Asymptotically Optimal Decision Rules for Joint Detection and Source Coding

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

The problem of joint detection and lossless source coding is considered. We derive asymptotically optimal decision rules for deciding whether or not a sequence of observations has emerged from a desired information source, and to compress it if has. In particular, our decision rules asymptotically minimize the cost of compression in the case that the data has been classified as 'desirable', subject to given constraints on the two kinds of the probability of error. In another version of this performance criterion, the constraint on the false alarm probability is replaced by the a constraint on the cost of compression in the false alarm event. We then analyze the asymptotic performance of these decision rules and demonstrate that they may exhibit certain phase transitions. We also derive universal decision rules for the case where the underlying sources (under either hypothesis or both) are unknown, and training sequences from each source may or may not be available. Finally, we discuss how our framework can be extended in several directions.

Index Terms: Error exponent, hypothesis testing, false alarm, misdetection, source coding, universal schemes.

1 Introduction

Classical hypothesis testing theory, based on the Neyman–Pearson theorem (see, e.g., [2, Sect. 11.7]), provides the optimal rule for deciding between two hypotheses concerning the distribution or density of a given observation or sequence of observations. It tells us that best trade-off between the two kinds of probability of error is achieved by the likelihood ratio test.

In certain situations, however, this decision between the two hypotheses might be only one of the tasks to be carried out. For example, consider a scenario where under hypothesis $\mathcal{H}_0$, the sequence of observations that we receive is just pure noise, which contains no useful information that may interest us, whereas under hypothesis $\mathcal{H}_1$, the data that we have at hand has emerged from a desirable information source, and in this case, further processing is called for, such as lossless or lossy data compression, parameter estimation [9], [10], channel decoding [7], [11], [12], encryption, further classification, etc.

The straightforward approach to this problem would be to first apply Neyman–Pearson hypothesis testing, and then, if hypothesis $\mathcal{H}_1$ is accepted, perform the corresponding task using the best strategy available. This approach *separates* between optimal decision and the optimality of the subsequent task. A more sophisticated approach, however, is to solve the two problems jointly, namely, to devise a decision rule that takes into account also the cost of the subsequent task (in case it is to be carried out), and on the other hand, optimize the strategy of the following task, taking into account that the data belongs to the decision region of $\mathcal{H}_1$.

For the case where the second task is Bayesian parameter estimation, Moustakides [9] and Moustakides *et al.* [10] have derived an optimal solution for the combined problem. In particular, in these articles, the problem of joint detection and estimation was posed and solved under the criterion of minimizing the conditional expected cost of the estimation error, given that the data is classified into $\mathcal{H}_1$ subject to certain constraints on the false alarm (FA) and misdetection (MD) probabilities (or related constraints). The optimal decision rule, under this criterion, is interesting, but it turns out to be rather complicated and non–trivial in three respects: (i) the proof of optimality is quite long and not easy, (ii) the insight behind this decision rule is not obvious, and (iii) it may be difficult to implement.

In this paper, we propose a modified criterion,[1] which is asymptotically equivalent for a large number of observations, at least in the relevant regime, where the MD probability is constrained to tend to zero. The point of this modification in the criterion is that it allows us to use a slightly extended version of the Neyman–Pearson lemma in order to derive the optimal decision rule in a fairly simple and easy manner. It is also rather easy to implement, or at least to approximate by an easily implementable decision rule. Finally, the intuition behind this decision rule is easier to grasp. We focus, in this paper, on memoryless sources and on the case where the second task to be performed, after the detection, is lossless data compression, but this should be considered only as an example, as the methodology proposed is applicable for a wide variety of tasks, as will be discussed. In fact, the same methodology has already been used in [7], where under $\mathcal{H}_1$, the observed data is the output of a noisy channel fed by a codeword, and the second task after detection is channel decoding, with application to (slotted) asynchronous communication (see also [11] and [12] for earlier work).

In addition to the derivation of the optimal decision rule under our modified criterion, we also analyze its performance in terms of asymptotic exponents. One of our findings is that these asymptotic exponents may exhibit "phase transitions" in the sense of having discontinuous derivatives as functions of the parameters of the problem. Such phase transitions do not occur in the ordinary Neyman–Pearson decision rule. Finally, we derive universal versions of our decision rule that are suitable for scenarios where at least one of the probability distributions (under $\mathcal{H}_0$ and/or $\mathcal{H}_1$) is unknown (yet they are still known to be memoryless), and we might have access to a training sequence from one of the sources or both. We also discuss, as mentioned earlier, how our method applies to tasks other than lossless source coding as well as more general classes of sources.

The outline of the remaining part of this paper is as follows. In Section 2, we establish notation conventions and define the problem in several different versions. In Section 3, we present the above–mentioned extension of the Neyman–Pearson lemma. In Section 4, we apply this lemma to the solution of one version of the joint detection and compression problem, and in Section 5 we analyze its performance and discuss it. In Section 6, we show how to apply Lemma 1 to a number of other variants of the problem. Section 7 is devoted to universal decision rules, and finally, in Section 8 we conclude.

[1] Details will follow in the sequel.

2 Notation Conventions and Problem Formulation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. In this paper, $\boldsymbol{X}$ emerges from either one of two sources, P_0 or P_1. The probability of an event $\mathcal{E}$ under P_i will be denoted by $P_i(\mathcal{E})$ and the expectation operator w.r.t. P_i will be denoted by $\boldsymbol{E}_i\{\cdot\}$, $i = 0, 1$. The entropy of a generic distribution Q on $\mathcal{X}$ will be denoted by $H(Q)$. The notation $H(P_i)$ will be shortened to H_i, $i = 0, 1$. For two positive sequences a_n and b_n, the notation $a_n \doteq b_n$ will stand for equality in the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{E\}$. The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies $\hat{P}_{\boldsymbol{x}}(x)$ of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$. The type class of $\boldsymbol{x} \in \mathcal{X}^n$, denoted $\mathcal{T}_{\boldsymbol{x}}$, is the set of all vectors $\boldsymbol{x}'$ with $\hat{P}_{\boldsymbol{x}'} = \hat{P}_{\boldsymbol{x}}$. When we wish to emphasize the dependence of the type class on the empirical distribution $\hat{P}$, we denote it by $\mathcal{T}(\hat{P})$.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sequence of random variables drawn from a finite alphabet memoryless source. There are two hypotheses concerning the probability distribution of the underlying source: Under hypothesis $\mathcal{H}_0$, the source is $P_0 = \{P_0(x),\ x \in \mathcal{X}\}$, whereas under hypothesis $\mathcal{H}_1$, the source is $P_1 = \{P_1(x),\ x \in \mathcal{X}\}$. The source P_0 designates unwanted data (e.g., pure noise, spam, meaningless or unimportant data), while the source P_1 represents useful, desirable information, which we would like to keep for further processing. In this paper, this further processing is lossless data compression (source coding).

