Joint Estimation and Detection Against Independence

Gil Katz, Pablo Piantanida, Romain Couillet and Mérouane Debbah
SUPELEC
Gif sur Yvette, France
Email: gil.katz@supelec.fr, pablo.piantanida@supelec.fr, romain.couillet@supelec.fr, merouane.debbah@supelec.fr

***Abstract*—A receiver in a two-node system is required to make a decision of relevance as to received information, using side information that may or may not be correlated with the received signal. In case the information is judged to be relevant, the receiver is then required to estimate the source with average distortion D. Focusing on the case of testing against independence, a single-letter expression for the rate-error-distortion region is proposed and proven. The resulting region ports a surprising resemblance to a seemingly non-associated classification problem, known as the information-bottleneck. The optimal region is then calculated for a binary symmetric example. Results demonstrate an interesting trade-off between the achievable error-exponent for the decision and the distortion at the decoder.**

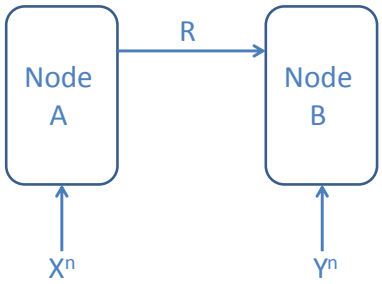

Fig. 1: Simple Detection and Estimation Model

I. Introduction

The problem of hypothesis testing (HT) is very familiar in statistics. Using a list of i.i.d. realizations, a statistician is required to determine the probability distribution (or "law") of the random variable (RV) X. In the binary HT problem, it is assumed that the probability distribution is one of two possible laws (commonly called hypothesis H_0 and hypothesis H_1), both of which are known to the statistician. Two error events are commonly defined in HT problems. The error of the first type, whose probability is dependent on the number of available realizations and is denoted by α_n, is defined to be the event in which H_1 is chosen despite H_0 being true. Likewise, the error of the second type, with probability β_n, is defined to be the event in which H_0 is chosen despite H_1 being true.

Obviously, there is a trade-off between the probabilities of the two types of error events. *Stein's Lemma* (see e.g. [1]) determines the optimal exponential rate in which the probability of the second type decays to zero, under any fixed and positive constraint over the probability of error of the first type ($\alpha_n \leq \epsilon$, $\epsilon > 0$). It turns out, that the optimal exponential rate of decay of β_n *does not depend* on the specific constraint over α_n, and is equal to the Kullback-Leiber divergence between the two possible probability distributions:

$$\theta \triangleq -\frac{1}{n} \lim_{n\to\infty} \beta_n^* = D(P_0||P_1) \ , \qquad (1)$$

where P_0 and P_1 are the probability distributions implied by hypotheses H_0 and H_1, respectively.

The distributed HT problem [2]–[5] assumes the existence of several RVs, that are commonly distributed according to one of two (or more) hypotheses. It is assumed that each of the RVs is received at a different location ("node") in the system. Nodes are allowed to communicate (under restrictions depending on the specific layout of the system), and are required to reach a common decision as to the law governing the RVs they see. While [5] defines the L-encoder general HT problem and proposes both lower and upper bounds for the performance of an $L+1$ node system, most work on the subject is confined to the two-node HT problem, in which two RVs, X and Y, are seen by node A and node B, respectively. In this paper, we choose to confine ourselves to the popular assumption (see e.g. [2], [3]) that the decision is done by one of the two original nodes (namely node B, without loss of generality). This assumption implies that the realizations of Y can always be used by the statistician. We assume that information about the realizations of X is transferred from node A to node B through an error-free link with rate $R < H(X)$. See Fig. 1 for a visual depiction of the system model assumed throughout this paper.

The two-node HT against independence problem is defined to be the problem in which the alternative hypothesis H_1 is the multiplication of the marginal distributions of X and Y, according to hypothesis H_0. In [2], it has been shown that the optimal approach for this case is to first choose a function of the source of X that complies to the rate constraint. Then, Stein's Lemma can be applied over the joint distribution of $f(\mathbf{X}^n)$ and $\mathbf{Y}^n$. Optimality is achieved by optimizing over all functions such that the rate constraint is respected.

In this paper, we investigate a scenario in which, after testing against independence, node B also wishes to estimate the realizations of X, with average distortion D. It is assumed that estimation is only relevant in case node B *correctly* decides H_0. If the decision is H_1, the realizations of X are considered irrelevant, and estimation is not attempted. In case the decision H_0 is incorrect, however, estimation is attempted, but defining the distortion in this case may be hard in many practical systems, and it is assumed that the failure of the system is already encapsulated by the probability of error in the HT stage.

By building on the results of [2], [3], we propose and prove a single-letter expression for the rate-error-distortion region of the system. It is shown that the optimal region is attained by first performing HT as in [2], and then using known results for source coding with side information at the decoder [6], while ignoring the information received by node B at the HT stage.

Using the single-letter region for the rate-error-distortion problem, an interesting relation to a known classification problem is uncovered [7]–[9]. This problem, called the information bottleneck problem, is focused on the possibility of conveying to a receiver some *relevant information* about a source at the transmitter's end. The relationship between the two problems, while surprising at first, is explained, offering intriguing prospects for future research.

The remainder of this paper is organized as follows. In Section II the system model is presented. Section III presents the proposed single-letter rate-error-distortion region, with Section IV and Section V containing the achievability and converse parts of the proof, respectively. In Section VI the information bottleneck problem is presented, and its relationship with the binary HT problem is brought forth and explained. Finally, a specific example of HT against independence and estimation for a binary symmetric source is presented in Section VII, before some concluding remarks are given in Section VIII.

