terms of an error exponent rate $\Omega(\alpha)$, which replaces the canonical rate $\Omega_{\text{lab}}(\alpha)$ of the labeled case. It is proven that, when type I error tends to zero as $\exp[-n\alpha]$ with the data size n, type II error may converge to zero as $\exp[-n\Omega(\alpha)]$ but not faster. The rate difference $\Omega_{\text{lab}}(\alpha) - \Omega(\alpha)$ quantifies the loss of information induced by the loss of data labels.

The second part of this paper addresses the practical question of how to solve the test by algorithms of affordable computational complexity and good performance. The ULR detector makes no attempts to estimate the labels and is very efficient computationally. The GLRT solution for unlabeled data boils down to an assignment problem, for which a tailored form of the auction algorithm can be exploited. We also propose two alternative detection algorithms with good trade-off between performance and complexity, as we show by computer experiments.

For future studies it would be interesting to relate the performance and computational cost of the detection algorithms to the statistical distribution of the data, thus providing a theoretical assessment of their relative merits. Large sensor networks with severely quantized data represent the motivating application area for the theory developed in this article, as mentioned in the introduction. Accordingly, the case of binary alphabets deserves special attention. Preliminary investigations in this direction can be found in [50], but a more in-depth analysis is certainly desirable. On the opposite side, generalization of the theoretical results to continuous random variables and designing practical algorithms for the continuous case are two important open problems. Furthermore, in some applications the observation vector undergoes some form of local perturbation of the entry positions (e.g., an entry can be moved only a given number of positions away). These kinds of problems, characterized by more structured perturbations, are interesting subjects for future investigations. Finally, extensions to the case of multi-hypothesis, sequential, and universal tests are also worth of investigation.

Appendix A
Propositions 1 and 2: Sketch of proof

The assertion claimed in Proposition 1, when data are *iid* under both hypotheses, can be found, e.g., in [51]. A sketch of the proof in the case where data are not necessarily identically distributed exploits results from [38], as follows. Let

$$\Omega_{\text{lab}}^{(n)}(\alpha) \triangleq \inf_{\omega_{1:n} \in \mathcal{P}(\mathcal{X}^n):\, \frac{1}{n}\sum_{i=1}^n D(\omega_i \| q_i) \le \alpha} \frac{1}{n} \sum_{i=1}^n D(\omega_i \| p_i), \tag{A.1}$$

so that $\Omega_{\text{lab}}(\alpha) = \lim_{n\to\infty} \Omega_{\text{lab}}^{(n)}(\alpha)$, except for the fact that our definition of $\Omega_{\text{lab}}(\alpha)$ in (5) involves open divergence balls. Suppose that $\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_0(\mathbf{X}^n \notin A_n) \ge \alpha$, and fix a small $\epsilon > 0$. For all sufficiently large n, one has $-\frac{1}{n}\log \mathbb{P}_0(\mathbf{X}^n \notin A_n) \ge \alpha - \epsilon$. For any $\gamma \in (0,1)$, $\alpha - \epsilon \ge \alpha - 2\epsilon - \frac{1}{n}\log\gamma$, provided that n is large enough. Then, Corollary 2 in [38] gives

$$\begin{aligned} &-\frac{1}{n}\log \mathbb{P}_1(\mathbf{X}^n \in A_n) \\ &\quad\le -\frac{1}{n}\log\left(1 - \frac{C(\alpha,\epsilon)}{n\epsilon^2} - \gamma\right) + \Omega_{\text{lab}}^{(n)}(\alpha - 3\epsilon) + \epsilon, \end{aligned} \tag{A.2}$$

where the constant $C(\alpha,\epsilon) \ge 0$ is made explicit in [38] and is of no concern to us. Taking the limit superior, (A.2) implies $\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(\mathbf{X}^n \in A_n) \le \Omega_{\text{lab}}(\alpha)$, because ϵ can be made arbitrarily small, yielding (6).

Part b) follows immediately by Corollary 1 in [38], where it is shown that there exists a sequence A_n^* of acceptance regions for $\mathcal{H}_0$ such that

$$-\frac{1}{n}\log \mathbb{P}_0(\mathbf{X}^n \notin A_n^*) \ge \alpha, \quad -\frac{1}{n}\log \mathbb{P}_1(\mathbf{X}^n \in A_n^*) \ge \Omega_{\text{lab}}^{(n)}(\alpha).$$

The former yields (7a) and, by part a), also implies $\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(\mathbf{X}^n \in A_n^*) \le \Omega_{\text{lab}}(\alpha)$. The latter gives $\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(\mathbf{X}^n \in A_n^*) \ge \Omega_{\text{lab}}(\alpha)$. Combining these two bounds gives (7b).

From the definition shown in (A.1) it is also seen that $\Omega_{\text{lab}}^{(n)}(0) = \frac{1}{n}\sum_{i=1}^n D(q_i \| p_i)$, and $\Omega_{\text{lab}}^{(n)}(\alpha) = 0$ for $\alpha \ge \frac{1}{n}\sum_{i=1}^n D(p_i \| q_i)$. Letting $n \to \infty$, these relationships give $\Omega_{\text{lab}}(0) = \bar{D}(q_{1:\infty} \| p_{1:\infty})$, and $\Omega_{\text{lab}}(\alpha) = 0$ for $\alpha \ge \bar{D}(p_{1:\infty} \| q_{1:\infty})$.