A decision rule is a partition of $\mathcal{X}^n$, the space of source vectors of length n, into two complementary regions $\Omega \subseteq \mathcal{X}^n$ and $\Omega^c = \mathcal{X}^n \setminus \Omega$, where Ω is the region where we accept $\boldsymbol{X}$ as having emerged from P_1, and Ω^c is the region where we classify it as having been generated by P_0. Thus, only source vectors that fall in Ω are to be compressed. Since the decision rule is fully defined by the choice of the subset Ω, we will sometimes use expressions like "the decision rule Ω" as shorthand for "the decision rule associated with Ω," with a slight abuse of formal preciseness.

Our aim is to find a decision rule and a compression strategy that jointly optimize the compression performance within Ω subject to constraints on the error probabilities of the two kinds: $P_0(\Omega)$ – the probability of false alarm (FA), and $P_1(\Omega^c)$ – the probability of misdetection (MD). In particular, let $L : \mathcal{X}^n \to \{0, 1, 2, \ldots\}$ be a length function of a lossless fixed–to–variable length code that satisfies Kraft's inequality

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} 2^{-L(\boldsymbol{x})} \leq 1. \tag{1}$$

A seemingly natural goal (in the spirit of [10]) would be to solve the problem:

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{E}_1\{L(\boldsymbol{X})|\boldsymbol{X} \in \Omega\} \\ \text{subject to} \quad & P_0(\Omega) \leq \epsilon_{\text{FA}} \\ & P_1(\Omega^c) \leq \epsilon_{\text{MD}} \end{aligned} \tag{2}$$

where the minimization is over the length function $L(\cdot)$ and the choice of Ω, and where ϵ_{FA} and ϵ_{MD} are prescribed numbers designating the maximum tolerable FA and MD probabilities, respectively. Of course, ϵ_{FA} and ϵ_{MD} should not be chosen both too small, otherwise, the two constraints may become contradictory (the minimum achievable ϵ_{MD} for a given ϵ_{FA} is achieved by the performance of the ordinary likelihood ratio test).

Now, it makes sense to let ϵ_{MD} and ϵ_{FA} decay exponentially with n. We let then $\epsilon_{\text{MD}} = \exp(-nE_{\text{MD}})$ and $\epsilon_{\text{FA}} = \exp(-nE_{\text{FA}})$, where E_{MD} and E_{FA} are positive constants, independent of n. In this regime, $P_1(\Omega) \geq 1 - \exp(-nE_{\text{MD}})$ tends to unity, and so, the conditioning on $\boldsymbol{X} \in \Omega$, that appears in the objective function of (2) has an asymptotically vanishing effect, as $P_1(\boldsymbol{x}|\boldsymbol{x} \in \Omega) = P_1(\boldsymbol{x})/P_1(\Omega) \approx P_1(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$. This means that the best achievable compression performance in the sense of (2) is roughly the entropy of $\boldsymbol{X}$ under P_1, essentially independently of the choice of Ω, whenever $E_{\text{MD}} > 0$, which makes (2) somewhat less interesting than it might seem at first glance.

It is therefore more interesting to examine objective functions with stronger sensitivity to the choice of Ω. This would be the case with a large deviations criterion, like $P_1\{L(\boldsymbol{X}) \geq nR|\boldsymbol{X} \in \Omega\}$, or the related criterion of the exponential moment, $\boldsymbol{E}_1[\exp\{\theta L(\boldsymbol{X})\}|\boldsymbol{X} \in \Omega]$, where $\theta > 0$ is a given parameter. These objective functions are not new and they are interesting on their own right (see, e.g., [6, Introduction] for a discussion on the motivation). In another version of our problem, we will replace the FA constraint $P_0(\Omega) \leq \exp(-nE_{\text{FA}})$, by a constraint on the cost of compression in the FA event, namely, a constraint on $P_0\{L(\boldsymbol{X}) \geq nR|\boldsymbol{X} \in \Omega\}$ or $\boldsymbol{E}_0[\exp\{\theta L(\boldsymbol{X})\}|\boldsymbol{X} \in \Omega]$. In this paper, we focus on these performance criteria, as well as on issues of universality, that is, how to confront uncertainty in P_0 and/or P_1. When dealing with these universality issues, we will find it more convenient to switch the roles between the objective function and one of the constraints, for example, minimize $P_1(\Omega^c)$ subject to constraints on $P_0(\Omega)$ and $P_1\{L(\boldsymbol{X}) \geq nR|\boldsymbol{X} \in \Omega\}$ or on $\boldsymbol{E}_1[\exp\{\theta L(\boldsymbol{X})\}|\boldsymbol{X} \in \Omega]$.

3 Preliminaries: A Simple Extension of the Neyman–Pearson Lemma

The following lemma will turn out to be useful for our purposes (see also [7] for a similar lemma).

Lemma 1 *Let f, g and h be any three functions from $\mathcal{X}^n$ to $\mathbb{R}$ and let*

$$\Omega_\star = \{\boldsymbol{x} :\ f(\boldsymbol{x}) + a \cdot g(\boldsymbol{x}) \leq b \cdot h(\boldsymbol{x})\}, \tag{3}$$

where $a \geq 0$ and $b \geq 0$ are fixed numbers. Let Ω be any other subset of $\mathcal{X}^n$. If

$$\sum_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x}) \leq \sum_{\boldsymbol{x}\in\Omega_\star} g(\boldsymbol{x}) \tag{4}$$

and

$$\sum_{\boldsymbol{x}\in\Omega^c} h(\boldsymbol{x}) \leq \sum_{\boldsymbol{x}\in\Omega_\star^c} h(\boldsymbol{x}) \tag{5}$$

then

$$\sum_{\boldsymbol{x}\in\Omega_\star} f(\boldsymbol{x}) \leq \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}). \tag{6}$$

The lemma tells us that the decision rule defined by $\Omega_\star$ is optimal in the sense that no other competing rule Ω gives strictly smaller values of all three quantities, $\sum_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x})$, $\sum_{\boldsymbol{x}\in\Omega^c} h(\boldsymbol{x})$, and $\sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$. The paramaters a and b can be thought of as Lagrange multipliers that control the magnitudes of $\sum_{\boldsymbol{x}\in\Omega_\star} g(\boldsymbol{x})$ and $\sum_{\boldsymbol{x}\in\Omega_\star^c} h(\boldsymbol{x})$. Note that Lemma 1 (similarly as the classical Neyman–Pearson lemma) does not require f, g and h to be probability distributions. These can be any functions from $\mathcal{X}^n$ to $\mathbb{R}$, in fact, not necessarily even positive functions.