II. System Model and Preliminaries

A. Notation

We use upper-case letters to denote RVs and lower-case letters to denote realizations of RVs. Vectors are denoted by bold-face letters. The length of the vector appears as a super-script, and may be omitted when it is clear from the context. $\mathbf{X}_a^b$ denotes the random vector X from place a to place b. We use $H(\cdot)$ to refer to the general (discrete) entropy function and $H_2(\cdot)$ to refer to the binary entropy function. $I(\cdot;\cdot)$ is the mutual information function and $D(\cdot||\cdot)$ refers to the Kullback-Leiber divergence.

B. Distributed Detection & Estimation Model

We focus here on a simple detection and estimation model, comprising two nodes, as shown in Fig. 1. Nodes A and B each see n i.i.d. realizations of the RVs X and Y, respectively. We assume testing against independence throughout this paper. Thus,

$$\begin{aligned} H_0: &\quad P_0(x,y) = P_{XY}(x,y)\ , \\ H_1: &\quad P_1(x,y) = P_X(x)P_y(y)\ . \end{aligned} \tag{2}$$

We assume that node A can send information to node B over an error-free link with rate R bits per source-symbol. Having received the information from node A, node B is then required to make a decision between the two possible hypotheses. Throughout this paper we use the widespread definition of the two types of error probabilities, defined by

$$\begin{aligned} \alpha_n &\triangleq \Pr\left(H_1|XY \sim P_0(x,y)\right)\ , \\ \beta_n &\triangleq \Pr\left(H_0|XY \sim P_1(x,y)\right)\ . \end{aligned} \tag{3}$$

Only in case node B detects the RVs are distributed according to H_0, it then attempts to estimate the original realizations of X, with average distortion D.

In [2] (see also [3]), the authors show that when testing against independence, the optimal approach at node B is to apply Stein's Lemma over the common distribution of Y^n and the received version of X^n, $f(X^n)$. By optimizing over the compressing function f, the resulting asymptotic behavior of the smallest possible probability of error of type 2 (for a fixed constraint over the probability of error of type 1, $\alpha_n \leq \epsilon$) is

$$\theta(R) = \sup_{k\leq n} \theta_k(R)\ , \tag{4}$$

where

$$\begin{aligned} \theta_k(R) &= \sup_f \left\{ \frac{1}{k} D\left(P_{f(\mathbf{X}^k)\mathbf{Y}^k} || P_{f(\mathbf{X}^k)} P_{\mathbf{Y}^k}\right) | \log||f|| \leq kR \right\} \\ &= \sup_f \left\{ \frac{1}{k} I\left(f(\mathbf{X}^k); \mathbf{Y}^k\right) | \log||f|| \leq kR \right\}\ . \end{aligned} \tag{5}$$

This result implies that, much like in the single-node HT case, the optimal exponential decay of β_n is not dependent upon the chosen constraint over the error probability of the first type, $\alpha_n \leq \epsilon$ ($\epsilon > 0$).

Using this result, the rate-error-distortion region of the system depicted in Fig. 1 can be described by

$$\begin{cases} \frac{1}{n}\log||f|| \leq R \\ \theta(R) \geq E \\ \mathbb{E}_{H_0}\left[d(g(f(\mathbf{X}^n), \mathbf{Y}^n), \mathbf{X}^n)|H_0\right] \leq D \end{cases}\ , \tag{6}$$

with $d(\cdot,\cdot)$ being a distortion measure, assumed to be additive $d(a^n, b^n) = \sum_{i=1}^{n} d(a_i, b_i)$, and $g : \hat{\mathcal{X}}^n \times \mathcal{Y}^n \to \mathcal{Z}^n$ is the decoding function, from the encoded version of x^n and y^n to some arbitrary alphabet $\mathcal{Z}$. Note that we only measure the distortion when Node B *correctly* decides H_0. Finally, we note that from [2, Lemma 1.a], when n is large enough, $\theta(R) = \theta_n(R)$, as defined by setting $k = n$ in (4). We thus use the expression

$$\theta(R) = \sup_f \left\{ \frac{1}{n} I\left(f(\mathbf{X}^n); \mathbf{Y}^n\right) | \log||f|| \leq nR \right\} \tag{7}$$

for the remainder of this paper.

III. Single-Letter Rate-Error-Distortion-Region

In this section we give our main result, being a single-letter expression for the rate-error-distortion region in (6).

Proposition 1. *The point (R, E, D) is achievable for the two-node detection and estimation problem as defined in* (6)*, if and only if two RVs can be found, such that*

$$\begin{cases} I(U;X) + I(V;X|UY) \leq R \\ I(U;Y) \geq E \\ \mathbb{E}_{H_0}\left[d(g(UVY), X)|H_0\right] \leq D \end{cases}\ , \tag{8}$$

with U and V being those RVs such that $U - V - X - Y$ form a Markov chain.

The proof is presented in the following sections. Section IV presents the proof of achievability, which consists of dividing the available rate into two parts. The first part is used for detection, while the second part allows node B

to estimate X. The converse part of the proof is given in Section V.

Two final remarks are in order before the presentation of the proof of Proposition 1. First, note that the expression for the rate can be evaluated as follows:

$$\begin{aligned} R &\geq I(U;X) + I(V;X|UY) \\ &= I(U;X) + I(V;XY|U) - I(V;Y|U) \\ &= I(U;X) + I(V;X|U) - I(V;Y|U) \\ &= I(U;X) + I(UV;X) - I(U;X) \\ &\qquad - I(UV;Y) + I(U;Y) \\ &= I(U;Y) + [I(V;X) - I(V;Y)] \ , \end{aligned} \tag{9}$$

where the final equality stems from the Markov chain formed by the RVs. Note that the rate can now be seen as comprised of two distinguished parts. The first part of the resulting expression in (9) is consecrated to detection, and is in fact identical to the expression of the error exponent given in (8) (which is in agreement with previous results [2], [3]). The second part of the rate is consecrated to decoding. This part is *independent of* U, which implies, rather surprisingly, that the information used for detection is useless for the sake of estimation, after the decision has been made. We expect this result to change for the general case (Where HT is not necessarily against independence).

