Détection binaire distribuée sous contraintes de communication

Gil Katz

université PARIS-SACLAY

CentraleSupélec

N° national de thèse : 2017SACLC001

THESE DE DOCTORAT DE L'UNIVERSITÉ PARIS-SACLAY PREPARÉE À CENTRALESUPÉLEC

ECOLE DOCTORALE N°580
Sciences et technologies de l'information et de la communication (STIC)

Spécialité : Réseaux, Information et Communications

Par

M. Gil Katz

Distributed Binary Detection with Communication Constraints

Thèse présentée et soutenue à Gif-sur-Yvette, le 6 Janvier 2017

Composition du jury :

Prof. Te Sun Han,	University of Tokyo	Rapporteur
Prof. Deniz Gunduz,	London Imperial College	Rapporteur
Dr. Pierre Duhamel,	CNRS/CentraleSupélec	Examinateur, Président du Jury
Prof. Stephane Boucheron,	Université Paris Diderot	Examinateur
Prof. Gersende Fort,	Institut Mathématiques de Toulouse	Examinatrice
Prof. Mérouane Debbah,	Chaire LANEAS - CentraleSupélec	Directeur de Thèse
Prof. Pablo Piantanida,	L2S, CentraleSupélec	Co-Directeur de Thèse

université PARIS-SACLAY

ÉCOLE DOCTORALE

Sciences et technologies de l'information et de la communication (STIC)

Title: Distributed Binary Detection with Communication Constraints

Keywords: Hypothesis testing, Error exponents, Lossy source coding, Re-distortion

Abstract: In recent years, interest has been growing in research of different autonomous systems. From the self-driving car to the Internet of Things (IoT), it is clear that the ability of automated systems to make autonomous decisions in a timely manner is crucial in the 21st century. These systems will often operate under strict constraints over their resources. In this thesis, an information-theoretic approach is taken to this problem, in hope that a fundamental understanding of the limitations and perspectives of such systems can help future engineers in designing them.

Throughout this thesis, collaborative distributed binary decision problems are considered. Two statisticians are required to declare the correct probability measure of two jointly distributed memoryless processes, denoted by $\mathbf{X}^n = (X_1, \ldots, X_n)$ and $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$, out of two possible probability measures on finite alphabets, namely P_{XY} and $P_{\bar{X}\bar{Y}}$. The marginal samples given by $\mathbf{X}^n$ and $\mathbf{Y}^n$ are assumed to be available at different locations.

The statisticians are allowed to exchange limited amounts of data over a perfect channel with a maximum-rate constraint. Throughout the thesis, the nature of communication varies. First, only unidirectional communication is allowed. Using its own observations, the receiver of this communication is required to first identify the legitimacy of its sender by declaring the joint distribution of the process, and then depending on such authentication it generates an adequate reconstruction of the observations, satisfying an average per-letter distortion. The performance of this setup is investigated through the corresponding *rate-error-distortion region* describing the trade-off between: the communication rate, the error exponent induced by the detection and the distortion incurred by the source reconstruction.

In the special case of *testing against independence*, where the alternative hypothesis implies that the sources are independent, the optimal rate-error-distortion region is characterized. The case of "general hypotheses" is also investigated. A new achievable rate-error-distortion region is derived based on the use of non-asymptotic *binning*, improving the quality of communicated descriptions. It is shown that the error exponent is further improved through the introduction of a new approach. Benefits of the proposed methods are demonstrated through numerical analysis.

A different scenario is then considered, by which the statisticians are required to reach a conclusion through a bidirectional link. This allows for the consideration of multiple rounds of interactions, which differs from previous work. A single round of interaction is considered before the result is generalized to any finite number of communication rounds. A feasibility result is shown, guaranteeing the achievability of an error exponent for general hypotheses, through information-theoretic methods. The special case of testing against independence is revisited as being an instance of this result for which also an unfeasibility result is proven, thus proving optimality, at least for one round of communication. A second special case is studied where zero-rate communication is imposed, for which it is shown that interaction does not improve asymptotic performance.

Acknowledgments

Firstly, I would like to thank Pablo Piantanida for his guidance and for working closely with me on the research in this thesis. Your input was invaluable and working with you was a pleasure. I would also like to thank Mérouane Debbah for accepting me to this PhD, as well as the advice given along the way. Thank you also to Shlomo Shamai, for recommending me for this PhD as well as his input along the way.

I would like to thank Eliza Dias for taking care of everything from resident visas to conference travel arrangements. The conditions may have not always been easy, but it was always a pleasure to come into your office. Thank you also to Anne Batalie, without whom I wouldn't have survived the annual re-inscription to the thesis.

A big thank you to my friends and colleagues in Supélec. Axel Müller, Luca Rose, Apostolos Karadimitrakis, Matthieu de Mari, Matha Deghel, Romain Couillet, Evgeny Kusmenko, Stephne Mijovic, Hafiz Tiomoko Ali, Clément Feutry, Jérôme Gaveau, Meryem Benammar, Julien Floquet, Meysam Sadeghi, Karim Tadrist, Apostolos Destounis, German Bassi, I cherish your friendship and wish you all the best. Kenza Hamidouche, thank you especially for the emotional support through the naturalization process, hopefully it will pay off for both of us soon.

I would also like to thank the entire team of CrossFit Original Addicts Paris as well as all of its members, for maintaining what has been my second home throughout this period. The toughest days were made easier by the knowledge that they would end in a magical place, where I get to meet great people, have fun and move. In three years I spent more than 550 hours with you and I regret none of them (ok, maybe just one).

I would like to thank my friends in Israel who kept in touch with me despite the distance, and made the transition easier. Thank you Shani Recher, Lior Zagiel, Yehonatan and Osnat Gida, Roi Karasik, Edan Leventhal, Erez Druk, Jonathan Fisher and Yonathan Aflalo.

Finally, I would like to thank my parents, David and Nitza Katz, for putting up with my nonsense on Skype. Thank you to my entire family for supporting me, and especially to Maayan Dana, who is the smartest kid I know.

Thank you all.

Contents

Acronyms and Abbreviations

AEP	asymptotic equipartition property
BSC	binary symmetric channel
BSS	binary symmetric source
HT	hypothesis testing
IB	Information Bottleneck
KL	Kullback-Leiber (divergence)
MAC	multiple access channel
PM	probability measure
pmf	probability mass function
RV	random variable
SPRT	sequential probability ration test
wrt	with relation to
WSN	wireless sensor networks

List of Figures

Chapter 1

Introduction

1.1 Overview

The field of hypothesis testing (HT) is comprised of different problems, in which the goal is to determine the probability measure (PM) of one or more random variables (RVs), based on a number of available observations. Considering binary HT problems, it is assumed that this choice is made out of two possible hypotheses, denoted the null hypothesis H_0 and the alternative hypothesis H_1. Each of the hypotheses implies a different probability distribution, usually denoted P_0 and P_1, respectively. An overview of different HT problems and the approaches to their solutions can be found in [1].

The problem of Binary HT is formally defined by two types of error events, formally known as Type I and Type II. Denote by α_n the probability of error of Type I, defined to be the event that H_1 is chosen despite H_0 being true. The probability of an error event of type II, defined to be the event that H_0 is chosen despite H_1 being true, is denoted by β_n. Clearly, there is a trade-off between these two probabilities - enforcing $\alpha_n = 0$, for example, can be done easily if we are willing to contend with $\beta_n = 1$. One common way of investigating that trade-off, which will be considered throughout this thesis, is to examine the exponential rate of decay of the error probability of the second type, i.e., $-\lim_{n\to\infty} \frac{1}{n} \log \beta_n(\epsilon)$, while imposing a fixed constraint over the error probability of the first type, i.e., $\alpha_n \leq \epsilon$ ($\epsilon > 0$).

Let $\{X_i\}_{i=1}^{\infty}$ be an independent and identically distributed (i.i.d) process, commonly refereed to as a *memoryless process*, taking values in a countably finite alphabet $\mathcal{X}$ equipped with probability measures P_0 or P_1 defined on the measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, where $\mathcal{B}_{\mathcal{X}} = 2^{\mathcal{X}}$. Denote $\mathbf{X}^n = (X_1, \ldots, X_n)$ the finite block of the process following the product measures P_0^n or P_1^n on $(\mathcal{X}^n, \mathcal{B}_{\mathcal{X}^n})$. Let us denote by $\mathcal{P}(\mathcal{X})$ the family of probability measures in $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, where for every $\mu \in \mathcal{P}(\mathcal{X})$, $f_\mu(x) \coloneqq \frac{d\mu}{d\lambda}(x) = \mu(\{x\})$ is a short-hand for its probability mass function (pmf). The optimal error exponent for the Type II error probability of the binary HT problem is well-known and given by *Stein's*

Lemma [1–3] to be:

$$-\lim_{n\to\infty}\frac{1}{n}\log\beta_n^\star(\epsilon)=\mathcal{D}(P_0\|P_1)\ , \tag{1.1}$$

(see Lemma 11), where P_0 and P_1 are the probability distributions implied by hypotheses H_0 and H_1, respectively, and $\mathcal{D}(\cdot\|\cdot)$ is the *Kullbeck-Leiber divergence* satisfying that the measure P_0 is *absolutely continuous* with respect to (wrt) P_1, $P_0 \ll P_1$ (i.e., $P_0(a)=0$ for every $a\in\mathcal{X}$ such that $P_1(a)=0$). Note that in this case, the optimal error-exponent of the probability of error of Type II does not depend on the constraint ϵ, imposed upon the probability of error of Type I. This characteristic is referred to as a *strong property*. It is discussed thoroughly in Section 1.4 of this introduction.