Proof. Let $\Omega_\star$ be defined as in Theorem 1 and let Ω be any competing decision rule. First, observe that for every $\boldsymbol{x} \in \mathcal{X}^n$

$$[\mathcal{I}\{\boldsymbol{x} \in \Omega_\star\} - \mathcal{I}\{\boldsymbol{x} \in \Omega\}] \cdot [b \cdot h(\boldsymbol{x}) - a \cdot g(\boldsymbol{x}) - f(\boldsymbol{x})] \geq 0. \tag{7}$$

This is true because, by definition of $\Omega_\star$, the two factors of the product at the left–hand side (l.h.s.) are either both non–positive or both non–negative. Thus, taking the summation over all $\boldsymbol{x} \in \mathcal{X}^n$, we have:

$$b \cdot \left[\sum_{\boldsymbol{x}\in\Omega_\star} h(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x})\right] - a \cdot \left[\sum_{\boldsymbol{x}\in\Omega_\star} g(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x})\right] - \left[\sum_{\boldsymbol{x}\in\Omega_\star} f(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})\right] \geq 0 \tag{8}$$

or, equivalently,

$$\sum_{\boldsymbol{x}\in\Omega_\star} f(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \leq a \cdot \left[\sum_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega_\star} g(\boldsymbol{x})\right] + b \cdot \left[\sum_{\boldsymbol{x}\in\Omega^c} h(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega_\star^c} h(\boldsymbol{x})\right]. \tag{9}$$

Since $a \geq 0$ and $b \geq 0$, then

$$\sum_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega_\star} g(\boldsymbol{x}) \leq 0 \tag{10}$$

and

$$\sum_{\boldsymbol{x}\in\Omega^c} h(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega_\star^c} h(\boldsymbol{x}) \leq 0 \tag{11}$$

imply

$$\sum_{\boldsymbol{x}\in\Omega_\star} f(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \leq 0, \tag{12}$$

which completes the proof of Lemma 1.

4 Applying Lemma 1 to Joint Detection and Compression

Lemma 1 is almost applicable for solving one version of the problem defined in Section 2. A simple modification will make it completely applicable. First, concerning the constraints, it is clear that the assignments should be $g(\boldsymbol{x}) = P_0(\boldsymbol{x})$ and $h(\boldsymbol{x}) = P_1(\boldsymbol{x})$, for the case of a constraint on $P_0(\Omega)$. Regarding the objective function, for a given choice of Ω, the minimization of $\boldsymbol{E}_1\{\exp[\theta L(\boldsymbol{X})|\boldsymbol{X} \in \Omega\}$ over all uniquely decodable length functions, $L(\cdot)$, gives (ignoring integer length constraints):

$$L^*(\boldsymbol{x}) = -\log\left[\frac{[P_1(\boldsymbol{x})]^{1/(1+\theta)}}{\sum_{\boldsymbol{x}'\in\Omega}[P_1(\boldsymbol{x}')]^{1/(1+\theta)}}\right], \qquad \boldsymbol{x} \in \Omega \tag{13}$$

which yields

$$\boldsymbol{E}_1\{\exp[\theta L^*(\boldsymbol{X})]|\boldsymbol{X}\in\Omega\}=\left(\sum_{\boldsymbol{x}\in\Omega}\left[\frac{P_1(\boldsymbol{x})}{P_1(\Omega)}\right]^{1/(1+\theta)}\right)^{1+\theta}. \tag{14}$$

Thus, the minimization of $\boldsymbol{E}_1\{\exp[\theta L^*(\boldsymbol{X})]|\boldsymbol{X}\in\Omega\}$ over Ω is equivalent to the minimization of

$$\sum_{\boldsymbol{x}\in\Omega}\left[\frac{P_1(\boldsymbol{x})}{P_1(\Omega)}\right]^{1/(1+\theta)}.$$

It is tempting now to use Lemma 1 with the additional assignment

$$f(\boldsymbol{x})=\left[\frac{P_1(\boldsymbol{x})}{P_1(\Omega)}\right]^{1/(1+\theta)}, \tag{15}$$

but this is not quite a legitimate choice for using Lemma 1, since this function depends on Ω. Nonetheless, as observed in Section 2, in the regime where $P_1(\Omega^c)\geq 1-\exp(-nE_{\text{MD}})\to 1$, the factor $P_1(\Omega)$ has an asymptotically negligible effect, and we can uniformly approximate by choosing

$$f(\boldsymbol{x})=[P_1(\boldsymbol{x})]^{1/(1+\theta)}. \tag{16}$$

Also, in order for the coefficients a and b to influence the asymptotic exponents of the objective function and the constraints, we let them be exponential functions of n, i.e., $a=e^{n\alpha}$ and $b=e^{n\beta}$, where α and β are fixed real numbers, independent of n, which are dictated by E_{FA} and E_{MD}. The asymptotically optimal decision rule now reads

$$\Omega_\star=\{\boldsymbol{x}:\ [P_1(\boldsymbol{x})]^{1/(1+\theta)}+e^{n\alpha}P_0(\boldsymbol{x})\leq e^{n\beta}P_1(\boldsymbol{x})\}. \tag{17}$$