Finally, defining the message sent from node A to node B as $W \triangleq f(\mathbf{X}^n)$, the constraint over the error exponent in (6) can be rewritten as

$$\sup_f \frac{1}{n} I(W;\mathbf{Y}^n) \geq E \ . \tag{10}$$

Using the Markovian relation between the message W and both random sequences, $W - \mathbf{X}^n - \mathbf{Y}^n$, and the fact that $I(\mathbf{X}^n;\mathbf{Y}^n)$ is given and fixed, the same constraint can be written as follows:

$$\begin{aligned} &\frac{1}{n}\left[I(\mathbf{X}^n;\mathbf{Y}^n) - I(W;\mathbf{Y}^n)\right] \leq \mu \ , \\ &\frac{1}{n}\mathbb{E}\underbrace{\left[\log\frac{P_{\mathbf{X}^n\mathbf{Y}^n}}{P_{\mathbf{X}^n}P_{\mathbf{Y}^n}} - \log\frac{P_{W\mathbf{Y}^n}}{P_W P_{\mathbf{Y}^n}}\right]}_{\hat{d}((\mathbf{X}^n\mathbf{Y}^n),(W\mathbf{Y}^n))} \leq \mu \ . \end{aligned} \tag{11}$$

Thus, the joint problem of detection against independence and estimation can be viewed as a re-distortion problem, in which the quality of the estimated message is measured by two different distortion functions - the first is an additive function d as defined in (6), while the second is a general non-additive function $\hat{d}$, as defined above.

IV. Proof of Achievability

Codebook: Divide the available rate into two parts. First choose a RV U such that $U - X - Y$ form a Markov chain. Randomly pick 2^{nS_1} sequences $\mathbf{u}^n(s_1)$ from the typical set $T_\delta^n(U)$, with typicality defined as in [3]. Define $R' \triangleq R - \hat{R}$. Choose a RV V such that $U - V - X - Y$. For each codeword in U's codebook $\mathbf{u}(s_1)$, randomly pick 2^{nS_2} sequences $\mathbf{v}^n(s_1,s_2)$ from the conditional typical set $T_\delta^n(V|\mathbf{u}(s_1))$ and divide them into $2^{nR'}$ bins, such that each bin contains roughly $2^{n(S_2-R')}$ sequences.

Encoding: Assuming that the sequence $\mathbf{x}^n$ was produced by the source of X, look for the first codeword in U's codebook such that $(\mathbf{u}^n(s_1),\mathbf{x}^n) \in T_\delta^n(UX)$. Then, look for the first codeword $\mathbf{v}^n(s_1,s_2)$ s.t. $(\mathbf{v}^n(s_1,s_2),\mathbf{x}^n) \in T_\delta^n(VX|\mathbf{u}(s_1))$. Let b be the bin of $\mathbf{v}^n(s_1,s_2)$. Send the message $f(\mathbf{x}^n) = (s_1,b)$ to node B.

Decoding: Given $\mathbf{u}(s_1), b$ and $\mathbf{y}^n$, the decoder first checks if $(\mathbf{u}^n(s_1),\mathbf{y}^n) \in T_\delta^n(UY)$. If so, it declares H_0. Else it declares H_1. If the decoder decides H_0 it then attempts to decode the message (with average distortion D) by using $\mathbf{v}(s_1,s_2)$. This codeword is first recovered by looking in bin b for the unique codeword such that $\mathbf{v}^n(s_1,s_2) \in T_\delta^n(V|\mathbf{u}(s_1),\mathbf{y}^n)$. Then, a function $g(\cdot)$ can be used over the entire available information (U, V and Y) in order to decode.

Errors and Constraints: We start with the HT part, and the relation between the expression $I(U;X)$ and the achievable error exponent. Denoting by ϵ_1 the event "an error occurred during encoding" (of the HT part U), we expend its probability as $\Pr(\epsilon_1) \leq P_0 + P_1$ with:

$$\begin{aligned} P_0 &\triangleq \Pr\{\mathbf{X}^n \notin T_\delta^n(X)\} \ , \\ P_1 &\triangleq \{\nexists s_1 \text{ s.t. } (\mathbf{u}(s_1),\mathbf{X}^n) \in T_\delta^n(UX) | \mathbf{X}^n \in T_\delta^n(X)\} \ , \end{aligned} \tag{12}$$

being the probabilities that the source X produces a non-typical sequence, and that (for a typical source sequence) the codebook doesn't contain an appropriate codeword, respectively. From the asymptotic equipartition property (AEP), $P_0 \leq \eta_n^{(1)} \underset{n\to\infty}{\longrightarrow} 0$. As for P_1:

$$\begin{aligned} P_1 &= \left(\Pr\{(\mathbf{U}^n,\mathbf{X}^n) \notin T_\delta^n(UX) \right. \\ &\qquad \left. |\mathbf{U}^n \in T_\delta^n(U), \mathbf{X}^n \in T_\delta^n(X)\}\right)^{2^{nS_1}} \\ &= \left(1 - \Pr\{(\mathbf{U}^n,\mathbf{X}^n) \in T_\delta^n(UX) \right. \\ &\qquad \left. |\mathbf{U}^n \in T_\delta^n(U), \mathbf{X}^n \in T_\delta^n(X)\}\right)^{2^{nS_1}} \\ &\overset{(a)}{\leq} 2^{-2^{nS_1}\Pr\{(\mathbf{U}^n,\mathbf{X}^n)\in T_\delta^n(UX)|\mathbf{U}^n\in T_\delta^n(U),\mathbf{X}^n\in T_\delta^n(X)\}} \\ &\leq 2^{-2^{nS_1}2^{-n\left(I(U;X)+\eta_n^{(2)}\right)}} \\ &= 2^{-2^{-n\left(I(U;X)-S_1+\eta_n^{(2)}\right)}} \ . \end{aligned} \tag{13}$$

Here, inequality (a) is due to the inequality $(1-a)^n \leq 2^{an}$ [10]. Since $\eta_n^{(2)} \underset{n\to\infty}{\longrightarrow} 0$, $P_1 \to 0$ if $S_1 > I(U;X)$. In this part of the coding scheme we send the index s_1 without binning. Thus, this result implies that $\hat{R} > I(U;X)$ is necessary for the achievability of this coding scheme.