In many scenarios, the realizations of different parts of a random process are available at different physical locations (with different statisticians) in the system (see Fig. 2.1). If it were possible to transmit all signals to some central location with negligible cost and delay, then the previous theory is in principle applicable. However, due to practical considerations such as energy cost, reliability, survivability, communication bandwidth, compartmentalization, there is never total centralization of information in practice [4]. In this thesis, we focus on the problem of distributed hypothesis testing where it is assumed that realizations of different memoryless sources are observed at different physical locations and thus, nodes are subject to satisfy different types of communication constraints. In this case, a new question arises –for different types of constraints over the data exchange between the nodes, what is the optimal error exponent to the error probability of Type II, under a fixed constraint over the error probability of Type I? This work attempts a modest step in the direction of a theory for distributed testing based on lossy data compression, which seems to offer a formidable mathematical complexity (see [5] and references therein).

In this thesis, we compose together two stories. One is from statistics concerning binary HT originating in the works of Wald [6, 7]. The other story is from information theory concerning the case of *unidirectional* data exchanges where only one statistician can share information with the other one, due to [8, 9]. The problems in HT we choose to focus on differ both in the *nature of communication* which is imposed, as well as the *task* the statisticians are required to complete. In the first problem, only *unidirectional* communication is allowed, from node A to node B. In this case, node B can be thought of as the statistician, while node A performs as a "helper". Having received the communication from the helper, the statistician is required to both *detect* (by declaring the correct hypothesis) and *estimate* the vector of realizations seen by the helper. Note that while the metric by which the performance of the detection step is assessed is its error exponent, the assessment of the estimation step is done through the average distortion of the signal [10]. By making this choice we allow ourselves to explore the trade-offs between two demands of different nature –a *timely* decision and an *average* distortion.

In many practical cases, the scenario discussed above can be thought of as one where the decoder is required to perform *user authentication* before estimating the information sent from the encoder. Consider for example the case where two users (known in literature as "Alice" and "Bob") attempt to agree on a common key by using a *perfect public channel*

and *correlated sources* (see e.g., [11–15] and references therein). Assume now that, unlike in most cases in literature, the attacker "Eve" tries to masquerade as Alice, and send Bob information over the channel, such that she establishes a private key with Bob, instead of Alice. Eve may hold realizations of a source that is also correlated with Bob's (although more loosely than Alice's), or none at all (which would correspond to the special case of *testing against independence*, discussed throughly in this thesis. In this case, Eve may use the same marginal distribution as Alice, however she cannot artificially create a correlation with Bob's realizations). The model presented here answers the question: How *quickly* can Bob declare he is being contacted by an imposter, while still being able to establish a common key with Alice, in case the transmission originated from her? This threat of compromise of the receiver's authentication data is motivated by situations in multiuser networks –such as automatic fault diagnosis– where the receiver is often the system itself, and which cannot be treated by conventional cryptography, requiring recourse to new techniques (e.g., image authentication [16, 17] and smart grids [18, 19]).

The second scenario discussed in this thesis involves *bidirectional collaborative* binary HT. In this scenario (see Figure 4.1), the two nodes are assumed to be connected by a perfect bidirectional link with a *sum-rate* constraint. It is further assumed that the available resources for interaction can be divided between the statisticians in any way that would benefit performance, and that without loss of generality no importance is given to the location at which the decision is made – as the decision can always be transmitted with sub-exponential resources. First, we concentrate on a special case where only one "round of interaction" (only a query and its reply) is allowed between the statisticians, i.e., a decision is made after each statistician communicates one statistics, which will be commonly referred to as a message. This scenario was first studied in [20] for a special case called *testing against independence*. While the scenario studied in this thesis borrows ideas from [20], the mathematical tools are fundamentally different since these rely on the *method of types* [21], as it was the case to deal with general hypotheses in [9]. The results are then extended for any finite number of interaction rounds.

Much like in the case of unidirectional communication in Chapter 3, optimality can be shown here for the special case of testing against independence. Interestingly, we can only show optimality for one round of communication. While it cannot be said that the strategy proposed in Chapter 4 is necessarily *suboptimal* when testing over multiple rounds, it can be shown that the tools used to prove optimality for a single round of communication do not suffice for the more general case of any finite number of rounds. Intuitively, it can be said that after the first round of communication is completed, the *information* available at each of the sides of the system is *no longer independent*, even under hypothesis 1, which assumes independent sources. Thus, it is easy to accept that optimality should remain allusive for this case, as long as it is not shown for the case of general hypotheses.

The bidirectional scenario, while apt to continue to constitute a model for questions in secret key distribution (this time over a perfect bidirectional public channel) as described above, may also be used as a model for complex automatic decision-making in a timely manner. The Internet of Things (IoT) (see e.g., [22–24] and references therein

for an overview of the subject) emerges recently as a practical "playground" for relevant scenarios. [25], for example, focuses on emergency response systems (ERS), which may assist governments' capability in responding to severe events. The importance of timely decisions is clear in this case, while constraints on the resources are mostly due to limited funds. Naturally, the well-established field of wireless sensor networks (WSN, see e.g., [26–29]), along with work on HT, may be critical in modeling such practical scenarios.

1.2 Summary of Related Works

Some of the first contributions on binary HT are due to Wald [6, 7], where an optimal course of action is given, by which a sequential probability ratio test (SPRT) is used. It was shown that the expected number of observations required to reach a conclusion is lower than any other approach, when a similar constraint over the probabilities of error is enforced.

Definition 1 (SPRT, [7])**.** *A sequential probability ratio test of two hypotheses H_0 and H_1, implying probability distributions P_0 and P_1, respectively, is defined with the aid of two positive numbers $A^\star > 1$ and $B^\star < 1$, as follows: Write the probability*

$$P_{ij} = \prod_{k=1}^{j} P_i(x_k) \ , \tag{1.2}$$

with x_k being the k-th argument of the vector $\mathbf{x}$ being tested. We say that the number of necessary realizations $n = j$ if j is the lowest natural number such that $\frac{P_{1j}}{P_{0j}} > A^\star$ or $\frac{P_{1j}}{P_{0j}} < B^\star$. If $\frac{P_{1j}}{P_{0j}} > A^\star$ hypothesis H_1 is accepted, and if $\frac{P_{1j}}{P_{0j}} < B^\star$ hypothesis H_0 is accepted.

As stated above, the SPRT is the optimal test, in the sense that out of all the tests that lead to the same probabilities of error, the SPRT would do so with the least amount of realizations (or, equivalently, the "quickest").

Among the first works that started enforcing constraints on the basic HT problem, which are independent from the statistical nature of the data, are references [30, 31]. The single-variable HT is considered, and the enforced constraint is related to the *memory* of the system, rather than to communication between different locations. It is assumed that a realistic system cannot hold a large number of observations for future use, and thus at each step a function must be used that would best encapsulate the "knowledge" gained from the new observation, combined with the compressed representation of previous observations. Namely, assuming that after each observation the data must be summarized by an m-valued statistic $T_n \in \{1, 2, \ldots, m\}$, updating this statistic is done through a function $T_{n+1} = f(T_n, X_{n+1})$. By presenting an algorithm and defining the two probabilities of error discussed above, α_n and β_n, the author of [30] shows that a 4-valued statistic is enough in order to bring these two values to zero as $n \to \infty$. This problem was then

revisited in [32,33], which are motivated by new scenarios in which memory efficiency is an important aspect, such as satellite communication systems. [33] explores the possibilities of simultaneous exponential decays for both error probabilities.

Stein's Lemma (see e.g., [1, 2]) is the first result that takes an information-theoretic approach to the subject of HT. By considering the limit where the number of observations $n \to \infty$, it is shown that the optimal error exponent for the error probability of Type II, under any fixed constraint over the error probability of Type I, is given by the Kullback-Leiber (KL) divergence. Later [34] proves an important property by which when $\alpha_n \equiv \exp(-nc) \to 0$ as $n \to \infty$, then $\beta_n \to 0$ or $\beta_n \to 1$, exponentially depending on the rate of decay $c > 0$. Blahut [35] investigates a similar scenario where both error probabilities are required to decrease exponentially with n, and proposes a function $e(c)$, non-increasing and convex, such that $\alpha_n \leq \exp\{-ne(c)\}$ and $\beta_n \leq \exp\{-nc\}$ are simultaneously satisfied, for n large enough. The function $e(c)$ is defined through the KL-divergence (referred to in [35] as the discrimination) as follows:

Definition 2. *Let P_0 and P_1 be two probability distributions in $\mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is assumed to be a finite alphabet. Let*

$$\mathcal{D}(P_0||P_1) = \sum_{a\in\mathcal{X}} P_0(a) \log \frac{P_0(a)}{P_1(a)} \tag{1.3}$$

be the KL-divergence (or discrimination) between the two probability distributions. The function $e(c)$, mentioned above, is defined as follows:

$$e(c) = \min_{\tilde{P}\in\mathcal{P}_c} \mathcal{D}(\tilde{P}||P_1) \ , \tag{1.4}$$

where

$$\mathcal{P}_c = \{\tilde{P} \in \mathcal{P}(\mathcal{X}) : \mathcal{D}(\tilde{P}||P_0) \leq c\} \ . \tag{1.5}$$

An information-theoretic approach becomes a natural choice when communication constraints are introduced to the model. Distributed HT with communication constraints was the focus of the seminal works [8, 9]. Both of these works investigated binary decisions in presence of a helper, i.e., unidirectional communication, and propose a feasible error exponent for β_n while enforcing a strict constraint over α_n. Although both of these approaches achieve optimality for the case of testing against independence, where it is assumed that under the alternative hypothesis H_1 the samples from (X, Y) are independent with the same marginal measures implied by H_0, optimal results for the case of general hypotheses remain allusive until this day. As both of these works are eminent for the research performed throughout this thesis, we discuss the result presented in them in depth in Chapter 2. In [36] a similar scenario is considered for parameter estimation with unidirectional communication. Here, the mean square-error loss in estimating the parameter θ was considered instead of exponential decay of the error probability. A Cramér-Rao type bound is established, and its asymptotic achievability is proven under certain conditions, in case of a finite alphabet $\mathcal{X}$, for the realizations observed by the helper.