5 Discussion and Analysis of the Decision Rule

Let us now examine the decision rule $\Omega_\star$, defined in eq. (17). Since

$$\begin{aligned}\max\left\{[P_1(\boldsymbol{x})]^{1/(1+\theta)},e^{n\alpha}P_0(\boldsymbol{x})\right\} &\leq [P_1(\boldsymbol{x})]^{1/(1+\theta)}+e^{n\alpha}P_0(\boldsymbol{x}) &(18)\\ &\leq 2\cdot\max\left\{[P_1(\boldsymbol{x})]^{1/(1+\theta)},e^{n\alpha}P_0(\boldsymbol{x})\right\}, &(19)\end{aligned}$$

the performance of $\Omega_\star$ is asymptotically equivalent (in terms of asymptotic exponents of the objective function, the FA probability and the MD probability) to that of

$$\hat{\Omega}\ \stackrel{\triangle}{=}\ \left\{\boldsymbol{x}:\ \max\{[P_1(\boldsymbol{x})]^{1/(1+\theta)},e^{n\alpha}P_0(\boldsymbol{x})\}\leq e^{n\beta}P_1(\boldsymbol{x})\right\}$$

$$
\begin{aligned}
&= \{\boldsymbol{x}:\ [P_1(\boldsymbol{x})]^{1/(1+\theta)} \le e^{n\beta}P_1(\boldsymbol{x}),\ e^{n\alpha}P_0(\boldsymbol{x}) \le e^{n\beta}P_1(\boldsymbol{x})\} \\
&= \left\{\boldsymbol{x}:\ -\ln P_1(\boldsymbol{x}) \le n\beta\left(1+\frac{1}{\theta}\right),\ \ln\left[\frac{P_1(\boldsymbol{x})}{P_0(\boldsymbol{x})}\right] \ge n(\alpha-\beta)\right\} \qquad (20)
\end{aligned}
$$

The form of $\hat{\Omega}$ is more convenient than that of $\Omega_\star$, both for understanding the behavior, and for implementation (since it allows passage to the logarithmic domain as is shown in the last line of eq. (20)). The test $\hat{\Omega}$ can be thought of as a combination of two tests: (i) the test $-\ln P_1(\boldsymbol{x}) \le n\beta(1+1/\theta)$, which guarantees that the code–length associated with $\boldsymbol{x}$ is small enough, and (ii) the test $\ln[P_1(\boldsymbol{x})/P_0(\boldsymbol{x})] \ge n(\alpha-\beta)$, which is the ordinary likelihood ratio test that distinguishes between P_0 and P_1. The test $\hat{\Omega}$ also lends itself more conveniently to standard asymptotic exponent analysis using the method of types [3]. The results are as follows.

Consider the MD probability first.

$$
P_1(\Omega_\star^c) \doteq P_1(\hat{\Omega}^c) \doteq \exp\{-ne_{\mathrm{MD}}\} \qquad (21)
$$

where

$$
\begin{aligned}
e_{\mathrm{MD}} &= \min_Q\{\mathcal{D}(Q\|P_1):\ \boldsymbol{E}_Q \ln P_1(X) \le -\beta(1+1/\theta) \text{ or } \boldsymbol{E}_Q \ln[P_1(X)/P_0(X)] \le \alpha-\beta\} \qquad (22) \\
&= \min\{e_1(\beta), e_2(\alpha-\beta)\} \qquad (23)
\end{aligned}
$$

with

$$
e_1(\beta) = \min_Q\{\mathcal{D}(Q\|P_1):\ \boldsymbol{E}_Q \ln P_1(X) \le -\beta(1+1/\theta)\} \qquad (24)
$$

and

$$
e_2(\alpha-\beta) = \min_Q\{\mathcal{D}(Q\|P_1):\ \boldsymbol{E}_Q \ln[P_1(X)/P_0(X)] \le \alpha-\beta\}. \qquad (25)
$$

Here $\boldsymbol{E}_Q\{\cdot\}$ denotes the expectation operator w.r.t. a generic probability distribution Q on $\mathcal{X}$ and $\mathcal{D}(Q\|P)$ is the Kullback–Leibler divergence between Q and P. Both $e_1(\beta)$ and $e_2(\alpha-\beta)$ must be no smaller than E_{MD}. Both minimization problems can easily be solved using Lagrange multipliers. The minimizing Q for e_1 is of the form

$$
Q_1(x) = \frac{[P_1(x)]^\lambda}{\sum_{x'\in\mathcal{X}}[P_1(x')]^\lambda}, \qquad \lambda \le 1 \qquad (26)
$$

where λ is chosen to satisfy the constraint $\boldsymbol{E}_Q \ln P_1(X) \le -\beta(1+1/\theta)$. Clearly, $e_1(\beta)$ is a monotonically increasing function, and due to its convexity, strictly so in the range where it is non–zero and finite, which is $\theta H_1/(1+\theta) < \beta \leq -\theta[\ln \min_x P_1(x)]/(1+\theta)$, H_1 being the entropy of P_1. Thus, we must choose $\beta \geq e_1^{-1}(E_{\rm MD})$. Similarly, the minimzing Q for e_2 is of the form

$$Q_2(x) = \frac{[P_0(x)]^{\nu}[P_1(x)]^{1-\nu}}{Z(\nu)} \tag{27}$$

where $\nu \geq 0$ is chosen to satisfy the constraint $\boldsymbol{E}_Q \ln[P_1(X)/P_0(X)] \leq \alpha - \beta$ and $Z(\nu)$ is a normalization constant. The convex function e_2 is strictly decreasing in $\alpha - \beta$ in the range where it is positive and finite, $\min_x \ln[P_1(x)/P_0(x)] \leq \alpha - \beta < \mathcal{D}(P_1\|P_0)$. Thus, we must choose $\alpha - \beta \leq e_2^{-1}(E_{\rm MD})$. Clearly, once we have selected some $\beta \geq e_1^{-1}(E_{\rm MD})$, the best choice of α (that would maximally shrink $\Omega_\star$, or $\hat{\Omega}$) would be the maximum allowed value, $\alpha = \beta + e_2^{-1}(E_{\rm MD})$. The choice of β will then be dictated by the FA constraint. This simple observation reduces the original space of trade-offs with two degrees of freedom (α and β) to one degree of freedom (β only).