Consider next Proposition 2, which is a version of Chernoff-Stein's Lemma [37] applied to test (1). The proof is based on defining, for a given integer n and some $\epsilon > 0$, the set of "typical" sequences $\mathbf{x}^n$ that verify $\left| \frac{1}{n}\log \frac{q_{1:n}(\mathbf{x}^n)}{p_{1:n}(\mathbf{x}^n)} - \bar{D}(q_{1:\infty} \| p_{1:\infty}) \right| \le \epsilon$. Then, straightforward modifications of the arguments provided in [37, Chap. 11.8] prove the claim.

Appendix B
Proof of Theorem 1

A. Properties of $\varphi_{\mathcal{H}_h}(\lambda; r_i)$ and its Legendre transform

To simplify the notation, in the following we occasionally write $\varphi_{\mathcal{H}_h}(\lambda)$ in place of $\varphi_{\mathcal{H}_h}(\lambda; r_i)$, and similar. Recall the definition $\mathcal{X}' = \mathcal{X} \backslash \{x'\}$, and the convention $\lambda(x') = 0$. Under hypothesis $\mathcal{H}_h$, $h = 0, 1$, let us consider the function $\varphi_{\mathcal{H}_h}(\lambda)$ defined over $\Re^{|\mathcal{X}|-1}$, see (11). The entries of gradient vector $\nabla\varphi_{\mathcal{H}_h}(\lambda)$ are, for $x \in \mathcal{X}'$,

$$[\nabla\varphi_{\mathcal{H}_h}(\lambda)]_x = \frac{r_i(x)e^{\lambda(x)}}{\sum_{y\in\mathcal{X}} r_i(y)e^{\lambda(y)}}. \tag{B.1}$$

Let us regard the gradient vector as a mapping. Its domain is $\Re^{|\mathcal{X}|-1}$ and its range is the convex set $\{\omega \in \Re^{|\mathcal{X}|-1} : \omega(x) > 0, \sum_{x\in\mathcal{X}'} \omega(x) < 1\}$ which is the projection of $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$ onto its $|\mathcal{X}| - 1$ coordinates $\omega(x)$, $x \in \mathcal{X}'$, where $\mathrm{rin}(A)$ denotes the relative interior of set A [40]. By adding the coordinate $\omega(x') = 1 - \sum_{x\in\mathcal{X}'} \omega(x)$, the range of the mapping becomes $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$, namely $\nabla\varphi_{\mathcal{H}_h} : \Re^{|\mathcal{X}|-1} \mapsto \mathrm{rin}(\mathcal{P}(\mathcal{X}))$. Note also that, with this notational convention, $\varphi_{\mathcal{H}_h}(0) = r_i \in \mathrm{rin}(\mathcal{P}(\mathcal{X}))$.

To compute the Hessian matrix $\nabla^2\varphi_{\mathcal{H}_h}(\lambda)$, let us first differentiate (B.1) with respect to $\lambda(x)$, which is the x-th entry of vector λ. This gives $\left[\nabla^2\varphi_{\mathcal{H}_h}(\lambda)\right]_{x,x}$ in the form

$$\frac{r_i(x)e^{\lambda(x)}\left[\sum_{y\in\mathcal{X}} r_i(y)e^{\lambda(y)} - r_i(x)e^{\lambda(x)}\right]}{\left[\sum_{y\in\mathcal{X}} r_i(y)e^{\lambda(y)}\right]^2},$$

where the term in brackets at the numerator can be rewritten as $\sum_{y\in\mathcal{X}\setminus\{x\}} r_i(y)e^{\lambda(y)} = r_i(x') + \sum_{y\in\mathcal{X}'\setminus\{x\}} r_i(y)e^{\lambda(y)}$, because $\lambda(x') = 0$. Differentiating (B.1) with respect to $\lambda(z)$, $z \neq x$, gives the off-diagonal entries, and we get the Hessian matrix as follows:

$$\begin{aligned}\left[\nabla^2\varphi_{\mathcal{H}_h}(\lambda)\right]_{x,z} &= \left[\sum_{y\in\mathcal{X}} r_i(y)e^{\lambda(y)}\right]^{-2} \\ &\times \begin{cases} r_i(x)e^{\lambda(x)}\left[r_i(x') + \sum_{\substack{y\in\mathcal{X}'\\ y\neq x}} r_i(y)e^{\lambda(y)}\right], & x = z, \\ -r_i(x)r_i(z)e^{\lambda(x)}e^{\lambda(z)}, & x \neq z. \end{cases}\end{aligned} \tag{B.2}$$

This shows that $\varphi_{\mathcal{H}_h}(\lambda)$ is twice continuously differentiable throughout $\Re^{|\mathcal{X}|-1}$. Straightforward algebra also shows that the Hessian matrix in (B.2) is strictly diagonally dominant, which implies that it is positive definite [52]. This proves that $\varphi_{\mathcal{H}_h}(\lambda)$ is strictly convex over $\Re^{|\mathcal{X}|-1}$ [40].

A convex function is proper if it is $< \infty$ for at least one point and never takes the value $-\infty$; for a proper convex function, closedness is the same as lower semi-continuity [41]. Thus, the convex function $\varphi_{\mathcal{H}_h}(\lambda)$ is proper and closed because finite and everywhere continuous throughout $\Re^{|\mathcal{X}|-1}$.