Improving upon the results of [8, 9] by using further randomization of the codebooks, referred to as "random binning", was first briefly suggested in [37], but never fully analyzed in the general case, to the best of our knowledge. [38] proposes binning as an optimal approach in a special case called *testing against conditional independence*. Here, it is assumed that given a third RV Z (available at the decoder), (X, Y) are independent under hypothesis H_1. [38] is also the first paper that discusses the "dangers" of employing binning in problems in HT. Since the binning process may also induce errors, it is unclear if the benefits of using this method outweigh the losses, when the error exponent is concerned. That is because reliable decoding of the "bin index" is required in the presence of side information uncertainty (e.g., similarly to problems under channel uncertainty [39]). We discuss this trade-off thoroughly in Chapter 3. Note that despite proving optimality in the special case of testing against conditional independence, it is still not clear whether examples exist or not, in which binning is *strictly beneficial*, when compared to the traditional non-binning approach.

A special case referred to as HT under "complete data compression" was studied in [9]. In this case, it is assumed that node A is allowed to communicate with node B by sending only one bit of information. A feasible scheme was proposed and its optimality proved. The much broader scenario, by which codebooks are allowed to grow with n, but not exponentially fast, was studied in [40]. Interestingly, it was shown that this scenario does not offer any advantage with relation to complete data compression, in terms of the error-exponent of β_n. This setting, referred to as zero-rate communication, was recently revisited in [41], where both α_n and β_n are required to decrease exponentially with n.

As was mentioned above, the task in Chapter 3 is comprised of two parts. First, the statistician at node B needs to declare the correct probability distribution governing the pair of RVs, and then he is required to reproduce the vector of realizations seen by node A, with maximum average distortion D. This problem is closely connected to the case of successive refinement for the Wyner-Ziv problem [42]. Here, the Wyner-Ziv problem for source estimation with side information at the decoder [43] is investigated for a system with multiple decoders, each having different side information. The problem presented in this thesis can be thought of as a complication of this problem – not only does the encoder needs to encode to satisfy both requirements as before, but the decoder himself also does not know the "value" of its own realizations and the correct way to use them, before communication starts. Unfortunately, even for the simpler case of successive-refinement for the Wyner-Ziv problem, optimality results remain allusive [42, 44, 45]. Optimality can be achieved, however, for the special case where side information may be absent [46, 47]. In this case, similar to the joint problem of testing against independence and source reconstruction investigated in this thesis, out of the two decoders, one holds side information, while the other does not. Still, each decoder knows its "identity", which differs from our problem. The optimal region in this case is given by:

Lemma 1 (Rate-distortion when side information may be absent [46])**.** *Let $(X, Y, P(x, y))$ be a discrete memoryless 2-source with generic RVs X and Y. for $i \in \{0, 1\}$ let $\hat{\mathcal{X}}_i$ be the construction alphabet and let*

$$d_i : \mathcal{X} \times \hat{\mathcal{X}}_i \to [0, \infty) \tag{1.6}$$

be a distortion measure. The expected distortion $\mathbf{D} = (D_0, D_1)$ *for a given code (consisting in this case of an encoding function and two decoding functions) is given by*

$$D_i = \mathbb{E}d_i(\mathbf{X}, \hat{\mathbf{X}}_i) = \mathbb{E}\frac{1}{n}\sum_{k=1}^{n} d_i(x_k, \hat{x}_{ik}) \ , i = \{0, 1\} \ . \tag{1.7}$$

The rate-distortion function in this case is

$$R(\mathbf{D}) = \min_{\mathcal{P}(\mathbf{D})} \left[I(X; W) + I(X; U|WY) \right] \ , \tag{1.8}$$

where the minimum is over the set $\mathcal{P}(\mathbf{D})$ *of all the RVs* $(W, U) \in \mathcal{W} \times \mathcal{U}$*, jointly distributed with the generic RVs* (X, Y) *such that* $(W, U) - X - Y$ *form a Markov chain and there exist functions* $\hat{X}_0(W, U, Y)$ *and* $\hat{X}_1(W)$ *such that* $\mathbb{E}_i d_i(X, \hat{X}_i) \leq D_i$ *for* $i \in \{0, 1\}$*. The cardinalities of the alphabets of the auxiliary RVs satisfy the conditions* $|\mathcal{W}| \leq |\mathcal{X}| + 2$*,* $|\mathcal{U}| \leq (|\mathcal{X}| + 1)^2$*.*

Proof. Refer to reference [46]. □

In Lemma 1, the term "rate-distortion function" signifies the rate $R(\mathbf{D})$ is the minimum rate that achieves the vector of distortions $\mathbf{D}$. Note that this result is an *asymptotic one*, which differs from the nature of the results discussed above, studying the error exponent as a function of the block-length. The tension between non-asymptotic and asymptotic performance measures is one of the main points of interest of Chapter 3. [47] adds to this result the case where the side information may also be available at the encoder's end. Benefits of successive refinement for testing against independence are studied in [48]. Another special case studied in literature was the one of joint vector-Gaussian source and side information (at both decoders), under mutual information and distortion constraints, to be found in [49].

Hypothesis testing with interactive communication has unfortunately seen less treatment in literature, for the best of our knowledge. Interactive communication was considered for the problem of distributed binary HT within the framework of testing against independence in [20]. An achievable strategy was proposed for the case of single-round of communication (in which each node sends one statistic, before a decision is made), based on a coding scheme inspired by the seminal work of Kaspi [50]. In addition, it was claimed that the performance achieved by this strategy is optimal. Unfortunately, the proof of converse turned out to be problematic. We revisit this proof and show optimality for the case of single-round interaction against independence in this work. [51] attempted an extension of previous results to a scenario of multi-round testing against independence. Despite the claims of [51], it is still unclear if optimality in this case is possible. We try to explain the reasons for this in this thesis.

Another interesting branch of HT problems is the one focusing on tests of more than 2 hypotheses, commonly known as M-ary tests (see e.g., [52,53] for centralized scenarios and [54,55] for distributed ones). [56] points to an interesting connection between *Bayesian* M-ary tests and non-Bayesian binary ones, as studied in this thesis. The Bayesian framework

assumes that a probability distribution, referred to as the *prior* can be associated with the different hypotheses:

Definition 3. *Let H_i, $i \in \{0, \ldots, M-1\}$ be M hypotheses, each implying a different probability distribution $P_i \in \mathcal{P}(\mathcal{X})$. Let $Q_i \in \mathcal{P}(\{0, \ldots, M-1\})$ be a prior distribution of the hypotheses, i.e., the probability that hypothesis H_i is the correct one is Q_i. Finally, given a specific strategy $f(\cdot)$, operating on vectors $\mathbf{x}$ of length n, define the probabilities of erroneous detection to be*

$$\theta_i^n = \Pr\{\hat{H} \neq H_i | H_i\} = \sum_{\mathbf{x} \in \mathcal{X}^n} P_i(\mathbf{x}) \Pr\{f(\mathbf{x}) \neq i\} \ , \tag{1.9}$$

where $\hat{H}$ denotes the hypothesis chosen by the system, and the conditioning means that H_i is the true hypothesis in effect. The smallest average error probability for testing over vectors of length n is

$$\bar{\theta}^n = \min_f \left[\sum_{i=0}^{M-1} Q_i \theta_i^n \right] \ . \tag{1.10}$$

Clearly, if $M = 2$ calculating the probabilities of error α_n and β_n as defined above easily leads to the Bayesian average error, in case the priors Q_i are known. The opposite, however, is not true. In the case of [56], an equivalence was shown between Bayesian M-ary HT problems to non-Bayesian binary ones as follows:

Lemma 2. *Let an M-ary hypothesis testing problem be defined by M probability distributions $P_i \in \mathcal{P}(\mathcal{X})$. Define the RV V with alphabet $\mathcal{V} = \{0, \ldots, M-1\}$, denoting the true hypothesis in affect. Thus, the probability distribution $P_V(v)$ is the prior. Let any test be defined by a (possibly random) transformation $P_{\hat{V}|X} : \mathcal{X} \to \mathcal{V}$, where $\hat{V}$ denotes the RV associated to the test output. Denote the average error probability of the test as $\bar{\theta}(P_{\hat{V}|X})$.Minimizing over all possible conditional distributions $P_{\hat{V}|X}$ gives the smallest average error probability, namely*

$$\bar{\theta} = \min_{P_{\hat{V}|X}} \bar{\theta}(P_{\hat{V}|X}) \ . \tag{1.11}$$

The minimum error probability can be expressed as

$$\bar{\theta} = \max_{Q_Y} \alpha^{(\frac{1}{M})}(P_{VY}, Q_V \times Q_Y) = \max_{Q_Y} \sup_{\gamma \geq 0} \left\{ \Pr\left[\frac{P_{VY}(V,Y)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\} \ . \tag{1.12}$$

Here, $\alpha^{(\frac{1}{M})}(P_{VY}, Q_V \times Q_Y)$ denotes the optimal error probability of Type I, such that the error probability of Type II is at most $\frac{1}{M}$. $Q_V(v) \triangleq \frac{1}{M}$, $\forall v \in \mathcal{V}$, and the probability is computed with respect to P_{VY}. Moreover, a maximizing distribution Q_Y is given by

$$Q_Y^\star(y) = \frac{1}{\mu} \max_{v'} P_{VY}(v', y) \ , \tag{1.13}$$

where $\mu = \sum_y \max_{v'} P_{VY}(v', y)$ is a normalizing constant.

Proof. Refer to reference [56]. □

Through Lemma 2 it becomes clear that a good understanding of binary HT problems can be beneficial to the understanding of Bayesian M-ary HT problems. Specifically, a connection with problems in binary HT against independence is apparent. Recently, [57] investigated the error probabilities of M-ary tests (in contrast to previous work, that usually focus on error exponents [58, 59], and as is the case also in this thesis).