Concerning the FA probability, we have

$$P_0(\Omega_\star) \doteq P_0(\hat{\Omega}) \doteq \exp\{-ne_{\rm FA}\} \tag{28}$$

where

$$\begin{aligned} e_{\rm FA} &= \min_Q\{\mathcal{D}(Q\|P_0) : -\boldsymbol{E}_Q \ln P_1(X) \leq \beta(1+1/\theta), \boldsymbol{E}_Q \ln[P_0(X)/P_1(X)] \leq \beta - \alpha\} \\ &= \min_Q\{\mathcal{D}(Q\|P_0) : -\boldsymbol{E}_Q \ln P_1(X) \leq \beta(1+1/\theta), \boldsymbol{E}_Q \ln[P_0(X)/P_1(X)] \leq -e_2^{-1}(E_{\rm MD})\}. \end{aligned}$$

Similarly as before, the minimizing Q, denoted Q^*, is of the form

$$Q^*(x) = \frac{[P_0(x)]^{1-\eta}[P_1(x)]^{\eta+\xi}}{Z(\eta,\xi)}, \tag{29}$$

where $Z(\eta,\xi)$ is a normalization constant, and where $\eta \geq 0$ and $\xi \geq 0$ are chosen to satisfy the constraints, $-\boldsymbol{E}_Q \ln P_1(X) \leq \beta(1+1/\theta)$ and $\boldsymbol{E}_Q \ln[P_0(X)/P_1(X)] \leq -e_2^{-1}(E_{\rm MD})$. Here, $e_{\rm FA}$ is a decreasing function of β, and so, the constraint $e_{\rm FA}(\beta) \geq E_{\rm FA}$ dictates the choice $\beta \leq e_{\rm FA}^{-1}(E_{\rm FA})$, which is feasible (in view of the earlier MD exponent analysis) provided that $e_{\rm FA}^{-1}(E_{\rm FA}) \geq e_1^{-1}(E_{\rm MD})$. Under this condition, it is possible to assign

$$\alpha = e_{\rm FA}^{-1}(E_{\rm FA}) + e_2^{-1}(E_{\rm MD}) \tag{30}$$

and

$$\beta = e_{\rm FA}^{-1}(E_{\rm FA}). \tag{31}$$

Finally, using the method of types once again, the exponent associated with the objective function is given by

$$\boldsymbol{E}_1[\exp\{\theta L^*(\boldsymbol{X})\}|\boldsymbol{X}\in\Omega_\star] = \left(\sum_{\boldsymbol{x}\in\hat{\Omega}}[P_1(\boldsymbol{x})]^{1/(1+\theta)}\right)^{1+\theta} \doteq \exp\{ne_c\}, \tag{32}$$

where

$$e_c = \max_Q\{\theta H(Q) - \mathcal{D}(Q\|P_1):\ \boldsymbol{E}_Q \ln P_1(X) \geq -\beta(1+1/\theta),\ \boldsymbol{E}_Q \ln[P_1(X)/P_0(X)] \geq \alpha-\beta\}, \tag{33}$$

with α and β as in eqs. (30) and (31), and with $H(Q)$ being the entropy associated with a distribution Q on $\mathcal{X}$. Once again, this is a convex programming problem that can be solved using Lagrange multipliers. This completes the analysis of asymptotic exponents associated with $\Omega_\star$.

As α, β and θ vary, it is expected that these exponents may exhibit certain phase transitions, because of possible abrupt passages between regions where one of the constraints is active to regions where the other one becomes active (or both). The following is a simple example that demonstrates this point.

Example. Let $\mathcal{X}=\{0,1\}$, define P_0 to be the binary symmetric source (BSS) and let P_1 be defined by $P_1(1) = 1-P_1(0) = 3/4$. In this case, it is straightforward to verify that $\hat{\Omega}$ is the set of all source vectors $\{\boldsymbol{x}\}$ for which the relative frequency of 1's is at least as large as

$$q_{\alpha,\beta} = \frac{\max\{\ln 4 - \beta(1+1/\theta), \ln 2 + \alpha - \beta\}}{\ln 3}. \tag{34}$$

As long as $q_{\alpha,\beta}\in(1/2,3/4)$, the error exponents are simply

$$e_{\text{FA}} = D\left(q_{\alpha,\beta}\|\frac{1}{2}\right), \quad e_{\text{MD}} = D\left(q_{\alpha,\beta}\|\frac{3}{4}\right), \tag{35}$$

where for $s,t\in[0,1]$, $D(s\|t)$ denotes the binary divergence, i.e., $D(s\|t) = s\ln(s/t)+(1-s)\ln[(1-s)/(1-t)]$. It is assumed, of course, that E_{FA} and E_{MD} are small enough such that there exist α and β with $D(q_{\alpha,\beta}\|\frac{1}{2})\geq E_{\text{FA}}$ and $D(q_{\alpha,\beta}\|\frac{3}{4})\geq E_{\text{MD}}$. The derivatives of the exponents e_{FA} and e_{MD}, as functions of α, β and θ, are discontinuous at the points where

$$\ln 4 - \beta\left(1+\frac{1}{\theta}\right) = \ln 2 + \alpha - \beta, \tag{36}$$

because at these points, the achiever of the maximum on the r.h.s. of (34) switches between the two arguments of the max operator. These are therefore points of phase transitions.

6 Other Variants of the Problem

As mentioned in Section 2, it makes sense to replace the FA constraint by a constraint that quantifies the true cost of the FA error, namely, superfluous data compression. This suggests to replace $g(\boldsymbol{x}) = P_0(\boldsymbol{x})$ by $g(\boldsymbol{x}) = P_0(\boldsymbol{x})e^{\theta L^*(\boldsymbol{x})}$, where $L^*(\boldsymbol{x})$ is still defined as above because when $\boldsymbol{x} \in \Omega$, we believe that the underlying source is P_1. This amounts to

$$g(\boldsymbol{x}) = P_0(\boldsymbol{x})[P_1(\boldsymbol{x})]^{-\theta/(1+\theta)}\left(\sum_{\boldsymbol{x}'\in\Omega}[P_1(\boldsymbol{x}')]^{1/(1+\theta)}\right)^{\theta}. \tag{37}$$

The problem is that now, similarly as in (15), Lemma 1 is not directly applicable since g depends on Ω and in a non–trivial manner.

There is, however, a way to circumvent this difficulty, that both improves performance and allows to use Lemma 1. Let us replace $L^*(\boldsymbol{x})$ by the length function of a *universal* encoder, which will be nearly optimal no matter whether P_0 or P_1 (or any other memoryless source) is the true underlying source. The best we can do is use a universal code whose length function, $L_U(\boldsymbol{x})$, is essentially as small as $n\hat{H}_{\boldsymbol{x}}(X)$ (up to a sub-linear additional term), where $\hat{H}_{\boldsymbol{x}}(X)$ stands for the empirical entropy of $\boldsymbol{x}$, namely, the entropy associated with the empirical distribution of $\boldsymbol{x}$.[2] Such a code is known to be asymptotically optimal, not only in the sense of the expected code–length, but also for a very wide class of additional criteria (see [13]), including $\boldsymbol{E}_1\exp\{\theta L(\boldsymbol{X})|\boldsymbol{X} \in \Omega\}$ and $P_1\{L(\boldsymbol{X}) \geq n|\boldsymbol{X} \in \Omega\}$.[3] We can now apply Lemma 1 with the choice

$$g(\boldsymbol{x}) = P_0(\boldsymbol{x})\exp\{n\theta\hat{H}_{\boldsymbol{x}}(X)\}. \tag{38}$$