Next, we look at the achievable error exponent with the proposed encoding scheme:

$$\begin{aligned} \frac{1}{n} I\left(f(\mathbf{X}^n);\mathbf{Y}^n\right) &= \frac{1}{n}\left[H(\mathbf{Y}^n) - H(\mathbf{Y}^n|f(\mathbf{X}^n))\right] \\ &= H(Y) - \frac{1}{n}H(\mathbf{Y}^n|f(\mathbf{X}^n)) \ . \end{aligned} \tag{14}$$

The second term here can be evaluated by defining the RV

$$\hat{\mathbf{Y}}^n = \begin{cases} \mathbf{Y}^n & \text{if } (\mathbf{u}^n(s_1),\mathbf{Y}^n) \in T_\delta^n(UY) \\ \emptyset & \text{else} \end{cases} \ , \tag{15}$$

and writing

$$\begin{aligned}\frac{1}{n}H(\mathbf{Y}^n|f(\mathbf{X}^n)) &\overset{(b)}{\leq} \frac{1}{n}H(\mathbf{Y}^n|S_1)\\ &= \frac{1}{n}\sum_{j=1}^{2^{nS_1}} H(\mathbf{Y}^n|S_1=j)\Pr(S_1=j)\\ &= \frac{1}{n}\sum_{j=1}^{2^{nS_1}} H(\mathbf{Y}^n\hat{\mathbf{Y}}^n|S_1=j)\Pr(S_1=j)\\ &= \frac{1}{n}\sum_{j=1}^{2^{nS_1}} \left(\underbrace{H(\hat{\mathbf{Y}}^n|S_1=j)}_{(*)}+\underbrace{H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n,S_1=j)}_{(**)}\right)\cdot \\ &\qquad \Pr(S_1=j)\ . \end{aligned} \tag{16}$$

Here, inequality (b) stems from the fact that $f(\mathbf{X}^n)$ contains (but is not limited to) the information S_1, and side information makes entropy smaller. We bound this expression further by treating each part separately:

$$\begin{aligned}(*) &= \frac{1}{n}\sum_{j=1}^{2^{nS_1}} H(\hat{\mathbf{Y}}^n|S_1=j)\Pr(S_1=j)\\ &\overset{(c)}{\leq} \frac{1}{n}\sum_{j=1}^{2^{nS_1}} \log\left(||T_\delta^n(Y|\mathbf{u}^n(j))||+1\right)\Pr(S_1=j)\\ &\overset{(d)}{\leq} \sum_{j=1}^{2^{nS_1}} \left(H(Y|U)+\eta_n^{(3)}\right)\Pr(S_1=j)\\ &= H(Y|U)+\eta_n^{(3)}\ , \end{aligned} \tag{17}$$

where (c) is due to the fact that uniform distribution maximizes entropy and (d) stems from [11, Lemma 2].

$$\begin{aligned}(**) &= \frac{1}{n}\sum_{j=1}^{2^{nS_1}} H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n,S_1=j)\Pr(S_1=j)\\ &\overset{(e)}{\leq} \frac{1}{n}\sum_{j=1}^{2^{nS_1}} \left(1+\Pr\{\mathbf{Y}^n\neq\hat{\mathbf{Y}}^n|S_1=j\}\log||\mathcal{Y}||^n\right)\\ &\qquad\qquad \cdot\Pr(S_1=j)\\ &\leq \frac{1}{n}+\sum_{j=1}^{2^{nS_1}} \Pr\{(\mathbf{u}^n(S_1),\mathbf{Y}^n)\notin T_\delta^n(UY)|S_1=j\}\\ &\qquad\qquad \cdot\log||\mathcal{Y}||\Pr(S_1=j)\\ &\leq \frac{1}{n}+(P_0+P_1)\log||\mathcal{Y}||\ . \end{aligned} \tag{18}$$

Here, (e) stems from Fano's inequality. As was already shown, if $S_1 > I(U;X)$ both P_0 and P_1 go to 0 when $n\to\infty$. Thus

$$H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n,S_1=j)\Pr(S_1=j)\leq\eta_n^{(4)}\underset{n\to\infty}{\longrightarrow}0\ . \tag{19}$$

All in all:

$$\frac{1}{n}H(\mathbf{Y}^n|S_1)\leq H(Y|U)+\eta_n^{(3)}+\eta_n^{(4)}\ , \tag{20}$$

and

$$\begin{aligned}\frac{1}{n}I(f(\mathbf{X}^n);\mathbf{Y}^n) &\geq H(Y)-H(Y|U)-\eta_n^{(3)}-\eta_n^{(4)}\\ &= I(U;Y)-\eta_n^{(3)}-\eta_n^{(4)}\ . \end{aligned} \tag{21}$$

Thus if $I(U;Y)\geq E$ so is $\frac{1}{n}I(f(X^n);Y^n)$ and the achievability of the error exponent is complete.