Recall that a proper convex function f is *essentially smooth* if [41]: (i) $D_\lambda = \mathrm{in}(\mathrm{dom}(f))$ is nonempty, where $\mathrm{dom}(f)$ is the effective domain (the domain where f is finite) and $\mathrm{in}(\cdot)$ denotes the interior; (ii) f is differentiable throughout D_λ; (iii) $\lim_{k\to\infty} |\nabla f(\lambda_k)| = \infty$, whenever $\lambda_1, \lambda_2, \ldots$ is a sequence in D_λ converging to a boundary point of D_λ. A convex function on $\Re^{|\mathcal{X}|-1}$ which is everywhere finite and differentiable throughout $\Re^{|\mathcal{X}|-1}$ is essentially smooth. Therefore, $\varphi_{\mathcal{H}_h}(\lambda)$ is closed, proper, strictly convex, and essentially smooth, which allows us to invoke the following result, adapted from [41, Th. 26.5]:

THEOREM (Facts from convex analysis): *Let $f(\lambda)$ be a closed proper convex function, and let*

$$F(\omega) = \sup_{\lambda\in\Re^{|\mathcal{X}|-1}} \left\{\sum_{x\in\mathcal{X}'} \lambda(x)\omega(x) - f(\lambda)\right\}. \tag{B.3}$$

Let $D_\lambda = \mathrm{in}(\mathrm{dom}(f))$ and $D_\omega = \mathrm{in}(\mathrm{dom}(F))$. F is strictly convex and essentially smooth on D_ω if and only if f is strictly convex and essentially smooth on D_λ. In this case: (i) (D_ω, F) is the Legendre transformation of (D_λ, f), and vice versa; (ii) the gradient mapping ∇f is continuous and one-to-one from the open convex set D_λ onto the open convex set D_ω; (iii) ∇F is the continuous inverse mapping of ∇f, i.e., $\nabla F = (\nabla f)^{-1}$. Function $F(\omega)$ in (B.3) admits the representation: $F(\omega) = \sum_{x\in\mathcal{X}'} \omega(x)\nabla F(\omega) - f(\nabla F(\omega))$.

Let $\Phi_{\mathcal{H}_h}(\omega)$ be the Legendre transform of $\varphi_{\mathcal{H}_h}(\lambda)$ defined by (B.3). We see that $\mathrm{dom}(\Phi_{\mathcal{H}_h}) = \mathcal{P}(\mathcal{X})$ and $\Phi_{\mathcal{H}_h}(\omega)$ is strictly convex and essentially smooth on $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$. The mapping $\nabla\varphi_{\mathcal{H}_h} : \Re^{|\mathcal{X}|-1} \mapsto \mathrm{rin}(\mathcal{P}(\mathcal{X}))$ is one-to-one, with inverse $\nabla\Phi_{\mathcal{H}_h} = (\nabla\varphi_{\mathcal{H}_h})^{-1} : \mathrm{rin}(\mathcal{P}(\mathcal{X})) \mapsto \Re^{|\mathcal{X}|-1}$.

Assumption $\mathcal{A}$ in Sec. III ensures that all these conclusions apply to the pair $(\psi_{\mathcal{H}_h}(\lambda), \Psi_{\mathcal{H}_h}(\omega))$: Since $\psi_{\mathcal{H}_h}(\lambda)$ is finite, strictly convex and twice continuously differentiable throughout $\Re^{|\mathcal{X}|-1}$, we have that $\mathrm{dom}(\Psi_{\mathcal{H}_h}) = \mathcal{P}(\mathcal{X})$, $\Psi_{\mathcal{H}_h}(\omega)$ is strictly convex and essentially smooth on $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$, and the mapping $\nabla\psi_{\mathcal{H}_h} : \Re^{|\mathcal{X}|-1} \mapsto \mathrm{rin}(\mathcal{P}(\mathcal{X}))$ is one-to-one, with inverse $\nabla\Psi_{\mathcal{H}_h} = (\nabla\psi_{\mathcal{H}_h})^{-1} : \mathrm{rin}(\mathcal{P}(\mathcal{X})) \mapsto \Re^{|\mathcal{X}|-1}$.

Assumption $\mathcal{A}$ also ensures $\nabla\psi_{\mathcal{H}_h}(0) = \bar{r}$, which implies $\nabla\Psi_{\mathcal{H}_h}(\bar{r}) = 0$. Since $\Psi_{\mathcal{H}_h}(\omega)$ is strictly convex, this function has a unique global minimum at $\bar{r} \in \mathrm{rin}(\mathcal{P}(\mathcal{X}))$. By the representation of the Legendre transform given in the above theorem, it is also easily seen that $\Psi_{\mathcal{H}_h}(\bar{r}) = -\psi_{\mathcal{H}_h}(0) = 0$.

B. Properties of $\Omega(\alpha)$

Using the results of the previous section, the properties of the error exponent defined in (14) can be easily derived. First, note that, for $\alpha > 0$:

$$\Omega(\alpha) = \inf_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega)<\alpha} \Psi_{\mathcal{H}_1}(\omega) = \min_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega)\leq\alpha} \Psi_{\mathcal{H}_1}(\omega), \tag{B.4}$$

where the second equality in (B.4) follows by observing that $\Psi_{\mathcal{H}_1}(\omega)$ is continuous on $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$ and the set where the infimum is computed can be replaced by the compact set $\{\omega \in \mathcal{P}(\mathcal{X}) : \Psi_{\mathcal{H}_0}(\omega) \leq \alpha\}$, so that the minimum is attained at some point of this compact set [53].