By the same token, the choice of f can also be changed to

$$f(\boldsymbol{x}) = P_1(\boldsymbol{x})\exp\{n\theta\hat{H}_{\boldsymbol{x}}(X)\}. \tag{39}$$

More generally, one can use, of course, two different values of θ, say, θ_0 and θ_1 in eqs. (38) and (39), respectively, and finally, re–define $\Omega_\star$ accordingly to read

$$\Omega_\star = \left\{\boldsymbol{x}:\ P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} + e^{n\alpha}P_0(\boldsymbol{x})\exp\{n\theta_0\hat{H}_{\boldsymbol{x}}(X)\} \leq e^{n\beta}P_1(\boldsymbol{x})\right\}. \tag{40}$$

[2] For example, consider a two–part code that first describes the index of the type class and then the location of $\boldsymbol{x}$ within the type class.

[3] The fact that $L_U(\boldsymbol{x})$ asymptotically achieves the minimum of $\boldsymbol{E}_1\exp\{\theta L(\boldsymbol{x})|\boldsymbol{X} \in \Omega\}$, which is approximated by $[\sum_{\boldsymbol{x}\in\Omega}[P_1(\boldsymbol{x})]^{1/(1+\theta)}]^{1+\theta}$, can easily be verified using the method of types. Concerning the criterion $P_1\{L(\boldsymbol{X}) \geq nR|\boldsymbol{X} \in \Omega\}$, it achieves an error exponent of $\min\{\mathcal{D}(Q\|P_1):\ H(Q) \geq R,\ \mathcal{T}(Q) \subseteq \Omega\}$, which is the best possible, as can easily be shown by a straightforward modification of the converse part of [4, Theorem 1].

Similarly, we can now address directly the excess code–length criterion by choosing

$$g(\boldsymbol{x}) = P_0(\boldsymbol{x}) \cdot \mathcal{I}\{\boldsymbol{x}: \ \hat{H}_{\boldsymbol{x}}(X) \geq R\}, \tag{41}$$

$$f(\boldsymbol{x}) = P_1(\boldsymbol{x}) \cdot \mathcal{I}\{\boldsymbol{x}: \ \hat{H}_{\boldsymbol{x}}(X) \geq R\}, \tag{42}$$

and again, define $\Omega_\star$ accordingly. It should be emphasized that this passage from $L^*(\boldsymbol{x})$ to $n\hat{H}_{\boldsymbol{x}}(X)$ is not accompanied by loss in performance in terms of asymptotic exponents.

In the case of lossy source coding, $\hat{H}_{\boldsymbol{x}}(X)$, throughout this discussion, should be replaced by the empirical rate–distortion function, namely, the rate-distortion function associated with the empirical distribution induced by $\boldsymbol{x}$, or the empirical distortion–rate function, depending on the assumed regime, fixed–distortion and minimum rate or vice versa (see [1]). In all these variants, the asymptotic exponential performance can easily be assessed using the method types, similarly as before.

7 Universal Decision Rules

In the previous section, we discussed the use of universal lossless source coding, which facilitates the use of Lemma 1, and at the same time, makes sense even if P_0 and P_1 are known, because when $\boldsymbol{x} \in \Omega$, there is never full certainty that it has really emerged from P_1. But what happens if P_0 and P_1 are not both known (except for being memoryless)? The latest proposed version of $\Omega_\star$ (eq. (40)) still depends on P_0 and P_1, and hence not implementable in this case. We next turn to handle universality issues associated with the choice of the decision rule. The methodology here is similar to that of a few earlier papers on universal hypothesis testing (see, e.g., [5], [8], [14], [15], [16]). Lemma 1 is no longer used explicitly.

As a starting point, it will be more convenient to consider the problem

$$\begin{aligned} \text{minimize} \quad & P_1(\Omega^c) \\ \text{subject to} \quad & P_0(\Omega) \leq e^{-nE_{\text{FA}}} \\ & \sum_{\boldsymbol{x} \in \Omega} P_1(\boldsymbol{x}) \exp\{n\theta_1 \hat{H}_{\boldsymbol{x}}(X)\} \leq e^{\lambda_1 n}, \end{aligned} \tag{43}$$

which is equivalent to one of the versions of the earlier problem, except that the objective function and one of the constraints have interchanged their roles.

We begin with the case where P_0 is known but P_1 is not. Since P_1 is unknown, the second constraint must be imposed for *every* memoryless source P_1, that is,

$$\max_{P_1}\sum_{\boldsymbol{x}\in\Omega} P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\}\le e^{\lambda_1 n}. \tag{44}$$

First, observe that without loss of asymptotic optimality, every type class of source vectors, $\mathcal{T}_{\boldsymbol{x}}$, can be assumed to belong in its entirety to either Ω or Ω^c.[4] Accordingly, let $\mathcal{T}_{\boldsymbol{x}}\subseteq\Omega$. Then,

$$\begin{aligned}
e^{\lambda_1 n} &\ge \max_{P_1}\sum_{\boldsymbol{x}\in\Omega} P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} && (45)\\
&\ge \max_{P_1}\sum_{\boldsymbol{x}'\in\mathcal{T}_{\boldsymbol{x}}} P_1(\boldsymbol{x}')\exp\{n\theta_1\hat{H}_{\boldsymbol{x}'}(X)\} && (46)\\
&= \max_{P_1}|\mathcal{T}_{\boldsymbol{x}}|\cdot P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} && (47)\\
&= \exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)-O(\log n)\}. && (48)
\end{aligned}$$

The conclusion is then that $\mathcal{T}_{\boldsymbol{x}}\subseteq\Omega$ implies $\mathcal{T}_{\boldsymbol{x}}\subseteq\{\boldsymbol{x}:\ \hat{H}_{\boldsymbol{x}}(X)\le\lambda_1/\theta_1+O(\log n/n\}$, which means

$$\Omega\subseteq\{\boldsymbol{x}:\ \hat{H}_{\boldsymbol{x}}(X)\le\lambda_1/\theta_1+O(\log n/n)\}. \tag{49}$$