Other works in recent years evolve the problem of HT in many different directions. A few interesting examples are [60] (see references therein), which assumes a tighter control by the statistician throughout the process, allowing him to choose and evaluate the testing procedure through past information, and [61–63] which investigate HT in the framework of quantum statistical models. [64] considers an interesting distributed model, very different from the one considered in this thesis, by which the network grows with the number of realizations n. In this case, each node only sees a small part of the realizations, all belonging to a single RV. Here, detection is done through one-bit quantization at each node.

1.3 Related Problems

Before finishing the literary review, we would like to take a look at a few seemingly unrelated problems, which turn out to be surprisingly linked to the problem of distributed hypothesis testing. In this section we present two such problems, and attempt to explain the connection to the problem of HT, which may seem unintuitive at a first glance.

1.3.1 The Information Bottleneck

The information bottleneck (IB) was first introduced under this name by Tishbi *et al.* in [65] (see also [66]). Given some joint distribution $P_{XY}(x, y)$, the basic idea was to search for some compact description of X that maximally preserves the information about Y. In other words, Y can be thought of as the RV that represents the *characteristics* of X, which we would like to maintain through the compression process. As the joint distribution of X and Y is known (a fact which stands in contrast to the case of HT), the information bottleneck method proposes to compress X through a new RV U, such that $U - X - Y$ form a Markov chain (i.e., given X, U is independent of Y) and $R \geq I(U; X)$. Note that in this problem there is no transmission of information over a link. We use R to denote the constraint over the compression because of conventions in information theory, and not necessarily in order to denote the rate of communication. In order to maintain the maximum amount of information over the characteristics Y, under the constraints defined above, we set the goal as the maximization of $I(U; Y)$. Thus, the information bottleneck

problem can be summarized as follows:

$$\begin{aligned} \underset{P_{U|X}}{\text{maximize}} \quad & I(U;Y) \\ \text{subject to} \quad & I(U;X) \leq R \\ & U - X - Y \ . \end{aligned} \tag{1.14}$$

As the Markov chain $U - X - Y$ enforces the following inequality through the chain rule:

$$I(U;Y) \leq I(U;X) \ , \tag{1.15}$$

the name information *bottleneck* becomes clear.

Note that the information bottleneck approach differs from traditional rate-distortion. Specifically, when feature extraction is the desired goal as described above, it may be very difficult to define a suitable distortion measure. Taking the compression of photos as an example, rate-distortion can be a very successful approach when trying to transmit the photo over a link (or simply saving it to memory) under rate constraints, with the intent of reconstructing the photo for a later use. In that case, the rate-distortion approach may result in some noise (as a function of the available rate), but in most cases the photo could be reconstructed quite successfully. In some cases, however, the goal may not be reconstructing the photo. Instead, consider a case where the end-user is only interested in some characteristic of the photo, such as the types of objects in the picture (cars, animals, trees, people, etc.), the type of location (city, field, forest, etc.) or anything else. In such a case, it is very hard to define a distortion measure that would yield the desired result. This is one reason for the abundance of work on the information bottleneck in fields such as image representation ([67–69]), video ([70]), text classification ([71–73]), deep learning ([74]) and more.

Surprisingly, the formulation of the information bottleneck problem in (1.14) is identical to the *solution* of the distributed HT against independence problem with unidirectional communication (as seen in [8, 9] and in Chapter 3 of this thesis). At a first glance, this connection is hard to explain –The IB problem can be classified as a *non-binary* clustering problem, which stands in contrast to the very specific nature of a *binary* hypothesis testing problem against independence. We try to explain this connection intuitively here: Presented with a "list" (represented by a vector of realizations $\mathbf{x}$), the goal, as defined by the IB problem, is to be able to create a new list (represented by the vector $\mathbf{u}$), which represents, as best as possible, some characteristic of the data. In other words, the list $\mathbf{u}$ is highly connected to a third list, $\mathbf{y}$, which is hidden. When testing against independence over a distributed system, the two lists, $\mathbf{x}$ and $\mathbf{y}$ already exist. The question is –does the list $\mathbf{y}$ represent, with relation to the list $\mathbf{x}$, the characteristic we are interested in? Note that since we discuss the problem of testing against independence, the answer can only be either "yes" (dependent) or "no" (independent). Since both lists don't exist in the same physical place, and a rate constraint is imposed upon the communication from node A to node B, it is only natural that a solution that was good in order to represent $\mathbf{y}$ from $\mathbf{x}$ in the IB problem (where $\mathbf{y}$ is hidden) should also be good in order to check if $\mathbf{y}$ is related to $\mathbf{x}$ in the HT case.

Having clarified the relationship between the two problems, it is clear that the understanding of both could benefit from this result. The HT problem provides information theoretic formalism to the IB 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 [65, 75–77].

1.3.2 The Ahlswede-Gacs-Körner Bound

In [78], the authors investigate the following statistical problem: Consider a sequence (X_i, Y_i) of independent identically distributed pairs of RVs. For any pair of events $\{\mathbf{X}^n \in \mathcal{A}\}$ and $\{\mathbf{Y}^n \in \mathcal{B}\}$ satisfying $\Pr\{\mathbf{Y}^n \in \mathcal{B}|\mathbf{X}^n \in \mathcal{A}\} \geq 1-\epsilon$, and for non-negative real c, how small can $\Pr\{\mathbf{Y}^n \in \mathcal{B}\}$ be, in case $\Pr\{\mathbf{X}^n \in \mathcal{A}\} > \exp\{-nc\}$? In order to present the solution to this problem, we need the following definitions:

Definition 4. *Let $\mathcal{X}$ and $\mathcal{Y}$ be alphabets of the RVs X and Y, respectively, and let $W(y|x)$ be the transition probabilities for $x \in \mathcal{X}$, $y \in \mathcal{Y}$. For the n-th Cartesian power of $\mathcal{X}$ and $\mathcal{Y}$ define*

$$\mathbf{W}^n(\mathbf{y}^n|\mathbf{x}^n) = \prod_{i=1}^{n} W(y_i|x_i) \ . \tag{1.16}$$

The set $\mathcal{B} \in \mathcal{Y}^n$ is said to ϵ-decode the vector $\mathbf{x} \in \mathcal{X}^n$ if $\mathbf{W}^n(\mathcal{B}|\mathbf{x}) \geq 1-\epsilon$. Let $\Psi_\epsilon(\mathcal{B}) \in \mathcal{X}^n$ be the set of all the $\mathbf{x}$s which are ϵ-decoded by $\mathcal{B}$.

Let P_X be the measure given on X and P_Y the measure given on Y, both assumed to be i.i.d when vectors of length n are considered. Define

$$S_n(c,\epsilon) \triangleq -\frac{1}{n}\log \min_{-\frac{1}{n}\log P_X^n(\Psi_\epsilon(\mathcal{B}))\geq c} P_Y^n(\mathcal{B}) \ . \tag{1.17}$$

Lemma 3.

$$\lim_{n\to\infty} S_n(c,\epsilon) = \sup_{\substack{I(X;U)\leq c \\ U-X-Y}} I(U;Y) \ . \tag{1.18}$$

Proof. Refer to reference [78]. □

For a second time, it turns out that a problem, which seems quite different from testing against independence over a distributed system with a unidirectional link, turns out to be closely related to it. In fact, by looking closely, the two problems have many similarities and some differences. Here, for every set $\mathcal{B} \in \mathcal{Y}^n$, we want to find a corresponding set $\mathcal{A} \in \mathcal{X}^n$, such that two statements are simultaneously true:

$$\Pr\{\mathbf{Y}^n \in \mathcal{B}|\mathbf{X}^n \in \mathcal{A}\} \geq 1-\epsilon \ , \tag{1.19a}$$

$$\Pr\{\mathbf{X}^n \in \mathcal{A}\} > \exp\{-nc\} \ . \tag{1.19b}$$

Under these two constraints, we would like to find the smallest $\Pr(\mathbf{Y}^n \in \mathcal{B})$ possible. In fact, it could be said that we are looking for a good code for the problem of HT against independence, as discussed in [8]. If we succeed, we might be able to divide $\mathcal{X}^n$ into sets (representing codewords), such that the probability of each set is non-negligible (1.19b). Under hypothesis 0, the probability of error α_n is smaller then ϵ (1.19a), and under hypothesis 1, the probability of error β_n is as small as possible (minimizing $\Pr\{\mathbf{Y}^n \in \mathcal{B}\}$). Thus, the similarity of the two problems seems less surprising.

Nevertheless, note that there are indeed some differences between the two problems. Specifically, one could say that the approach here is to look at the trees (each codeword separately) but never at the forest (the codebook). We are not required to span $\mathcal{X}^n$ with the union of all sets $\mathcal{A}_i$ (in fact, there are no sets $\mathcal{A}_i$, only one set $\mathcal{A}$). Thus, a situation can be imagined, where while *for each* $\mathcal{B}$ a set $\mathcal{A}$ can be found that fulfills the constraints above, but many $\mathbf{x}^n \in \mathcal{X}$ do not belong to any of these sets, and thus cannot be encoded when considering the problem of HT against independence. If the probability of such sequences $\mathbf{x}^n$ does not become negligible with n, this indeed constitutes a problem for the HT case. Such small differences account for the single difference in the formula, namely that in [78], $|\mathcal{U}| \leq 3$ is enough. This stands in contrast to the case of HT (see Chapter 3).