Finally, we show that given a (correct) decision H_0, the RV V can be used to decode X^n with the desired distortion: Denoting by ϵ_2 the event "an error occurred during encoding or decoding" (of V), we expend its probability as follows $\Pr(\epsilon_2)\leq P_2+P_3$, with P_2 being the probability that no codeword $\mathbf{v}(s_1,s_2)$ could be found in the codebook for the given sequence $\mathbf{x}^n$ and the chosen codeword $\mathbf{u}(s_1)$, and P_3 being the probability that a different codeword in the same bin b is compatible with $\mathbf{y}^n$ and $\mathbf{u}(s_1)$.

$$\begin{aligned}P_2 &\triangleq \Pr\{\nexists s_2 \text{ s.t. } (\mathbf{v}^n(s_1,s_2),\mathbf{x}^n)\in T_\delta^n(VX|\mathbf{u}^n(s_1))\}\\ &= \big(\Pr\{(\mathbf{V}^n,\mathbf{X}^n)\notin T_\delta^n(VX|\mathbf{u}(s_1))\\ &\qquad |\mathbf{V}^n\in T_\delta^n(V|\mathbf{u}(s_1)),\mathbf{X}^n\in T_\delta^n(X|\mathbf{u}(s_1))\}\big)^{2^{nS_2}}\\ &\leq 2^{-2^{nS_2}2^{-n\left(I(V;X|U)+\eta_n^{(5)}\right)}}\\ &= 2^{-2^{-n\left(I(V;X|U)-S_2+\eta_n^{(5)}\right)}}\ . \end{aligned} \tag{22}$$

Thus, $P_2\underset{n\to\infty}{\longrightarrow}0$ if $S_2>I(V;X|U)$. Finally,

$$P_3\triangleq\Pr\{\exists s_2'\in b \text{ s.t. } \mathbf{v}^n(s_1,s_2')\in T_\delta^n(V|\mathbf{u}^n(s_1),\mathbf{y}^n)\}\ , \tag{23}$$

with b being the bin sent to node B.

$$\begin{aligned}P_3 &\leq 2^{n\left(S_2-R'+\epsilon\right)}\Pr\{\mathbf{V}^n\in T_\delta^n(V|\mathbf{u}^n(s_1),\mathbf{y}^n)\\ &\qquad\qquad |\mathbf{V}^n\in T_\delta^n(V|\mathbf{u}^n(s_1))\}\\ &\leq 2^{n\left(S_2-R'+\epsilon\right)}2^{-n(I(V;Y|U)+\eta_n(6))}\\ &= 2^{-n\left(I(V;Y|U)-(S_2-R')+\eta_n^{(6)}-\epsilon\right)}\ . \end{aligned} \tag{24}$$

Thus, $P_3\underset{n\to\infty}{\longrightarrow}0$ if $S_2-R'<I(V;Y|U)$, or equivalently

$$\begin{aligned}R' &> S_2-I(V;Y|U)>I(V;X|U)-I(V;Y|U)\\ &\overset{(f)}{=} I(V;XY|U)-I(V;Y|U)=I(V;X|UY)\ , \end{aligned} \tag{25}$$

where equality (f) stems from the Markov chain $U-V-X-Y$. Thus, since the total rate R is composed of $\hat{R}$ and R', we conclude that our scheme is achievable if $R>I(U;X)+I(V;X|UY)$.[1] Notice that we don't need to check the case that for the true s_2, $(\mathbf{v}^n(s_1,s_2),\mathbf{y}^n)\notin T_\delta^n(VY|\mathbf{u}^n(s_1))$. That is because we only decode under the decision H_0, and we are interested in the distortion only when this decision is correct. This means that $(\mathbf{x}^n,\mathbf{y}^n)\in T_\delta^n(XY)$. Together with the coding process and the Markov chain $U-V-X-Y$, the typicality of $\mathbf{v}^n,\mathbf{y}^n$ is assured through the Markov lemma.

We now know that our scheme allows the decoding of $\mathbf{v}^n$ with high probability when the rate is large enough. It remains to be shown that V (together with U and Y, which

[1] We ignored one more probability of error, which is the probability that $\mathbf{y}^n$ is not typical. This probability goes to 0 much like P_0, thanks to the AEM. In the calculation of P_3 it was inexplicitly assumed that $\mathbf{y}^n$ is typical.

are also known at node B) is enough to recover X with average distortion D. We choose a (possibly suboptimal) decoder, that decodes x_i only from u_i, v_i and y_i:

$$\begin{aligned} d(\mathbf{x}^n, \hat{\mathbf{X}}^n(\mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n)) &= \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{X}(u_i, v_i, y_i)) \\ &\overset{(g)}{=} \frac{1}{n}\sum d(x, \hat{X}(u,v,y)) N(x,u,v,y|\mathbf{x}^n, \mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n) \\ &\overset{(h)}{\leq} \mathbb{E}\left[d(X, \hat{X}(UVY))|H_0\right] \\ &\quad + \sum\left(\frac{1}{n}N(x,u,v,y|\mathbf{x}^n, \mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n) - p(x,u,v,y)\right) \\ &\overset{(i)}{\leq} \mathbb{E}\left[d(X, \hat{X}(UVY))\middle| H_0\right] + d_{\max}||\mathcal{X}||||\mathcal{U}||||\mathcal{V}||||\mathcal{Y}||\delta_n \ , \end{aligned} \tag{26}$$

where the summation in (g) and (h) is over all the possible letters in the respective alphabets of the RVs $(x,u,v,y) \in \mathcal{X}\times\mathcal{U}\times\mathcal{V}\times\mathcal{Y}$ and inequality (i) holds since $(\mathbf{x}^n, \mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n) \in T_\delta^n(XUVY)$. Since $\delta_n \underset{n\to\infty}{\longrightarrow} 0$, condition $D > \mathbb{E}\left[d(X, \hat{X}(UVY))|H_0\right]$ is sufficient to achieve distortion $D+\epsilon$ at node B. This concludes the proof of achievability.