From the first constraint of (43), we similarly have:

$$\Omega\subseteq\{\boldsymbol{x}:\ \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)\ge E_{\text{FA}}-O(\log n/n)\}, \tag{50}$$

where $\hat{P}_{\boldsymbol{x}}$ is the empirical distribution associated with $\boldsymbol{x}$. Combining the last two equations, we get:

$$\Omega\subseteq\Omega_u\stackrel{\Delta}{=}\{\boldsymbol{x}:\ \hat{H}_{\boldsymbol{x}}(X)\le\lambda_1/\theta_1+O(\log n/n),\ \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)\ge E_{\text{FA}}-O(\log n/n)\}. \tag{51}$$

We now propose Ω_u as our universal decision rule. First, observe that it asymptotically satisfies the constraints, as

$$P_0(\Omega_u)\le P_0\{\boldsymbol{x}:\ \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)\ge E_{\text{FA}}-O(\log n/n)\}\doteq\exp\{-n[E_{\text{FA}}-O(\log n/n)\}, \tag{52}$$

and

$$\sum_{\boldsymbol{x}\in\Omega_u} P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} \le \max_{P_1}\sum_{\boldsymbol{x}\in\Omega_u} P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\}$$

[4] If this is not the case, then at least half of the members of the type class belong to either Ω or Ω^c. By transferring the smaller part of each type class to the other decision region, to join the majority therein, one at most doubles the probability of that region, while reducing the probability of the other region. This has no negative impact on the asymptotic exponents.

$$\begin{aligned} &\leq \sum_{\boldsymbol{x}\in\Omega_u} \max_{P_1} P_1(\boldsymbol{x})\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} \\ &\leq \sum_{\boldsymbol{x}\in\Omega_u} \exp\{-n\hat{H}_{\boldsymbol{x}}(X)\}\cdot\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} \\ &= \sum_{\mathcal{T}_{\boldsymbol{x}}\subset\Omega_u} |\mathcal{T}_{\boldsymbol{x}}|\cdot\exp\{-n\hat{H}_{\boldsymbol{x}}(X)\}\cdot\exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} \\ &\doteq \max_{\mathcal{T}_{\boldsymbol{x}}\subset\Omega_u} \exp\{n\theta_1\hat{H}_{\boldsymbol{x}}(X)\} \\ &\doteq e^{\lambda_1 n}. \end{aligned} \tag{53}$$

On the other hand, since Ω_u is a super-set of any competing decision rule Ω that satisfies the constraints (see eq. (51)), then it follows that $\Omega_u^c \subseteq \Omega^c$, and so, $P_1(\Omega_u^c) \leq P_1(\Omega^c)$, for every P_1. This means that Ω_u minimizes the MD probability *uniformly* for every memoryless source P_1 and hence establishes the optimality of Ω_u.

The idea here is that Ω_u is essentially the largest subset of $\mathcal{X}^n$ that still satisfies the constraints, and hence its complement is the smallest possible. Once again, we see that membership in Ω_u consists of two requirements: the requirement on the empirical entropy, which limits the code length, and a requirement on the divergence, which means that $\boldsymbol{x}$ is far enough from being typical to P_0, in order to reject unwanted data that stems from P_0.

Universal counterparts of other variants of the problem, discussed in the previous section, can be derived in a similar manner. For example, if the constraint $P_0(\Omega) \leq e^{-nE_{\text{FA}}}$ is replaced by compression cost constraint

$$\sum_{\boldsymbol{x}\in\Omega} P_0(\boldsymbol{x})\exp\{n\theta_0\hat{H}_{\boldsymbol{x}}(X)\} \leq e^{\lambda_0 n} \tag{54}$$

then the set $\{\boldsymbol{x}:\ \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0) \geq E_{\text{FA}} - O(\log n/n)\}$, in eq. (50), should be replaced by $\{\boldsymbol{x}:\ \theta_0\hat{H}_{\boldsymbol{x}}(X) - \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0) \leq \lambda_0\}$ and Ω_u should, of course, be modified accordingly. If, in addition, P_0 is unknown as well, and this constraint is imposed for every memoryless source P_0 on $\mathcal{X}$, then this becomes $\{\boldsymbol{x}:\ \theta_0\hat{H}_{\boldsymbol{x}}(X) \leq \lambda_0\}$. Thus, overall Ω_u would be redefined as

$$\Omega_u = \{\boldsymbol{x}:\ \hat{H}_{\boldsymbol{x}}(X) \leq \min\{\lambda_0/\theta_0, \lambda_1/\theta_1\}\}. \tag{55}$$

Suppose next that both P_0 and P_1 are unknown but there are training sequences available from each one of these sources. In other words, in addition to the vector $\boldsymbol{x} \in \mathcal{X}^n$ as before, we also have a training sequence, $\boldsymbol{x}_0 \in \mathcal{X}^m$ from P_0, and a training sequence, $\boldsymbol{x}_1 \in \mathcal{X}^m$ from P_1. A natural approach would be the plug–in approach: First estimate each source from its corresponding training data and then use each estimate in place of the corresponding unknown, true source. This is a sub-optimal approach because it is based on separation and it does not use $\boldsymbol{x}$ for estimating either source. The best approach is to combine the training and the decision into a single step, which means that our decision rule classifies triples $\{(\boldsymbol{x}, \boldsymbol{x}_0, \boldsymbol{x}_1)\}$ rather than single vectors $\{\boldsymbol{x}\}$ as before. The compression cost constraints will now read

$$\sum_{\boldsymbol{x},\boldsymbol{x}_0,\boldsymbol{x}_1 \in \Omega} P_0(\boldsymbol{x}_0) P_1(\boldsymbol{x}_1) P_0(\boldsymbol{x}) \exp\{n\theta_0 \hat{H}_{\boldsymbol{x}}(X)\} \le \exp(\lambda_0 n) \tag{56}$$

and

$$\sum_{\boldsymbol{x},\boldsymbol{x}_0,\boldsymbol{x}_1 \in \Omega} P_0(\boldsymbol{x}_0) P_1(\boldsymbol{x}_1) P_1(\boldsymbol{x}) \exp\{n\theta_1 \hat{H}_{\boldsymbol{x}}(X)\} \le \exp(\lambda_1 n), \tag{57}$$

both imposed for every two memoryless sources P_0 and P_1. Here, we assume, again without loss of asymptotic optimality, that Ω is a union of Cartesian products of type classes $\mathcal{T}_{\boldsymbol{x}} \times \mathcal{T}_{\boldsymbol{x}_0} \times \mathcal{T}_{\boldsymbol{x}_1}$. As for the first constraint, we have