1.4 Strong vs. Weak in Hypothesis Testing

In information theory, it is common to distinguish between two types of converse claims for any theorem. Consider as an example the channel coding theorem [79], which claims that communication of information over a channel is possible with vanishing probability of error $P_e \to 0$ when the block-length n grows if and only if it is done below the capacity of the channel. A weak converse to this theorem [80] shows that if the capacity of the channel is surpassed, the probability of error cannot be brought to 0 ($P_e \not\to 0$). A strong converse, however, shows that necessarily in this case $P_e \to 1$. While the difference between the two seems negligible, it is in fact significant –through the strong converse we know that surpassing the capacity of the channel is never a good idea, even if we are willing to accept some finite probability of error. In other words, the strong converse proves that working with a code that does not have a vanishing probability of error is pointless, as necessarily if such a code exists, there also exists a code with a vanishing error probability and the same rate.

Strong converse proofs exist in literature for both of the two main branches of point-to-point information theory –channel coding ([81]) and source coding (see [82]). Strong converses to some cases of source coding with a fidelity criterion (rate-distortion) can be found in [83, 84]. Strong converse proofs in network information theory are harder to find (as indeed all converse proofs are). One such example is the strong converse to the multiple access channel (MAC) in [85].

In the field of HT, the definitions of weak and strong converse proofs must be adjusted, in order to fit this different scenario. Throughout this thesis, we define the two probabil-

ities of error (of Type I and Type II) with their respective probabilities (α_n and β_n) as defined above. The goal throughout the different scenarios is to find the error exponent of β_n, while α_n is kept below some threshold $\epsilon \in (0,1)$. To this end, a *weak converse* is defined to be a proof that $\lim_{n\to\infty} -\frac{1}{n}\log\beta_n \leq E$ as long as $\alpha_n \leq \cup\epsilon$ for n large enough, and the union is over the range $\epsilon \in (0,1)$ (i.e., $\alpha_n \to 0$ with n). A *strong converse* is a proof that $\lim_{n\to\infty} -\frac{1}{n}\log\beta_n \leq E$, for any constraint $\alpha_n \leq \epsilon$, with $\epsilon \in (0,1)$.

Note that the significance of a strong converse here is different from the usual information theoretic case. By proving a strong converse to a HT problem, the conclusion is that if a strategy is found, such that the error exponent of β_n is E and $\alpha_n \to 0$, there is *nothing to gain*, in terms of error exponent, by looking for another strategy such that $\alpha_n \leq \epsilon \not\to 0$. In other words, for any case where a strong converse has been proven, the existence of a strategy such that $\alpha_n \leq \epsilon$ and the error exponent of β_n is E implies that there also exist a strategy with the same error exponent and $\alpha_n \to 0$.

Remark 1. *Throughout this thesis, we occasionally abuse notation by stating that $\epsilon \to 0$. The meaning of this is of course that $\alpha_n \to 0$, and thus it can be said that $\alpha_n \leq \epsilon$ for any ϵ and n large enough.*

As was mentioned above, Stein's Lemma is a strong property. We show a proof for this, originating in tutorials of the subject [3], in Appendix A.1. [8] proves the strong property of a multi-letter expression, for distributed HT problems with a unidirectional link and general hypotheses. Note that a single-letter converse (of any kind) still alludes us for this case. [8] shows us that this single-letter expression, if ever found, cannot be dependent on ϵ. [48] shows a similar property (as well as a single-letter expression) for the problem of successive refinement of testing against independence over a unidirectional link.

1.5 Thesis Outline and Contributions

In this thesis, we focus mainly on two problems in distributed HT under communication constraints. The first problem, treated in Chapter 3, consists of a unidirectional communication link. Throughout most of the chapter, the statistician, assumed to be located at node B in this scenario, is required to both detect (i.e., declare the true hypothesis in effect) and estimate the realizations seen by node A. The second problem, which is the focus of Chapter 4, assumed a bidirectional communication link between the nodes. In the most general case, it is assumed that the communication resources can be divided freely between the two participants, in any way that would benefit performance.

Before diving into the scenarios of interest, we dedicate Chapter 2 to definitions and tools that would prove necessary in subsequent chapters. First, notation conventions are discussed in Section 2.1, that would be used throughout the thesis. Section 2.2 gives a general model for the system of interest. This model is then adjusted in each subsequent chapter to fit the specific scenario of reference. Finally, Section 2.3 details important tools

that are in use in this work, including important results on the method of types, typicality and in hypothesis testing.

Chapter 3 (unidirectional communication link) is divided into three main parts. In the first part (Section 3.2), we focus on the case of testing against independence where the alternative hypothesis H_1 is a disjoint "version" of H_0 that leads to $\mathbf{X}^n$ and $\mathbf{Y}^n$ being independent from each other while sharing the same marginal distributions as under H_0. By relying on the techniques introduced in [8], we offer an achievable (single-letter) expression for the trade-off between the encoding rate, the error exponent and the average per-letter distortion, referred to as *rate-error-distortion region*. In this setting, we simply assume that reconstruction is only attempted when H_0 is decided, since no effective side-information is available at the decoder when H_1 is the true hypothesis.

Interestingly, it is shown that the optimal rate-error-distortion region is attained by using *successive refinement* coding where the first layer performs HT, and the second layer uses well-known results for source coding with side information at the decoder [43], while ignoring the information received by node B at the HT stage. This result is quite surprising, as in general there is no reason to believe that such a separation between the two aspects of the problem should be optimal. Indeed, this approach leads to significant losses in subsequent sections of this chapter, when general hypotheses are considered. We explicitly evaluate the rate-error-distortion region for uniform Binary Sources where a Binary Symmetric Channel (BSC) is assumed between X and Y, and plot the resulting trade-offs between the three quantities of interest.

In the second part (Section 3.3), we derive an achievable rate-error-distortion region for the same system, under no specific assumptions on the two hypotheses. To this end, we allow the use of binning not only for source reconstruction but also for testing purposes. The resulting rate-error-distortion achievable region is in fact a quadruplet in this case, comprised of the rate of communication, the error exponent for an error of the second type, subject to a maximum probability of error of first type, and the average distortion corresponding to each hypothesis. The techniques required for this analysis are inspired by previous work on distributed HT [9] and recent work [86] on the study of the error exponent for the problem of lossy source coding with side information at the receiver. It should be mentioned here that although the use of binning for HT was first suggested in [37] as a possible approach to improve performance, the benefits of this were never demonstrated.

In the third part of Chapter 3 (Section 3.4), we concentrate on distributed HT without reconstruction constraints. We show that for the case of two general hypotheses, unlike the case of testing against independence, our previous two-stage coding approach leads to significant loss in performance. We do so by suggesting a new approach for testing without requiring the decoding of the involved descriptions. This approach turns out to be superior to the previous one in terms of error exponent, but prevents the decoder from providing a lossy reconstruction of the source. Thus, the separability principle, discussed above for the case of testing against independence, is no longer true in the general case. We use the example of the BSC again (where under both hypotheses the sources are

assumed to be correlated, this time) in order to compare both proposed approaches with each other, as well as with the performance attained by previous art.

Chapter 4 (bidirectional communication link) is also divided into several parts. In the first part (Section 4.3) the main result of this chapter is presented. Considering a distributed system with a bidirectional link of *sum-rate* constrain R, we give an achievable error-exponent for the error of type II, under a fixed constraint over the error of type I, when only one round of communication is allowed (i.e., each node is allowed to send one message before a decision must be taken). We use methods inspired by [9] in order to prove the achievability of this error exponent. Section 4.4 extends the result of Section 4.3 to include any *finite number* of communication rounds between the nodes. Note that both of the error exponents of Section 4.3 and Section 4.4 are given through a minimization over some set, which is called $\mathscr{L}$. Interestingly, it turns out that the choice to allow multiple communication rounds results in more degrees of freedom (by giving the statisticians a choice regarding the right way to distribute resources through the set $\mathscr{L}$) while not changing the expression being minimized. This choice is expressed through the apparition of "new" RVs, that adhere to "new" Markov chains.

After establishing an achievable error exponent for interactive communication and general hypotheses in the previous sections, Section 4.5 revisits the special case of testing against independence, studied in [20,51]. First, it is shown that the known achievable error exponent in this case can also be achieved through our proposed result of Section 4.3, when testing against independence is assumed. This is not a trivial result, as the approach of Section 4.3 neglects to count some possible "successes", which are counted in the approach of [20]. Then, a *weak converse* is proven for the case of testing against independence over one communication round, in order to establish optimality, at least in a weak sense, of the error exponent in this case. We try to explain why, despite being achievable, an extension of this error exponent to multiple communication rounds is probably not optimal.

Section 4.6 focuses on the case of distributed HT with a bidirectional link and *zero-rate communication*. Note that this does not mean that no communication is allowed between the participants, but only that the size of the codebook cannot grow exponentially with the number of observed realizations n. [9] gave an optimal error exponent in this case, when only *one bit communication* is allowed, from node A to node B. We show, through an approach similar to the one proposed in [40], that this result is in-fact optimal whenever zero-rate communication is assumed, even when it is not limited to one bit, and allowed to also be bidirectional. We do so by proving a *strong converse* statement for this case.

Finally, concluding remarks are given in Chapter 5, where directions for possible future work are also outlined. Throughout this manuscript, proofs are generally relegated to the appendices.