Now, pick two positive values α_1 and α_2, and let $\omega_1, \omega_2 \in \mathcal{P}(\mathcal{X})$ be the minimizers that attain $\Omega(\alpha_1)$ and $\Omega(\alpha_2)$, respectively. Let $\omega_\theta = \theta\omega_1 + (1 - \theta)\omega_2$, with $0 \leq \theta \leq 1$. Clearly $\omega_\theta \in \mathcal{P}(\mathcal{X})$ because $\mathcal{P}(\mathcal{X})$ is convex. From the convexity of $\Psi_{\mathcal{H}_0}(\omega)$, we have $\Psi_{\mathcal{H}_0}(\omega_\theta) \leq \theta\Psi_{\mathcal{H}_0}(\omega_1) + (1 - \theta)\Psi_{\mathcal{H}_0}(\omega_2) \leq \theta\alpha_1 + (1-\theta)\alpha_2 \triangleq \alpha_\theta$. Thus, $\omega_\theta \in \{\omega \in \mathcal{P}(\mathcal{X}) : \Psi_{\mathcal{H}_0}(\omega) \leq \alpha_\theta\}$. Then, from the convexity of $\Psi_{\mathcal{H}_1}(\omega)$,

$$\begin{aligned}\Omega(\alpha_\theta) &= \min_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega)\leq\alpha_\theta} \Psi_{\mathcal{H}_1}(\omega) \\ &\leq \Psi_{\mathcal{H}_1}(\omega_\theta) \leq \theta\Psi_{\mathcal{H}_1}(\omega_1) + (1 - \theta)\Psi_{\mathcal{H}_1}(\omega_2) \\ &= \theta\Omega(\alpha_1) + (1 - \theta)\Omega(\alpha_2),\end{aligned} \tag{B.5}$$

which proves the convexity of $\Omega(\alpha)$.

It is immediate to see that $\Omega(\alpha)$ is nonincreasing in α. We also see that $\alpha \geq \Psi_{\mathcal{H}_0}(\bar{p}) \Rightarrow \Omega(\alpha) = 0$, because in this case the set in (B.4) where the minimum is computed includes $\bar{p}$, and $\Psi_{\mathcal{H}_1}(\bar{p}) = 0$. At the origin, we set $\Omega(0) = \Psi_{\mathcal{H}_1}(\bar{q})$ by continuity. Combining convexity and the nonincreasing property, we conclude that $\Omega(\alpha)$ is convex and strictly decreasing for $0 < \alpha < \Psi_{\mathcal{H}_0}(\bar{p})$.

Next, by Jensen's inequality

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log\sum_{x\in\mathcal{X}} r_i(x)e^{\lambda(x)} \le \log\sum_{x\in\mathcal{X}}\bar{r}(x)e^{\lambda(x)}, \qquad \text{(B.6)}$$

which proves $\psi_{\mathcal{H}_h}(\lambda; r_{1:\infty}) \le \psi_{\mathcal{H}_h}(\lambda;\bar{r})$, $\forall\lambda\in\Re^{|\mathcal{X}|-1}$. For their Legendre transforms the inequality is reversed, yielding $\Psi_{\mathcal{H}_h}(\omega; r_{1:\infty}) \ge \Psi_{\mathcal{H}_h}(\omega;\bar{r})$, $\forall\omega\in\mathcal{P}(\mathcal{X})$. Noting that $\Psi_{\mathcal{H}_h}(\omega;\bar{r}) = \Phi_{\mathcal{H}_h}(\omega;\bar{r})$, we then get, for $\alpha>0$,

$$\begin{aligned}
\Omega(\alpha; p_{1:\infty}, q_{1:\infty}) &= \inf_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega;q_{1:\infty})<\alpha} \Psi_{\mathcal{H}_1}(\omega; p_{1:\infty}) \\
&\ge \inf_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega;q_{1:\infty})<\alpha} \Psi_{\mathcal{H}_1}(\omega;\bar{p}) \quad [=\Omega(\alpha;\bar{p},q_{1:\infty})] \\
&\ge \inf_{\omega\in\mathcal{P}(\mathcal{X})\,:\,\Psi_{\mathcal{H}_0}(\omega;\bar{q})<\alpha} \Psi_{\mathcal{H}_1}(\omega;\bar{p}) \qquad [=\Omega(\alpha;\bar{p},\bar{q})]
\end{aligned}$$

and similarly one ontains $\Omega(\alpha; p_{1:\infty}, q_{1:\infty}) \ge \Omega(\alpha; p_{1:\infty},\bar{q})$. All the inequalities of (15) are so proved, except the first, which follows by the operational meaning of the two rate error functions $\Omega(\alpha)$ and $\Omega_{\mathrm{lab}}(\alpha)$ provided by Proposition 1 and Theorem 2.

Finally, we show that $\Psi_{\mathcal{H}_h}(\omega;\bar{r}) = D(\omega\|\bar{r})$, $h=0,1$. For any $\omega\in\mathcal{P}(\mathcal{X})$, any $\lambda\in\Re^{|\mathcal{X}|-1}$, and $\lambda(x')=0$:

$$\begin{aligned}
\psi_{\mathcal{H}_h}(\lambda;\bar{r}) &= \log\sum_{x\in\mathcal{X}}\bar{r}(x)e^{\lambda(x)} = \log\sum_{x\in\mathcal{X}}\omega(x)\frac{\bar{r}(x)e^{\lambda(x)}}{\omega(x)} \\
&\ge \sum_{x\in\mathcal{X}}\omega(x)\log\frac{\bar{r}(x)e^{\lambda(x)}}{\omega(x)} = \sum_{x\in\mathcal{X}'}\omega(x)\lambda(x) - D(\omega\|\bar{r}),
\end{aligned}$$

yielding $\sum_{x\in\mathcal{X}'}\omega(x)\lambda(x) - \psi_{\mathcal{H}_h}(\lambda;\bar{r}) \le D(\omega\|\bar{r})$. If there exists a vector $\lambda\in\Re^{|\mathcal{X}|-1}$ that attains this upper bound, then $\Psi_{\mathcal{H}_h}(\omega;\bar{r}) = D(\omega\|\bar{r})$ because of the definition of Legendre transform, see (B.3). Direct substitution shows that such vector is $\lambda(x) = \log\frac{\omega(x)\,\bar{r}(x')}{\bar{r}(x)\,\omega(x')}$, $x\in\mathcal{X}'$.