V. Proof of Converse

Denote by $W = f(\mathbf{X}^n)$ the message sent from node A to node B. The rate can be bound as follows:

$$\begin{aligned} nR &\geq I(W;\mathbf{X}^n) \\ &\overset{(j)}{=} I(W;\mathbf{X}^n,\mathbf{Y}^n) = I(W;\mathbf{Y}^n) + I(W;\mathbf{X}^n|\mathbf{Y}^n) \\ &= \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) \\ &\qquad + \sum_{i=1}^{n} I(W;X_i|\mathbf{Y}^n,\mathbf{X}^{i-1}) \\ &= \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) \\ &\qquad + \sum_{i=1}^{n} I(W;X_i|Y_i,\mathbf{Y}_{i+1}^n,\mathbf{Y}^{i-1},\mathbf{X}^{i-1}) \\ &\overset{(k)}{=} \sum_{i=1}^{n} \big[I(W,\mathbf{Y}^{i-1};Y_i) \\ &\qquad + I(W,\mathbf{Y}_{i+1}^n,\mathbf{Y}^{i-1},\mathbf{X}^{i-1};X_i|Y_i)\big] \\ &= \sum_{i=1}^{n} \big[I(W,\mathbf{Y}^{i-1};Y_i) + I(W,\mathbf{Y}^{i-1};X_i|Y_i) \\ &\qquad + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\big] \\ &= \sum_{i=1}^{n} \big[I(W,\mathbf{Y}^{i-1};Y_i,X_i) \\ &\qquad + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\big] \\ &\overset{(l)}{=} \sum_{i=1}^{n} \big[I(W,\mathbf{Y}^{i-1};X_i) \\ &\qquad + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\big] \ . \end{aligned} \tag{27}$$

Here, (j) and (l) are due to the Markov chains $W - \mathbf{X}^n - \mathbf{Y}^n$ and $W - X_i - Y_i$, respectively. (k) stems from the fact that both sources X and Y are assumed to be jointly i.i.d. Defining $U_i \triangleq (W, \mathbf{Y}^{i-1})$ and $V_i \triangleq (U_i, \mathbf{Y}_{i+1}^n, \mathbf{X}^{i-1})$ the Markov chain $U_i - V_i - X_i - Y_i$ is satisfied since the sources X and Y are assumed to be jointly i.i.d, and the bound over the rate becomes

$$\begin{aligned} R &\geq \frac{1}{n}\sum_{i=1}^{n}\left[I(U_i;X_i) + I(V_i;X_i|U_i,Y_i)\right] \\ &= I(U;X) + I(V;X|UY) \ , \end{aligned} \tag{28}$$

with U and V defined through time-sharing as is subsequently shown in (31).

The error exponent can now be expressed as follows:

$$\begin{aligned} I\left(f(\mathbf{X}^n);\mathbf{Y}^n\right) = I(W;\mathbf{Y}^n) &= \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) \\ &= \sum_{i=1}^{n} I(U_i;Y_i) = nI(U;Y) \ , \end{aligned} \tag{29}$$

with the same definition of U_i. Thus, the converse over the error exponent is proved with equality.

Finally, the distortion at node B can be bound as follows. Define the function $\hat{X}_i$ as the ith coordinate of the estimate in node B:

$$\hat{X}_i(U_i,V_i,Y_i) \triangleq g_i(W,\mathbf{Y}^{i-1},Y_i,\mathbf{Y}_{i+1}^n) \ . \tag{30}$$

The component-wise mean distortion thus verifies

$$\begin{aligned} D+\epsilon &\geq \mathbb{E}\left[d(\mathbf{X}^n, g(W,\mathbf{Y}^n))\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d(X_i,\hat{X}_i(U_i,V_i,Y_i))\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d(X_Q,\hat{X}_Q(U_Q,V_Q,Y_Q))|Q=i\right] \\ &= \mathbb{E}\left[d(X_Q,\hat{X}_Q(U_Q,V_Q,Y_Q))\right] \\ &= \mathbb{E}\left[d(X,\hat{X}(U,V,Y))\right] \ . \end{aligned} \tag{31}$$

For the sake of this calculation, we use the fact that any U_i and V_i, as they were defined for this converse, contain the entire message W, as well as the past and future of Y. This concludes the converse of Proposition 1.

VI. Relation to the Information Bottleneck Problem

The information bottleneck problem (see e.g. [7]–[9] and references therein), is a well-known problem in the signal processing field. A RV X is produced by a source, whose distribution is known. It is assumed, that X cannot be sent losslessly to the desired destination, due to communication rate constraints. This Destination, however, is only interested in some specific features of the source. One interesting example is the speaker detection problem, in which the source produces short speech sequences. The destination is only interested in *identifying the speaker* of each sequence, out of a fixed number of possible speakers. Obviously, the information carried by X contains much more than the identity of the speaker, making the transmission of the entirety of the information carried by X redundant. Moreover, rate-distortion theory may not be the answer here, since the chosen distortion function is very likely to hide some information about the speaker (by

changing the pitch, for example), while conserving useless information to the decoder (like the words spoken).

The information bottleneck problem is thus defined as follows. Given a source X with alphabet $\mathcal{X}$ and a RV Y with alphabet $\mathcal{Y}$ representing the desired information at the decoder, find a RV U, such that $U-X-Y$ form a Markov chain, and which minimizes $I(U;X)$ while maximizing $I(U;Y)$. The Markovian relation between the three RVs implies that $I(U;X) \geq I(U;Y)$, which clarifies the name of the problem.

Ignoring the estimation part of rate-error-distortion region (for example by setting $D = D_{\max}$), the information bottleneck problem resembles the single-letter relation between rate and error exponent of the second type implied by Proposition 1. This result is quite surprising at a first glance, since the two problems are very different. The information bottleneck problem can be classified as a *non-binary* clustering problem, while Proposition 1 is the solution to a very specific *binary* hypothesis testing problem. More thorough investigation, however, can point to many similarities between the two problems. Intuitively, out of a list of realizations of X, the information bottleneck problem aims to find a mapping, such that a list $\mathbf{y}^n$ containing some relevant characteristic of the realizations in the original list $\mathbf{x}^n$, can be built at the decoder. This is done under the knowledge of all relevant probability distributions in the system. The HT against independence assumes that the decoder already acquired the list $\mathbf{y}^n$. The mapping U is now meant to check if both lists (of X and Y) "fit" the previously known joint probability. Obviously, the mapping that was used to build $\mathbf{y}^n$ from $\mathbf{x}^n$ under some known distribution could be helpful in making that decision.