1.6 Publications

During the work towards this thesis, the following papers were published or submitted to publication:

Journal Articles

[87] G. Katz, P. Piantanida and M. Debbah, "Distributed Binary Detection with Lossy Data Compression", *(submitted to) Information Theory, IEEE Transactions on, 2016.* (Available online: http://arxiv.org/abs/1601.01152)

[88] G. Katz, P. Piantanida and M. Debbah, "Collaborative Distributed Hypothesis Testing", *(submitted to) The Annals of Applied Probability.* (Available online: http://arxiv.org/abs/160

Conference Papers

[89] G. Katz, P. Piantanida, R. Couillet and M. Debbah, "Joint Estimation and Detection against Independence", *Communication, Control and Computing (Allerton), 52nd Annual Allerton Conference on, 2014.*

[90] G. Katz, P. Piantanida, R. Couillet and M. Debbah, "On the Necessity of Binning for the Distributed Hypothesis Testing Problem", *Inf. Theory, 2015 IEEE International Symposium on, (ISIT), Hong Kong.*

[91] G. Katz, P. Piantanida and M. Debbah, "Collaborative Distributed Hypothesis Testing, with General Hypotheses", *Inf. Theory, 2016 IEEE International Symposium on, (ISIT), Barcelona.*

[92] G. Katz, P. Piantanida and M. Debbah, "A New Approach to Distributed Hypothesis Testing, *(to be presented in) 50th Asilomar Conf. on Signals, Systems and Computers.*

Chapter 2

Definitions and Tools

2.1 Notation

We use upper-case letters to denote random variables (RVs) and lower-case letters to denote realizations of RVs. Vectors are denoted by boldface letters, with their length as a superscript, emitted when it is clear from the context. Let $\mathbf{X}_i^j$ denote the vector $\mathbf{X}$, from position i to position j, i.e., $\mathbf{X}_i^j = (X_i, X_{i+1}, \ldots, X_{j-1}, X_j)$. Sets, including alphabets of RVs, are denoted by calligraphic letters. Throughout this work we assume all RVs have an alphabet of finite cardinality. $P_X \in \mathcal{P}(\mathcal{X})$ denotes a probability measure (PM) for the RV $X \in \mathcal{P}(\mathcal{X})$ defined on the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, that belongs to the set of all possible PMs over $\mathcal{X}$; $X - Y - Z$ denotes that X, Y and Z form a Markov chain. We shall use tools from information theory. Notations generally comply with the ones introduced in [21]. Thus, for a RV X, distributed by $X \sim P_X(x)$, the *entropy* is defined to be $H(X) = H(P) \coloneqq -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x)$. Similarly, the *conditional entropy*:

$$H(Y|X) = H(V|P) \coloneqq -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{X}} P_X(x)V(y|x) \log V(y|x)$$

for a stochastic mapping $V : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$. The *conditional Kullback-Leiber (KL) divergence* between two stochastic mappings $P_{Y|X} : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$ and $Q_{Y|X} : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$, is:

$$\mathcal{D}(P_{Y|X}\|Q_{Y|X}|P_X) \coloneqq \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_X(x)P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \ , \tag{2.1}$$

satisfying that $P_{Y|X} \ll Q_{Y|X}$ *a.e.* with relation to P_X (i.e., for every $x \in \mathcal{X}$ such that $P_X(x) > 0$, $P_{Y|X}(y|x) > 0$ implies that $Q_{Y|X}(y|x) > 0$). For any two RVs, X and Y, whose measure is controlled by $XY \sim P_{XY}(x, y) = P_X(x)P_{Y|X}(y|x)$, the following is defined to be the *mutual information* between them: $I(X;Y) = I(P_X; P_{Y|X}) \coloneqq \mathcal{D}(P_{XY}\|P_X P_Y)$. Given a vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $N(a|\mathbf{x})$ be the *counting measure*, i.e., the number of times the letter $a \in \mathcal{X}$ appears in the vector X. The *type* of the vector $\mathbf{x}$, denoted by $Q_\mathbf{x}$, is defined through its *empirical measure*: $Q_\mathbf{x}(a) = n^{-1}N(a|\mathbf{x})$ with $a \in \mathcal{X}$. $\mathcal{P}_n(\mathcal{X})$ denotes

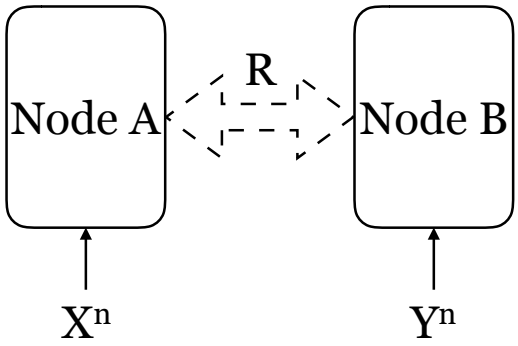


Figure 2.1: General distributed hypothesis testing model with two nodes.

the set of all possible types (or empirical measures) of length n over $\mathcal{X}$. We use type variables of the form $X^{(n)} \in \mathcal{P}_n(\mathcal{X})$ to denote a RV with a probability measure identical to the empirical measure induced by $\mathbf{x}$. The set of all vectors $\mathbf{x}$ that share this type is denoted by $\mathcal{T}(Q_{\mathbf{x}}) = \mathcal{T}_{[Q_{\mathbf{x}}]}$. Main definitions of *$\delta$-typical sets* and some of their properties are given in the following short tutorial. We denote the scalar convolution function by $a \star b \triangleq a(1-b) + b(1-a)$. All exponents and logarithms are assumed to be of base 2.

2.2 System Model and Definitions

In a system comprising two statisticians, as depicted in Fig. 2.1, each of them is assumed to observe the i.i.d. realizations of one RV. Let $\mathbf{X}^n\mathbf{Y}^n = (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random variables in $(\mathcal{X}^n \times \mathcal{Y}^n, \mathcal{B}_{\mathcal{X}^n \times \mathcal{Y}^n})$ that are jointly distributed in one of two ways, denoted by hypotheses H_0 and H_1, with probability measures as follows:

$$\begin{cases} H_0: & P_{0,XY}(x,y) \triangleq P_{XY}(x,y) \ , \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y} \ , \\ H_1: & P_{1,XY}(x,y) \triangleq P_{\bar{X}\bar{Y}}(x,y) \ , \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y} \ . \end{cases} \tag{2.2}$$

It is assumed that the two nodes of the system are connected through a perfect link, constrained by a *sum-rate constraint* $R \left[\frac{\text{bits}}{\text{symbol} \times \text{node}}\right]$. That is, if each node sees n realizations of its respective RV, 2^{nR} bits are allowed to pass on the link. No errors are introduced by the link to the transmitted information, as long as the rate constraint is respected.

Throughout this thesis, several problems in distributed hypothesis testing will be faced. Each of these problems will be defined by the *task* the statisticians are required to complete, as well as the *nature of communication*. In all cases, the statisticians will be required to declare the correct probability distribution controlling the observed RVs, out of the two possible options. We define two error events, with their respective probabilities, in accordance to the literature on the subject:

$$\begin{cases} \alpha_n \triangleq \Pr(H_1 \text{ is declared } | XY \sim P_{0,XY}) = P_{XY}(\mathcal{A}^c) \ , \\ \beta_n \triangleq \Pr(H_0 \text{ is declared } | XY \sim P_{1,XY}) = P_{\bar{X}\bar{Y}}(\mathcal{A}) \ . \end{cases} \tag{2.3}$$

Here, $\mathcal{A}$ is assumed to be the *acceptance region*, comprising pairs of vectors of length n, which prompts the statisticians to declare H_0. α_n and β_n are the probabilities of error of types I and II, respectively, when the statisticians see n realizations of their respective RVs.

2.3 Tools

The rest of this chapter is consecrated to a short tutorial on the different tools that will be in use throughout this thesis. We start with an overview of the method of types (see e.g., [21] for a more complete tutorial on the subject) and its relation with the notion of *typicality*, which plays a major role in the field of Information Theory in general, as well as in this thesis. Next, some important results in the field of Hypothesis Testing (both centralized and distributed) are discussed.

2.3.1 Types and Typicality

Definition 5 (Types [82])**.** *The type of a sequence $\mathbf{x} \in \mathcal{X}^n$ is the measure $\hat{P}_X$ on $\mathcal{X}$ defined by*

$$\hat{P}_X(a) \coloneqq \frac{1}{n} N(a|\mathbf{x}) \ , \quad \forall a \in \mathcal{X} \ , \tag{2.4}$$

where $N(a|\mathbf{x})$ is the counting measure of the letter a in $\mathbf{x}$. The joint type *of a pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ is the empirical measure $\hat{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ such that*

$$\hat{P}_{XY}(a,b) \coloneqq \frac{1}{n} N(a,b|\mathbf{x},\mathbf{y}) \ , \quad \forall (a,b) \in \mathcal{X} \times \mathcal{Y} \ , \tag{2.5}$$

where $N(a,b|\mathbf{x},\mathbf{y})$ is the joint counting measure of the pair (a,b) in $(\mathbf{x},\mathbf{y})$.

Definition 6 (Conditional Types [82])**.** *The vector $\mathbf{y} \in \mathcal{Y}^n$ is said to have* conditional type $V : \mathcal{X} \mapsto \mathcal{P}_n(\mathcal{Y})$ *given $\mathbf{x} \in \mathcal{X}^n$ if*

$$N(a,b|\mathbf{x},\mathbf{y}) = N(a|\mathbf{x}) V(b|a) \ , \quad \forall (a,b) \in \mathcal{X} \times \mathcal{Y} \ , \tag{2.6}$$

where V is a stochastic mapping.

The definition of types (and typicality, to be discussed later) is crucial to the field of Information Theory. When drawing n times independently from a probability distribution P_X, note that the probability of obtaining any sequence $\mathbf{x} \in \mathcal{X}^n$ is dependent only on its type. Specifically, this probability can be expressed as follows:

$$P_X^n(\mathbf{x}) = \prod_{i=1}^{n} P_X(x_i) = \prod_{a \in \mathcal{X}} P_X(a)^{N(a|\mathbf{x})} \ . \tag{2.7}$$

Here, x_i denotes the i-th component of the vector $\mathbf{x}$. As mentioned above, the set of all vectors in $\mathcal{X}$ that have the same type as $\mathbf{x}$ is denoted by $\mathcal{T}_{[Q_\mathbf{x}]}$. As the type of $\mathbf{x}$, $Q_\mathbf{x}$

fulfills all the requirement to being a probability distribution itself, we may treat it as such. For a given vector $\mathbf{x}$, we denote the corresponding *type variable*, which is a RV with probability distribution $Q_{\mathbf{x}}$, by $X^{(n)}$.