$$\begin{aligned} e^{\lambda_0 n} &\ge \max_{P_0,P_1} \sum_{\boldsymbol{x},\boldsymbol{x}_0,\boldsymbol{x}_1 \in \Omega} P_0(\boldsymbol{x}_0) P_1(\boldsymbol{x}_1) P_0(\boldsymbol{x}) \exp\{n\theta_0 \hat{H}_{\boldsymbol{x}}(X)\} && (58) \\ &\ge \exp\{n\theta_0 \hat{H}_{\boldsymbol{x}}(X) - n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_0}) - m\mathcal{D}(\hat{P}_{\boldsymbol{x}_0} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_0}) - O(\log n)\}, && (59) \end{aligned}$$

where $\hat{P}_{\boldsymbol{x}\boldsymbol{x}_i}$ denotes the empirical distribution associated with the concatenation of $\boldsymbol{x}$ and $\boldsymbol{x}_i$, $i = 0, 1$. Similarly, for the other constraint

$$\begin{aligned} e^{\lambda_1 n} &\ge \max_{P_0,P_1} \sum_{\boldsymbol{x},\boldsymbol{x}_0,\boldsymbol{x}_1 \in \Omega} P_0(\boldsymbol{x}_0) P_1(\boldsymbol{x}_1) P_1(\boldsymbol{x}) \exp\{n\theta_1 \hat{H}_{\boldsymbol{x}}(X)\} && (60) \\ &\ge \exp\{n\theta_1 \hat{H}_{\boldsymbol{x}}(X) - n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_1}) - m\mathcal{D}(\hat{P}_{\boldsymbol{x}_1} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_1}) - O(\log n)\} && (61) \end{aligned}$$

and then Ω_u is defined as

$$\begin{aligned} \Omega_u = \{\boldsymbol{x} :\ & \theta_0 \hat{H}_{\boldsymbol{x}}(X) - \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_0}) - \frac{m}{n}\mathcal{D}(\hat{P}_{\boldsymbol{x}_0} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_0}) - O(\log n/n) \le \lambda_0, \\ & \theta_1 \hat{H}_{\boldsymbol{x}}(X) - \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_1}) - \frac{m}{n}\mathcal{D}(\hat{P}_{\boldsymbol{x}_1} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_1}) - O(\log n/n) \le \lambda_1\}. \end{aligned} \tag{62}$$

The terms $\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_i}) + \frac{m}{n}\mathcal{D}(\hat{P}_{\boldsymbol{x}_1} \| \hat{P}_{\boldsymbol{x}\boldsymbol{x}_i})$ measure the 'distance' between $\hat{P}_{\boldsymbol{x}}$ and $\hat{P}_{\boldsymbol{x}_i}$, $i = 0, 1$. If they are close, these terms are small and we compare the code length to a threshold. If they are far apart, we can afford to be more tolerant concerning the length since this is a rare event anyway.

The empirical distributions $\hat{P}_{\boldsymbol{x}\boldsymbol{x}_0}$ and $\hat{P}_{\boldsymbol{x}\boldsymbol{x}_1}$ stand for the fact that, in some sense, $\boldsymbol{x}$ participates in the estimation of the two sources, unlike the 'plug-in' approach describe above.

8 Conclusion

We have addressed the problem of joint detection and lossless data compression in several variants, including the universal regime, where at least one of the sources is unknown, with and without training sequences from each source. The method of our derivations can also be carried out in several more general situations.

First, it is not difficult to extend our results from memoryless sources to Markov sources, or even more generally, to unifilar finite–state sources. This is possible because the method of types extends to these classes of sources as well. Moreover, in the universal setting, it is more interesting to consider the case where the Markov order is unknown (or in the case of unifilar finite–state sources, the state–transition diagram and the number of states are unknown). In this case, it is expected that the length function of the Lempel–Ziv algorithm can be invoked instead of the empirical entropy, similarly as was done in earlier work (see, e.g., [5], [8]).

Secondly, as mentioned already in the Abstract and the Introduction, one may consider tasks other than lossless data compression. One of them is lossy data compression, and we have already mentioned, at the end of Section 6, how to modify our decision rule to account for this case. Channel decoding is another important task that has already been addressed in [7]. Additional tasks may be quantization, estimation, encryption, and so on. The general guideline is always to try to present (or approximate) the objective function (pertaining to the optimal strategy of the task within Ω) as (a monotonic function of) the summation or integral of some function $f(\boldsymbol{x})$ over Ω, and then use this f in the decision rule $\Omega_\star$ of eq. (3). The function f should be independent of Ω.

References

[1] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Second Edition, Hoboken, New Jersey, U.S.A., 2006.

[3] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[4] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–20, no. 2, pp. 197–199, March 1974.

[5] N. Merhav, "Universal detection of messages via finite–state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242–2246, September 2000.

[6] N. Merhav, "On optimum strategies for minimizing exponential moments of a loss function," *Communications in Information and Systems*, vol. 11, no. 4, pp. 343–368, 2011.

[7] N. Merhav, "Codeword or noise? Exact random coding exponents for slotted asynchronism," submitted to *IEEE Trans. Inform. Theory*, August 2013. http://arxiv.org/pdf/1308.4572.pdf

[8] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1014–1019, September 1989.

[9] G. V. Moustakides, "Optimum joint detection and estimation," *Proc. 2011 IEEE Symposium on Information Theory (ISIT 2011)*, pp. 2984–2988, St. Petersburg, Russia, July 2011.

[10] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: optimum tests and applications," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4215–4229, July 2012.

[11] D. Wang, *Distinguishing Codes From Noise: Fundamental Limits and Applications to Sparse Communication*, Master thesis, Massachusetts Institute of Technology, Department of EECS, June 2010.

[12] D. Wang, V. Chandar, S.-Y. Chung, and G. Wornell, “Error exponents in asynchronous communication,” *Proc. 2011 IEEE International Symposium on Information Theory*, pp. 1071–1075, 2011.

[13] M. J. Weinberger, N. Merhav, and M. Feder, “Optimal sequential probability assignment for individual sequences,” *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 384-396, March 1994.

[14] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, September 1992.

[15] J. Ziv, “On classification with empirically-observed statistics and universal data compression,” *IEEE Trans. Inform. Theory*, vol. IT–34, no. 2, pp. 278–286, March 1988.

[16] J. Ziv, “Compression, tests for randomness, and estimating the statistical model of an individual sequence,” *Proc. Sequences*, R. M. Capocelli Ed., New York: Springer Verlag, pp. 366–373, 1990.