Having clarified the relationship between the two problems, it is clear that the understanding of both could benefit from this result. Our proposition provides information theoretic formalism to the information-bottleneck approach, which defined the problem directly through single-letter expressions. HT against independence gains through the progress achieved in understanding the information bottleneck problem by the signal processing community, such as efficient algorithms for producing the auxiliary RV U [7].

VII. Binary Symmetric Source

In some cases, the region defined by Proposition 1 can be calculated analytically. We present such an example here. Consider the following source:

$$\begin{aligned} &X \sim \text{Bern}\left(\frac{1}{2}\right) \\ &\begin{cases} H_0: & Y = X + Z, \quad Z \sim \text{Bern}(p) \\ H_1: & Y \sim \text{Bern}\left(\frac{1}{2}\right) \perp X \ , \end{cases} \end{aligned} \tag{32}$$

with $\text{Bern}(p)$ being a Bernoulli RV with probability p for being 1, and $\perp$ signifying that X and Y are independent of each other. Under both hypotheses, the marginal distributions of both X and Y are similar. Thus, a decision can be reached only through cooperation between the nodes. In the following, the rate-error-distortion region for this problem is derived from (8).

Proposition 2. *The rate-error-distortion region for the binary symmetric example is given by*

$$\begin{cases} R = 1 - H_2\left(\alpha * \beta * p\right) + \theta\left[H_2\left(\alpha * p\right) - H_2\left(\alpha\right)\right] \\ E = 1 - H_2\left(\alpha * \beta * p\right) \\ D = \theta\alpha - (1-\theta)\, p \end{cases} , \tag{33}$$

for any $0 \leq \alpha, \beta \leq \frac{1}{2}$ *and* $0 \leq \theta \leq 1$*, where* $a * b = a(1-b) + b(1-a)$ *is the scalar convolution function.*

The proof is given in the following. Section VII-A gives the proof of achievability of Proposition 2, while Section VII-B proves the converse.

A. Proof of Achievability

In order to achieve the region proposed in Proposition 2, choose V as the output of a binary symmetric channel (BSC) with cross-over probability α when the input is X. Choose U as the output of another BSC, with cross-over probability β, when the input is V:

$$\begin{aligned} V &= X + W_1, \quad W_1 \sim \text{Bern}\left(\alpha\right) \ , \\ U &= V + W_2, \quad W_2 \sim \text{Bern}\left(\beta\right) \ . \end{aligned} \tag{34}$$

Calculating the expression for the error exponent, U and Y can be thought of as connected through a BSC with cross-over probability $\alpha * \beta * p$, which yields:

$$I(U;Y) = H(U) = H(U|Y) = 1 - H_2(\alpha * \beta * p) \ . \tag{35}$$

This complies with the expression proposed in (33). The relation between the second term in the expression for the rate and the amount of distortion expected can be calculated through the following two steps:

a) Setting $\hat{X} = g(Y,V) = V$, we have $\mathbb{E}\left[d(X,\hat{X})\right] = \alpha$. Note that all expectations henceforth are taken over the distribution imposed by H_0, and under the assumption that the decision H_0 was correct. Y and V can be thought of as being connected through a BSC with cross-over probability $\alpha * p$. Thus (9) results in

$$\begin{aligned} R_a &= I(U;Y) + \left[I(V;X) - I(V;Y)\right] \\ &= 1 - H_2(\alpha * \beta * p) + \left[H_2(\alpha * p) - H_2(\alpha)\right] \ . \end{aligned} \tag{36}$$

b) In this part we let V be degenerate and $\hat{X} = g(Y,V) = Y$. At a first glance, this seems to contradict the requirement of the Markov chain $U-V-X-Y$. However, this is equivalent to defining V as in (34) and choosing not to transmit it. We then have $\mathbb{E}\left[d\left(X,\hat{X}\right)\right] = p$. Since in this case $I(V;X) - I(V;Y) = 0$, we have

$$R_b = I(U;Y) = 1 - H_2(\alpha * \beta * p) \ . \tag{37}$$

Now let $0 \leq D \leq p$ be given and say that θ, α are such that $D = \theta\alpha + (1-\theta)p$. Since $R(D)$ is convex (for a given error exponent E),

$$\begin{aligned} R(E,D) &= R(\theta\alpha + (1-\theta)p) \\ &\leq \theta R(\alpha) + (1-\theta)R(p) \\ &= \theta R_a + (1-\theta)R_b \\ &\leq 1 - H_2(\alpha * \beta * p) + \theta\left[H_2(\alpha * p) - H_2(\alpha)\right] \ . \end{aligned} \tag{38}$$

Thus, any triplet (R,E,D) that complies with Proposition 2 is achievable through this scheme, and the proof of achievability is complete.

B. Proof of Coverse

Proposition 1, along with the development in (9), implies that the optimal region, for any specific example of hypothesis testing against independence, is comprised of two RVs, such that the Markov chain $U - V - X - Y$ is respected. Moreover, it implies that with these optimal auxiliary RVs, the required rate is comprised of two independent parts - one part dedicated to detection and the other to estimation. Thus, the proof of the converse to Proposition 2 can be divided, much like the proof of achievability, into two separate parts - one defining the trade-off between the rate and the error exponent, while the other defines the trade-off between the rate and the distortion.