The following important properties of types will be used throughout this thesis:

Lemma 4 (Type Counting)**.** *Let $\mathcal{P}_n(\mathcal{X})$ be the set of all possible types of sequences in $\mathcal{X}^n$. Then,*

$$|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|} \ . \tag{2.8}$$

Proof. For every $a \in \mathcal{X}$, $N(a|\mathbf{x})$ can take up to $(n+1)$ different values (see reference [82, Lemma 2.2]). □

Lemma 5. *For any type $\hat{P} \in \mathcal{P}_n(\mathcal{X})$ of sequences in $\mathcal{X}^n$, denote by $\mathcal{T}_{[\hat{P}]}$ the set of all sequences with this type. Then,*

$$(n+1)^{-|\mathcal{X}|} \exp\left[nH(\hat{P})\right] \leq |\mathcal{T}_{[\hat{P}]}| \leq \exp\left[nH(\hat{P})\right] \ . \tag{2.9}$$

In a similar fashion, for every $\mathbf{x} \in \mathcal{X}^n$ and stochastic mapping $V : \mathcal{X} \mapsto \mathcal{P}_n(\mathcal{Y})$, let $\mathcal{T}_{[V]}(\mathbf{x})$ be the set of all sequences $\mathbf{y} \in \mathcal{Y}^n$ with the conditional type V given $\mathbf{x}$. Then,

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|} \exp\left[nH(V|\hat{P})\right] \leq |\mathcal{T}_{[V]}(\mathbf{x})| \leq \exp\left[nH(V|\hat{P})\right] \ , \tag{2.10}$$

where $H(V|\hat{P})$ is the conditional entropy function,

$$H(V|\hat{P}) = \sum_{x \in \mathcal{X}} \hat{P}(x) H(V(\cdot|x)) \ . \tag{2.11}$$

Proof. Refer to reference [82, Lemma 2.3, Lemma 2.5]. □

Lemma 6 (Inaccuracy)**.** *Let $\hat{P} \in \mathcal{P}_n(\mathcal{X})$ be the type of $\mathbf{x} \in \mathcal{X}^n$ ($X^{(n)} \sim \hat{P}$ is referred to as the* type variable*). Then, for any RV X on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_X)$,*

$$\begin{aligned} &P_X^n(\mathbf{X}^n = \mathbf{x}) = \exp\Big\{-n\Big[H(\hat{P}) + \mathcal{D}(\hat{P}\|P_X)\Big]\Big\} \ , \\ &(n+1)^{-|\mathcal{X}|} \exp\Big\{-n\mathcal{D}(\hat{P}\|P_X)\Big\} \leq P_X^n(\mathcal{T}_{[\hat{P}]}) \leq \exp\Big\{-n\mathcal{D}(\hat{P}\|P_X)\Big\} \ . \end{aligned} \tag{2.12}$$

Similarly, for every $\mathbf{x} \in \mathcal{X}^n$ and stochastic mappings $V : \mathcal{X} \mapsto \mathcal{P}_n(\mathcal{Y})$, $W : \mathcal{X} \mapsto \mathcal{P}_n(\mathcal{Y})$ such that $\mathcal{T}_{[V]}(\mathbf{x})$ is non-void,

$$W^n(\mathbf{y}|\mathbf{x}) = \exp\Big\{-n\Big[H(V|\hat{P}) + \mathcal{D}(V\|W|\hat{P})\Big]\Big\} \tag{2.13}$$

if $\mathbf{y} \in \mathcal{T}_{[V]}(\mathbf{x})$, and

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|} \exp\Big\{-n\mathcal{D}(V\|W|\hat{P})\Big\} \leq W_X^n(\mathcal{T}_{[V]}(\mathbf{x})|\mathbf{x}) \leq \exp\Big\{-n\mathcal{D}(V\|W|\hat{P})\Big\} \tag{2.14}$$

Proof. Refer to reference [9, Lemma 3],[82, Lemma 2.6]. □

Note the following important conclusion of the last three lemmas: For a RV X, drawn n times independently out of the same probability distribution $P_X(x)$, the *number of possible types* grows sub-exponentially with n, while the "size" of each type (i.e., the number of vectors in $\mathcal{X}^n$ that have the same type) grows exponentially with n. The probability of seeing a particular type after n draws diminishes exponentially with n, unless the type in question is identical to the probability distribution P_X. These facts will come in handy in proofs throughout this thesis.

Definition 7 (δ-Typicality [9])**.** *Let $\delta > 0$. An n-sequence $\mathbf{x}$ is called δ-typical, if $|\frac{N(a|\mathbf{x})}{n} - P_X(a)| \leq \delta$, $\forall a \in \mathcal{X}$, and $\hat{P}_X \ll P_X$. The set of all δ-typical sequences $\mathbf{x}$ is denoted by $\mathcal{T}_{[P_X]_\delta} = \mathcal{T}_{[X]_\delta}$. The set of jointly δ-typical sequences $\mathcal{T}_{[XY]_\delta}$ is defined in a similar manner.*

Remark 2. *Note that the δ-typical set can be expressed as a union of types in the following manner:*

$$\mathcal{T}^n_{[X]_\delta} = \bigcup_{\substack{|\hat{P}(a) - P_X(a)| \leq \delta \ ,\forall a \\ \hat{P} \ll P_X}} \mathcal{T}^n_{[\hat{P}]} \tag{2.15}$$

Definition 8 (Conditional Typicality [9, 82])**.** *Let $\delta > 0$ and X, Y be two RVs, jointly distributed according to $P_{XY}(x, y)$. An n-sequence $\mathbf{y}$ is called conditionally δ-typical, with relation to a vector $\mathbf{x}$, if $|\frac{N(a,b|\mathbf{x},\mathbf{y})}{n} - P_{Y|X}(b|a)Q_{\mathbf{x}}| \leq \delta$, $\forall (a, b) \in \mathcal{X} \times \mathcal{Y}$, and $N(a, b|\mathbf{x}, \mathbf{y}) \ll P_{Y|X}(b|a)$, for every $a \in \mathcal{X}$ such that $Q_{\mathbf{x}} > 0$. The set of all δ-typical sequences $\mathbf{y}$ with relation to the vector $\mathbf{x}$ is denoted by $\mathcal{T}_{[Y|X]_\delta}(\mathbf{x})$.*

Remark 3. *Note in particular that $\mathcal{T}_{[Y|X]_\delta}(\mathbf{x}) = \varnothing$ for any $\mathbf{x} \notin \mathcal{T}_{[X]_\delta}$.*

The concept of typicality has been essential to information theory from the beginning. We bring forth a few properties that will prove useful throughout this thesis:

Lemma 7. *Let $\mathcal{T}_{[X]_\delta}$, $\mathcal{T}_{[XY]_\delta}$ and $\mathcal{T}_{[Y|X]_\delta}$ denote the sets of typical, jointly typical and conditionally typical sequences, respectively. For any $\mathbf{x} \in \mathcal{T}_{[X]_\delta}$ and $\mathbf{y} \in \mathcal{T}_{[Y|X]_{\delta'}}$, then $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{[XY]_{\delta+\delta'}}$. Moreover, $\mathbf{y} \in \mathcal{T}_{[Y]_{\delta''}}$, with $\delta'' \coloneqq (\delta + \delta')|\mathcal{X}|$.*

Proof. Refer to reference [82]. □

Remark 4. *Note that the opposite direction is simpler. By definition, if $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}^n_{[XY]_\delta}$, then $\mathbf{x} \in \mathcal{T}^n_{[X]_\delta}$, $\mathbf{y} \in \mathcal{T}^n_{[Y]_\delta}$, $\mathbf{x} \in \mathcal{T}^n_{[Y|X]_\delta}(\mathbf{x})$ and $\mathbf{y} \in \mathcal{T}^n_{[X|Y]_\delta}(\mathbf{y})$.*

Lemma 8 (Generalized Markov Lemma)**.** *Let $p_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ be a probability measure that satisfies: $U - X - Y$. Consider $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}^n_{[XY]_{\epsilon'}}$ and random vectors $\mathbf{U}^n$ generated according to:*

$$\Pr\left\{\mathbf{U}^n = \mathbf{u} \middle| U^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(\mathbf{x}), \mathbf{x}, \mathbf{y}\right\} = \frac{\mathbb{1}\left\{\mathbf{u}^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(\mathbf{x})\right\}}{\left|\mathcal{T}^n_{[U|X]_{\epsilon''}}(\mathbf{x})\right|} \ . \tag{2.16}$$

For sufficiently small $\epsilon, \epsilon', \epsilon'' > 0$,

$$\Pr\left\{\mathbf{U}^n \notin \mathcal{T}^n_{[U|XY]_\epsilon}(\mathbf{x}, \mathbf{y}) \middle| \mathbf{U}^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(\mathbf{x}), \mathbf{x}, \mathbf{y}\right\} \equiv \mathcal{O}\left(c^{-n}\right) \tag{2.17}$$

holds uniformly on $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}^n_{[XY]_{\epsilon'}}$ where $c > 1$.

Proof. Refer to reference [93]. □

Lemma 9. *For every probability measure $P_X \in \mathcal{P}(\mathcal{X})$ and stochastic mapping $W : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$, there exist sequences $(\varepsilon_n)_{n\in\mathbb{N}_+}, (\varepsilon'_n)_{n\in\mathbb{N}_+} \to 0$ as $n \to \infty$ satisfying:*

$$\left|\frac{1}{n}\log|\mathcal{T}_{[X]_\delta}| - H(X)\right| \leq \varepsilon_n \ , \ \left|\frac{1}{n}\log|\mathcal{T}_{[Y|X]_\delta}(\mathbf{x})| - H(Y|X)\right| \leq \varepsilon_n \ , \tag{2.18}$$

for each $\mathbf{x} \in \mathcal{T}_{[X]_\delta}$ where $\varepsilon_n \equiv \mathcal{O}(n^{-1}\log n)$, and

$$P_X^n\big(\mathcal{T}_{[X]_\delta}\big) \geq 1 - \varepsilon'_n \ , \ W^n\big(\mathcal{T}_{[Y|X]_\delta}(\mathbf{x})|X^n = \mathbf{x}\big) \geq 1 - \varepsilon'_n \ , \tag{2.19}$$

for all $\mathbf{x} \in \mathcal{X}^n$ where $\varepsilon'_n \equiv \mathcal{O}\left(\frac{1}{n\delta^2}\right)$, provided that n is sufficiently large.