Appendix C
Proof of Theorem 2

The type vector $t_{\mathbf{X}^n}(x)$, $x\in\mathcal{X}$, defined in (10) contains only $|\mathcal{X}|-1$ independent components. Here we work with the *reduced* type vector $t'_{\mathbf{X}^n}$ obtained by deleting the entry $t_{\mathbf{X}^n}(x')$ from $t_{\mathbf{X}^n}$. The excluded value $x'\in\mathcal{X}$ is arbitrary. Accordingly, let us introduce the set $\mathcal{Q}(\mathcal{X}') = \{\omega(x), x\in\mathcal{X}' : \omega\in\mathcal{P}(\mathcal{X})\}$ of probability vectors $\omega\in\mathcal{P}(\mathcal{X})$ from which the entry $\omega(x')$ is deleted. For notational simplicity we loosely use the same symbol ω to denote vectors in $\mathcal{P}(\mathcal{X})$, vectors in $\mathcal{Q}(\mathcal{X}')$, and vectors in $\Re^{|\mathcal{X}|-1}$; the context should avoid any confusion. Also, as done in Appendix B, we occasionally omit to make explicit the dependence of the various functions on the underlying statistical distributions.

There exists an obvious one-to-one correspondence between $(|\mathcal{X}|-1)$-vectors in $\mathcal{Q}(\mathcal{X}')$ and $|\mathcal{X}|$-vectors in $\mathcal{P}(\mathcal{X})$, as well as between reduced type $t'_{\mathbf{X}^n}$ and type $t_{\mathbf{X}^n}$. Thus, the event $\{t'_{\mathbf{X}^n}\in E'\}$ is the same of $\{t_{\mathbf{X}^n}\in E\}$, provided that $E\subseteq\mathcal{P}(\mathcal{X})$ is the element that corresponds to $E'\subseteq\mathcal{Q}(\mathcal{X}')$.

Let $\mathcal{H}_h$ be the hypothesis in force, and $\mathbb{E}_h$ the expectation operator under $\mathcal{H}_h$, $h=0,1$. Recall that the distribution of X_i under $\mathcal{H}_h$ is denoted by $r_i\in\mathcal{P}(\mathcal{X})$, with $i=1,2,\dots$. Let $\lambda\in\Re^{|\mathcal{X}|-1}$, and consider the logarithmic moment generating function of the reduced type vector $t'_{\mathbf{X}^n}$:

$$\begin{aligned}
\Lambda_{\mathcal{H}_h,n}(\lambda) &\triangleq \log\mathbb{E}_h\exp\left\{\sum_{x\in\mathcal{X}'}\lambda(x)t'_{\mathbf{X}^n}(x)\right\} \\
&= \sum_{i=1}^{n}\log\sum_{x\in\mathcal{X}} r_i(x)e^{\frac{\lambda(x)}{n}}, \qquad \text{(C.1)}
\end{aligned}$$

where, we recall, $\lambda(x')=0$ by convention. Note that $\lim_{n\to\infty}\frac{1}{n}\Lambda_{\mathcal{H}_h,n}(n\lambda)$ is exactly the function $\psi_{\mathcal{H}_h}(\lambda)$ defined in (12). Its Legendre transform $\Psi_{\mathcal{H}_h}(\omega)$ is defined in (13) for $\omega\in\mathcal{P}(\mathcal{X})$. We use the same symbol $\Psi_{\mathcal{H}_h}(\omega)$ to denote the Legendre transform of $\psi_{\mathcal{H}_h}(\lambda)$ as function of the $(|\mathcal{X}|-1)$-vector $\omega\in\mathcal{Q}(\mathcal{X}')$, in which case the definition is extended to all $\Re^{|\mathcal{X}|-1}$ by setting $\Psi_{\mathcal{H}_h}(\omega)=\infty$ for $\omega\in\Re^{|\mathcal{X}|-1}\setminus\mathcal{Q}(\mathcal{X}')$. The context should avoid any confusion.

The proof of Theorem 2 is based on the following version of Gärtner-Ellis theorem, see [54, Th. 3.2.6] or [55, Th. V.6].

Theorem (Gärtner-Ellis) *Suppose that the function $\psi_{\mathcal{H}_h}(\lambda)$ in (12) is finite and differentiable throughout $\Re^{|\mathcal{X}|-1}$. Then for any set $A'\subseteq\Re^{|\mathcal{X}|-1}$ we have the large deviation principle:*

$$\begin{aligned}
&\inf_{\omega\in\mathrm{cl}(A')}\Psi_{\mathcal{H}_h}(\omega) \le \liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_h(t'_{\mathbf{X}^n}\in A') \\
&\quad\le \limsup_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_h(t'_{\mathbf{X}^n}\in A') \le \inf_{\omega\in\mathrm{in}(A')}\Psi_{\mathcal{H}_h}(\omega). \qquad \text{(C.2)}
\end{aligned}$$

In (C.2) and in what follows $\mathrm{in}(A')$ and $\mathrm{cl}(A')$ denote the interior and the closure of A', respectively. The complement of A' will be denoted by $\overline{A'}$. These operations are relative to $\Re^{|\mathcal{X}|-1}$.