Starting with the relation between the rate and the error exponent, Proposition 1 implies that

$$E \leq I(U;Y) = H(Y) - H(Y|U) = 1 - A \ , \qquad (39)$$

while

$$R \geq 1 - A + \theta \left[I(V;X) - I(V;Y) \right] \ , \qquad (40)$$

with A defined as $A \triangleq H(Y|U)$. Ignoring the second term in the expression for the rate, the trade-off between rate and error exponent is clear, and is given through A. Obviously, $A \leq H(Y) = 1$. In addition,

$$A \geq H\left(H^{-1}\left(H(X|U) \right) * p \right) \ , \qquad (41)$$

which stems from Ms. Gerber's Lemma (see e.g. [12]). In order to allow the exploration of the entire region defined by the bounds over A, we define $\gamma \triangleq H^{-1}\left(H(X|U) \right)$. Thus, trade-off between rate and error exponent becomes

$$\begin{aligned} E &\leq 1 - H_2(\gamma * p) \\ R &\geq 1 - H_2(\gamma * p) + \theta \left[I(V;X) - I(V;Y) \right] \ . \end{aligned} \qquad (42)$$

In the second part of the proof, it needs to be demonstrated that, once the decision H_0 has been (correctly) made, the optimal estimation region, defined by the rate-distortion relation $\min_{\mathbb{E}\left[d(X,\hat{X})\right] \leq D} \left[I(V;X) - I(Y;X) \right]$, is in agreement with Proposition 2. This proof has already been given in [6] and is thus omitted from this paper. Defining V as the output of a BSC with cross-over probability α when X is in the input of the channel, as was shown to be optimal in [6], and keeping in mind the Markov chain implied by Proposition 1, it is clear that $\gamma = H^{-1}\left(H(X|U) \right) \geq \alpha$. Thus, γ can be expressed as $\gamma = \alpha * \beta$ for some $0 \leq \beta \leq \frac{1}{2}$, which completes the proof.

C. Numerical Results

We now present numerical results for the BSC case of testing against independence. Fig. 2 shows six curves, each representing the trade-off between the desired error exponent of the second type and the resulting average distortion of the source estimation, for a fixed value of available rate and for $p = \frac{1}{4}$. Unsurprisingly, all curves are non-decreasing, meaning that when the probability of error is exponentially smaller, the amount of rate left for estimation is smaller, resulting in a more crude estimation.

The maximally achievable error exponent in this case is $E_{\max} = I(X;Y) = 1 - H_2(p) \approx 0.1887$. It can be seen that when $R < E_{\max}$, the average distortion reaches its maximal value $D = p = 0.25$ for $E < E_{\max}$. Any exponent bigger than the value for which this happens is unachievable with this rate, since the desired exponent would demand more rate than available. When $R > E_{\max}$, further enlarging the rate allows for better distortion, for the same values of error exponent.

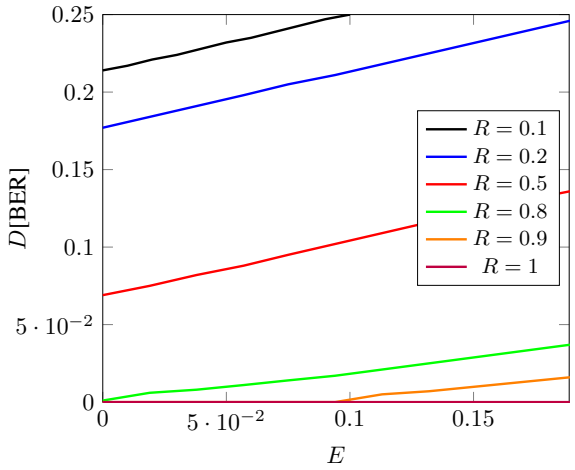

Fig. 2: Numerical results of the optimal average distortion as a function of the desired error exponent of the second type, for different amounts of available rate and for $p = \frac{1}{4}$

Note especially the curves of $R = 0.9$ and $R = 1$. Here, the rate complies with $R > H_2(p)$. According to the Slepian and Wolf principle (see e.g. [10]), this rate is enough to transmit $\mathbf{x}^n$ to node B without distortion, when no detection is necessary. Indeed, it can be seen that for any choice of error exponent that ensures enough available rate for estimation, zero-distortion is achievable. The curve for $R = 1$ is thus almost invisible, as in this case enough rate is available for estimation, for any achievable choice of error exponent.

VIII. Concluding Remarks

In this paper, binary hypothesis testing against independence was considered. The optimal rate-error-distortion region for detection and estimation was presented. It was shown than when testing against independence, the optimal solution is to divide the problem into two distinct problems. Detection is performed optimally as in [2], while estimation can be done by treating side information at the decoder as in [6]. It was shown that the information required for detection is useless for the estimation stage. An interesting and surprising relation between the binary hypothesis testing against independence and the information bottleneck problem was found, and promises interesting future research directions. Finally, a binary symmetric example was shown, for which the optimal region can be calculated explicitly.

IX. Acknowledgment

This research has been supported by the ERC Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering).

REFERENCES

[1] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, ser. Springer Texts in Statistics.

[2] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *Information Theory, IEEE Transactions on*, vol. 32, no. 4, pp. 533–542, Jul 1986.

[3] T. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, Nov 1987.

[4] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4666–4681, Oct 2008.

[5] S. Rahman and A. Wagner, "On the optimality of binning for distributed hypothesis testing," *Information Theory, IEEE Transactions on*, vol. 58, no. 10, pp. 6282–6303, Oct 2012.

[6] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 22, no. 1, pp. 1–10, Jan 1976.

[7] N. Tishbi, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37-th Annual Allerton Conference on Coomunication, Control and Computing*, 1999, pp. 368–377.

[8] S. Gordon, H. Greenspan, and J. Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct 2003, pp. 370–377 vol.1.

[9] C. R. Shalizi and J. P. Crutchfield, "Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction," *Advances in Complex Systems*, vol. 5, no. 01, pp. 91–95, 2002.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley & Sons, 1991.

[11] J. Villard and P. Piantanida, "Secure multiterminal source coding with side information at the eavesdropper," *Information Theory, IEEE Transactions on*, vol. 59, no. 6, pp. 3668–3692, Jun 2013.

[12] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.