By Assumption $\mathcal{A}$, $\psi_{\mathcal{H}_h}(\lambda)$ in (12) is finite and differentiable in $\Re^{|\mathcal{X}|-1}$, which allows us to apply the Gärtner-Ellis theorem. Let $E\subseteq\mathcal{P}(\mathcal{X})$ be an arbitrary closed acceptance region for $\mathcal{H}_0$ and let E' be the corresponding closed set in $\mathcal{Q}(\mathcal{X}')$. Note that $\overline{E'}$ is an open set $\subseteq\Re^{|\mathcal{X}|-1}$. Under $\mathcal{H}_0$:

$$\begin{aligned}
\liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_0(t_{\mathbf{X}^n}\in\overline{E}) &= \liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_0(t'_{\mathbf{X}^n}\in\overline{E'}) \\
&\le \limsup_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_0(t'_{\mathbf{X}^n}\in\overline{E'}) \qquad \text{(C.3a)} \\
&\le \inf_{\omega\in\overline{E'}\subseteq\Re^{|\mathcal{X}|-1}}\Psi_{\mathcal{H}_0}(\omega) \qquad \text{(C.3b)} \\
&\le \Psi_{\mathcal{H}_0}(\omega), \quad \forall\omega\in\overline{E'}\subseteq\Re^{|\mathcal{X}|-1}, \qquad \text{(C.3c)}
\end{aligned}$$

where (C.3b) follows by the upper bound in (C.2) for the open set $\overline{E'}$.

For $\alpha>0$, let us suppose $\liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{P}_0(t_{\mathbf{X}^n}\in\overline{E})\ge\alpha$. From (C.3c) this implies $\Psi_{\mathcal{H}_0}(\omega)\ge\alpha$, $\forall\omega\in\overline{E'}\subseteq\Re^{|\mathcal{X}|-1}$, so that $\Psi_{\mathcal{H}_0}(\omega)<\alpha \Rightarrow \omega\in E'\subseteq\mathcal{Q}(\mathcal{X}')$, and therefore

$$C'_\alpha \triangleq \{\omega\in\mathcal{Q}(\mathcal{X}') : \Psi_{\mathcal{H}_0}(\omega)<\alpha\}\subseteq E'. \qquad \text{(C.4)}$$

Under $\mathcal{H}_1$, we have:

$$\begin{aligned}
&\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(t_{\mathbf{X}^n}\in E) = \limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(t'_{\mathbf{X}^n}\in E') \\
&\le \limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(t'_{\mathbf{X}^n}\in C'_\alpha) \qquad \text{(C.5a)} \\
&\le \inf_{\omega\in \mathrm{in}(C'_\alpha)\subseteq \mathcal{Q}(\mathcal{X}')} \Psi_{\mathcal{H}_1}(\omega) \qquad \text{(C.5b)} \\
&= \inf_{\omega\in C'_\alpha\subseteq \mathcal{Q}(\mathcal{X}')} \Psi_{\mathcal{H}_1}(\omega) = \inf_{\omega\in \mathcal{P}(\mathcal{X}):\Psi_{\mathcal{H}_0}(\omega)<\alpha} \Psi_{\mathcal{H}_1}(\omega), \qquad \text{(C.5c)}
\end{aligned}$$

where (C.5a) follows by $C'_\alpha \subseteq E'$, (C.5b) follows by the upper bound in (C.2), and the first equality in (C.5c) is obtained by the continuity of $\Psi_{\mathcal{H}_1}(\omega)$ on $\mathrm{in}(\mathcal{Q}(\mathcal{X}'))$. This proves part a).

To prove part b), let us set $E'^* = \mathrm{cl}(C'_\alpha) = \{\omega \in \mathcal{Q}(\mathcal{X}') : \Psi_{\mathcal{H}_0}(\omega) \le \alpha\}$, and let E^* be the corresponding set in $\mathcal{P}(\mathcal{X})$. Under $\mathcal{H}_0$:

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_0(t_{\mathbf{X}^n}\in \overline{E^*}) = \liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_0(t'_{\mathbf{X}^n}\in \overline{E'^*}) \\
&\ge \inf_{\omega\in \Re^{|\mathcal{X}|-1}:\Psi_{\mathcal{H}_0}(\omega)\ge\alpha} \Psi_{\mathcal{H}_0}(\omega) = \inf_{\omega\in \mathcal{Q}(\mathcal{X}'):\Psi_{\mathcal{H}_0}(\omega)\ge\alpha} \Psi_{\mathcal{H}_0}(\omega) \qquad \text{(C.6a)} \\
&= \inf_{\omega\in \mathcal{P}(\mathcal{X}):\Psi_{\mathcal{H}_0}(\omega)\ge\alpha} \Psi_{\mathcal{H}_0}(\omega) \ge \alpha, \qquad \text{(C.6b)}
\end{aligned}$$

where the inequality in (C.6a) is the lower bound in (C.2), and the equality in (C.6a) can be verified by considering separately the two cases: α such that $\{\omega \in \mathcal{Q}(\mathcal{X}') : \Psi_{\mathcal{H}_0}(\omega) \ge \alpha\} = \emptyset$ (the infimum over the empty set being ∞ by definition), and $\neq \emptyset$. This proves (17a). Finally, under $\mathcal{H}_1$, note that

$$\inf_{\omega\in \mathrm{in}(E'^*)} \Psi_{\mathcal{H}_1}(\omega) = \inf_{\omega\in \mathrm{cl}(\mathrm{in}(E'^*))} \Psi_{\mathcal{H}_1}(\omega) = \inf_{\omega\in \mathrm{cl}(E'^*)} \Psi_{\mathcal{H}_1}(\omega), \qquad \text{(C.7)}$$

where the first equality follows by the continuity of $\Psi_{\mathcal{H}_1}(\omega)$ on $\mathrm{in}(\mathcal{Q}(\mathcal{X}'))$ and the second follows by $\mathrm{cl}(E'^*) = \mathrm{cl}(\mathrm{in}(E'^*))$. From (C.7) we see that the lower and the upper bounds in (C.2) coincide, and the large deviation principle gives a precise limit:

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(t_{\mathbf{X}^n}\in E^*) = \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1(t'_{\mathbf{X}^n}\in E'^*) \\
&= \inf_{\omega\in \mathcal{Q}(\mathcal{X}'):\Psi_{\mathcal{H}_0}(\omega)\le\alpha} \Psi_{\mathcal{H}_1}(\omega) \qquad \text{(C.8a)} \\
&= \inf_{\omega\in \mathcal{P}(\mathcal{X}):\Psi_{\mathcal{H}_0}(\omega)\le\alpha} \Psi_{\mathcal{H}_1}(\omega) = \inf_{\omega\in \mathcal{P}(\mathcal{X}):\Psi_{\mathcal{H}_0}(\omega)<\alpha} \Psi_{\mathcal{H}_1}(\omega), \qquad \text{(C.8b)}
\end{aligned}$$

where (C.8a) follows by (C.2) and (C.7), while the second equality in (C.8b) follows by the continuity of $\Psi_{\mathcal{H}_1}(\omega)$ on $\mathrm{rin}(\mathcal{P}(\mathcal{X}))$.

Appendix D
Asymptotics of $t_{\mathbf{X}^n}$

Let $\widetilde{\mathbf{X}}^n = (\widetilde{X}_1, \dots, \widetilde{X}_n)$, with the entries $\widetilde{X}_i$ drawn *iid* from $\bar{r}$, and let $t_{\widetilde{\mathbf{X}}^n}$ be the corresponding type. For $t_{\widetilde{\mathbf{X}}^n}$ the standard strong law of large numbers [56, Th. 22.1] gives, $\forall x \in \mathcal{X}$, $t_{\widetilde{\mathbf{X}}^n}(x) \to \bar{r}(x)$ with probability one. As to $t_{\mathbf{X}^n}$ — the type when data are drawn from $r_{1:n}$ — note that $\mathrm{VAR}_h[\mathbb{I}(X_i = x)] = r_i(x)(1 - r_i(x))$, where VAR_h denotes the variance computed under $\mathcal{H}_h$, and $\sum_{i=1}^{\infty} \mathrm{VAR}_h[\mathbb{I}(X_i = x)]/i^2 \le \sum_{i=1}^{\infty} i^{-2}/4 = \pi^2/24 < \infty$. This implies the following convergence with probability one [57, Th. 1.14]:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = x) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_h[\mathbb{I}(X_i = x)] \to 0, \qquad \text{(D.1)}$$

from which we see that $t_{\mathbf{X}^n}(x) \to \bar{r}(x)$ exactly as does $t_{\widetilde{\mathbf{X}}^n}(x)$. By triangular inequality, for any $\epsilon > 0$, and all sufficiently large n, $|t_{\mathbf{X}^n}(x) - t_{\widetilde{\mathbf{X}}^n}(x)| < \epsilon$ with probability one.

References

[1] S. Marano and P. Willett, “Sometimes they come back: Testing two simple hypotheses (in the realm of unlabeled data),” in *Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, Alberta, Canada, April 15-20 2018.

[2] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, “Unlabeled sensing with random linear measurements,” *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3237–3253, May 2018.

[3] ——, “Unlabeled sensing: Solving a linear system with unordered measurements,” in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 786–793.

[4] G. Wang, J. Zhu, R. S. Blum, P. Willett, S. Marano, V. Matta, and P. Braca, “Signal amplitude estimation and detection from unlabeled binary quantized samples,” *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4291–4303, Aug. 2018.

[5] J. Zhu, H. Cao, C. Song, and Z. Xu, “Parameter estimation via unlabeled sensing using distributed sensors,” *IEEE Communications Letters*, vol. 21, no. 10, pp. 2130–2133, Oct 2017.

[6] S. Marano, V. Matta, P. Willett, P. Braca, and R. Blum, “Hypothesis testing in the presence of Maxwell’s daemon: Signal detection by unlabeled observations,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, 5-9 Mar. 2017.

[7] R. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Artech House, 2007.

[8] ——, “Statistics 101 for multisensor, multitarget data fusions,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 19, no. 1, pp. 53–64, Jan. 2004.

[9] T. E. Humphreys, B. M. Ledvina, M. L. Psiaki, B. W. O’Hanlon, and P. M. Kintner, Jr., “Assessing the spoofing threat: Development of a portable GPS civilian spoofer,” in *2016 IEEE Conference on Communications and Network Security (CNS)*, Savanna, GA, Sep. 16-19 2008, pp. 2314–2325.

[10] Q. Zeng, H. Li, and L. Qian, “GPS spoofing attack on time synchronization in wireless networks and detection scheme design,” in *2012 IEEE Military Communications Conf. (MILCOM 2012)*, Oct 2012, pp. 1–5.

[11] P. Pradhan, K. Nagananda, P. Venkitasubramaniam, S. Kishore, and R. S. Blum, “GPS spoofing attack characterization and detection in smart grids,” in *2016 IEEE Conference on Communications and Network Security (CNS)*, Oct 2016, pp. 391–395.

[12] Z. Zhang, S. Gong, A. D. Dimitrovski, and H. Li, “Time synchronization attack in smart grid: Impact and analysis,” *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 87–98, March 2013.

[13] A. B. Poore and S. Gadaleta, “Some assignment problems arising from multiple target tracking,” *Mathematical and Computer Modelling*, vol. 43, no. 9, pp. 1074–1091, 2006.

[14] L. Schenato, “Optimal estimation in networked control systems subject to random delay and packet drop,” *IEEE Transactions on Automatic Control*, vol. 53, no. 5, pp. 1311–1317, June 2008.

[15] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, “Adaptive filtering of imprecisely time-stamped measurements with application to AIS networks,” in *2015 18th International Conference on Information Fusion (Fusion)*, July 2015, pp. 359–365.

[16] L. Keller, M. J. Siavoshani, C. Fragouli, K. Argyraki, and S. Diggavi, “Identity aware sensor networks,” in *Proc. of the 26th IEEE International Conference on Computer Communications (INFOCOM 2009)*, Rio De Janeiro, Brazil, April, 19-25 2009, pp. 2177–2185.

[17] P. David, D. Dementhon, R. Duraiswami, and H. Samet., “SoftPOSIT: Simultaneous pose and correspondence determination,” *International Journal of Computer Vision*, vol. 59, no. 3, pp. 259–284, 2004.

[18] M. Marques, M. Stošić, and J. Costeira, “Subspace matching: Unique solution to point matching with geometric constraints,” in *Proc. of the 2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1288–1294.

[19] X. Huang and A. Madan, “CAP3: A DNA sequence assembly program,” *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[20] W. S. Robinson, “A method for chronologically ordering archaeological deposits,” *American Antiquity*, pp. 293–301, 1951.

[21] T. Nakano, A. W. Eckford, and T. Haraguchi, *Molecular Communication*. UK: Cambridge University Press, 2013.

[22] W. Haselmayr, N. Varshney, A. T. Asyhari, A. Springer, and W. Guo. (2018, Jul.) On the impact of transposition errors in diffusion-based channels. [Online]. Available: arXiv:1701.02971v2

[23] L. J. Schulman and D. Zuckerman, "Asymptotically good codes correcting insertions, deletions, and transpositions," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2552–2557, 1999.

[24] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 452–468, Jan. 2015.

[25] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, "On credibility estimation tradeoffs in assured social sensing," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1026–1037, Jun. 2013.

[26] D. Wang, L. Kaplan, and T. Abdelzaher, "Maximum likelihood analysis of conflicting observations in social sensing," *ACM Trans. on Sensor Networks*, vol. 10, no. 2, Jan. 2014.

[27] S. Marano, V. Matta, and P. Willett, "The importance of being earnest: Social sensing with unknown agent quality," *IEEE Trans. Signal and Info. Process. over Networks*, vol. 2, no. 3, pp. 306–320, Sep. 2016.

[28] S. Marano and P. Willett, "Profiling agents in networks by simply knowing how to count," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 2, pp. 629–641, Apr. 2018.

[29] Y. C. Eldar, *Sampling Theory, Beyond Bandlimited Systems*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2015.

[30] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1040–1044.

[31] S. Thrun and J. J. Leonard, "Simultaneous localization and mapping," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Heidelberg: Springer Berlin, 2008, pp. 871–889.

[32] S. Haghighatshoar and G. Caire, "Signal recovery from unlabeled samples," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1242–1257, March 2018.

[33] A. Abid, A. Poon, and J. Zou. (2017, May 4) Linear regression with shuffled labels. [Online]. Available: http://arxiv.org/abs/1705.01342

[34] A. Pananjady, M. J. Wainwright, and T. A. Courtade. (2017, April 24) Denoising linear models with permutated data. [Online]. Available: http://arxiv.org/abs/1704.07461

[35] ——, "Linear regression with an unknown permutation: Statistical and computational limits," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 417–424.

[36] Z. Liu and J. Zhu, "Signal detection from unlabeled ordered samples," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2431–2434, Dec. 2018.

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New Jersey, USA: Wiley-Interscience, 2006.

[38] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 405–417, Jul. 1974.

[39] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

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

[41] R. T. Rockafellar, *Convex analysis*. Princeton, NJ: Princeton University Press, 1970.

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Englewood Cliffs, New Jersey: Prentice Hall, 1998.

[43] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[44] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Log. Q.*, vol. 2, pp. 83–97, 1955.

[45] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5(1), pp. 32–38, 1957.

[46] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[47] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*. Belmont, MA: Athena Scientific, 1998.

[48] D. P. Bertsekas and D. A. Castanon, "The auction algorithm for the transportation problem," *Annals of Operations Research*, vol. 20, pp. 67–96, 1989.

[49] R. Jonker and A. Volgenant, "Improving the Hungarian assignment algorithm," *Operations Research Letters*, vol. 5, pp. 171–175, 1986.

[50] S. Marano and P. Willett, "Making decisions with shuffled bits," in *Proc. of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, 12–17 May, 2019, *submitted.*

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

[52] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge University Press, 1985.

[53] V. A. Zorich, *Mathematical Analysis I*. Berlin: Springer, 2004.

[54] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer, 1998.

[55] F. den Hollander, *Large Deviations*, ser. Fields Institute Monographs. Providence, Rhode Island: American Mathematical Society, 2000.

[56] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley-Interscience, 1995.

[57] H. Shao, *Mathematical Statistics*, 2nd ed. Springer, 2003.