Proof. Refer to reference [82, Lemma 2.13]. □

Thus, while the *size* of the δ-typical set gets (in the single variable case, for example) arbitrarily close to $\exp\{nH(X)\}$, which may be much smaller than the amount of all possible sequences of length n, $|\mathcal{X}|^n$, the *probability* of the δ-typical set gets arbitrarily close to 1 as n grows. This observation is essential to many results in information theory in general, and in this thesis specifically.

Finally, the following lemma will prove useful in Chapter 3:

Lemma 10 (Set of sequences with small empirical entropy [86])**.** *For any pair of strings of length n, denoted by $(\mathbf{x}^n, \mathbf{y}^n)$, let*

$$\mathcal{S}(\mathbf{x},\mathbf{y}) = \Big\{(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in \mathcal{X}^n \times \mathcal{Y}^n \,\Big|\, H(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \leq H(\mathbf{x},\mathbf{y})\Big\} \ , \tag{2.20}$$

with $H(\mathbf{x},\mathbf{y})$ being the empirical entropy of the sequences,

$$H(\mathbf{x},\mathbf{y}) = -\sum_{a\in\mathcal{X},b\in\mathcal{Y}} Q_{\mathbf{xy}}(a,b)\log Q_{\mathbf{xy}}(a,b) \ . \tag{2.21}$$

Then

$$|\mathcal{S}(\mathbf{x},\mathbf{y})| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\big[nH(\mathbf{x},\mathbf{y})\big] \ . \tag{2.22}$$

In addition, let

$$\mathcal{S}(\mathbf{x}|\mathbf{y}) = \Big\{\tilde{\mathbf{x}} \in \mathcal{X}^n \,|\, H(\tilde{\mathbf{x}}|\mathbf{y}) \leq H(\mathbf{x}|\mathbf{y})\Big\} \ , \tag{2.23}$$

then

$$|\mathcal{S}(\mathbf{x}|\mathbf{y}))| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\big[H(\mathbf{x}|\mathbf{y})\big] \ . \tag{2.24}$$

Proof. Refer to reference [86]. □

2.3.2 Results in Hypothesis Testing

We now present some important known results in hypothesis testing, both in the context of a centralized scenario as well as a distributed scenario. We start with a fundamental result in centralized hypothesis testing, mentioned in the introduction:

Lemma 11 (Stein's Lemma)**.** *Let $X_1, X_2, \ldots, X_n$ be independently drawn from $P \in \mathcal{P}(\mathcal{X})$. Consider the hypothesis test*

$$\begin{cases} H_0: & P = P_0 \ , \\ H_1: & P = P_1 \ . \end{cases} \tag{2.25}$$

Assume that $\mathcal{D}(P_0||P_1) < \infty$, and let $\mathcal{A}_n$ be an acceptance region for hypothesis H_0. Let the probabilities of error be as defined above, $\alpha_n(\mathcal{A}_n) = P_0^n(\mathcal{A}_n^c)$ and $\beta_n(\mathcal{A}_n) = P_1^n(\mathcal{A}_n)$. For $\epsilon \in (0,1)$ define

$$\beta_n^\star(\epsilon) = \min_{\mathcal{A}_n \subseteq \mathcal{X}^n} \{\beta_n(\mathcal{A}_n) | \alpha_n \leq \epsilon\} \ . \tag{2.26}$$

For every $\epsilon \in (0,1)$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n^\star(\epsilon) = \mathcal{D}(P_0||P_1) \ . \tag{2.27}$$

Proof. Many proofs exist in literature for Stein's Lemma (also referred to in some places as the Chernoff-Stein Lemma). The problem is that in many cases these proofs are not strong. While [2] mentions that Stein's Lemma is a *strong* property, the proof only shows that

$$\mathcal{D}(P_0||P_1) - \epsilon \leq \lim_{n\to\infty} -\frac{1}{n} \log \beta_n \leq \mathcal{D}(P_0||P_1) + \epsilon \ , \tag{2.28}$$

with $\epsilon > 0$ being the constraint put over the error probability of Type I. Clearly, these bounds are dependent on ϵ, and are only tight if ϵ is arbitrarily small. We choose to present a different proof, common in tutorials on the subject and taken specifically from [3], that demonstrates the strong property of Stein's Lemma. This proof can be found in Appendix A.1. □

Consider now a two-node distributed system with two RVs, as depicted in Figure 2.1. A similar model, in which only unidirectional communication is allowed, from node A to node B (see Figure 3.1), was first presented and analyzed in [8]. For the general case, a *strong* property was demonstrated for a *multi-letter expression*, as summarized in the following lemma:

Lemma 12. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be pairs independently drawn from $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Consider the hypothesis test*

$$\begin{cases} H_0: & P = P_{XY} \ , \\ H_1: & P = P_{\bar{X}\bar{Y}} \ , \end{cases} \tag{2.29}$$

over a distributed system with a unidirectional communication link of rate R, as depicted in Figure 3.1. Let

$$\theta_n(R) = \sup_f \Big\{ \frac{1}{n} \mathcal{D}(P_{f(\mathbf{X})\mathbf{Y}} || P_{f(\bar{\mathbf{X}})\bar{\mathbf{Y}}}) : \log |f| \leq nR \Big\} \ , \tag{2.30}$$

where $|f|$ is the number of different values the function $f(\cdot)$ can present, and let

$$\theta(R) = \sup_n \theta_n(R) \ . \tag{2.31}$$

Defining the error events and their probabilities as above, the following is true:

$$\lim_{n\to\infty} \frac{1}{n} \beta_n^\star(R, \epsilon) = -\theta(R) \ , \tag{2.32}$$

where $\beta_n^\star(R, \epsilon)$ denotes the optimal error exponent of type II, under constraint $\epsilon \in (0, 1)$ over the error probability of type I and rate-constraint R.

Proof. Refer to reference [8]. □

Clearly, Lemma 12 constitutes a strong quality for the error exponent of Type II, as it is not dependent on the constraint ϵ. Note that this result leads to a simple but important conclusion: When unidirectional communication is considered for binary distributed HT problems, the problem boils down to the choice of a good encoding strategy $f(\cdot)$. Given this choice, the optimal approach is to apply Stein's Lemma to the entirety of the information available at node B, namely $(f(\mathbf{x}), \mathbf{y})$.

The multi-letter expression of Lemma 12 was calculated explicitly in a single-letter form in the same paper, for the case of testing against independence, as summarized in the following lemma.

Lemma 13. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be pairs independently drawn from $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Consider the hypothesis test*

$$\begin{cases} H_0 : & P = P_{XY} \ , \\ H_1 : & P = P_X P_Y \ , \end{cases} \tag{2.33}$$

over a distributed system with a unidirectional communication link of rate R, as depicted in Figure 3.1. For every $R \geq 0$,

$$\theta(R) = \max_U \Big\{ I(U;Y) : U - X - Y, I(U;X) \leq R, |\mathcal{U}| \leq |\mathcal{X}| + 1 \Big\} \ . \tag{2.34}$$

Proof. The proof, based on the Ahlswede-Körner solution to the problem of source coding with side information [78, 94], can be found in [8]. □

In [9], a new achievable error exponent was proposed for the same distributed system with a unidirectional link, for the case of general hypotheses, through the method of types [21]:

Lemma 14. *Under the assumptions of Lemma 12, define the two following sets:*

$$\begin{aligned}\mathscr{S}(R) &= \{U : I(U;X) \leq R \ , U - X - Y\} \ , \\ \mathscr{L}(U) &= \{\tilde{U}\tilde{X}\tilde{Y} : P_{\tilde{U}\tilde{X}}(u,x) = P_{UX}(u,x), P_{\tilde{U}\tilde{Y}}(u,y) = P_{UY}, \forall (u,x,y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y}\}\end{aligned} \tag{2.35}$$

and define the RV $\bar{U}$ so as to satisfy conditions $\bar{U} - \bar{X} - \bar{Y}$ and $P_{\bar{U}|\bar{X}} = P_{U|X}$, where the range $\mathcal{U}$ of U is over all finite sets. $\bar{U}$ is uniquely determined by these conditions when a $U \in \mathscr{S}(R)$ is given. Furthermore, define the non-decreasing function

$$\theta_L(R) = \sup_{U \in \mathscr{S}(R)} \inf_{\tilde{U}\tilde{X}\tilde{Y} \in \mathscr{L}(U)} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} || P_{\bar{U}\bar{X}\bar{Y}}) \ . \tag{2.36}$$

Let $\theta(R)$ be the error exponent of type II as defined in Lemma 12, then

$$\theta(R) \geq \theta_L(R) \ . \tag{2.37}$$

Proof. Refer to reference [9]. □

While the methods used in [9] differ from the ones in [8], it is straight-forward to see that the result of Lemma 13 for the case of testing against independence can be retrieved from the general achievable result of Lemma 14 as follows:

$$\begin{aligned}\theta_L(R) &= \sup_{U \in \mathscr{S}(R)} \inf_{\tilde{U}\tilde{X}\tilde{Y} \in \mathscr{L}(U)} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} || P_{\bar{U}\bar{X}\bar{Y}}) && \text{(2.38a)} \\ &\geq \sup_{U \in \mathscr{S}(R)} \inf_{\tilde{U}\tilde{X}\tilde{Y} \in \mathscr{L}(U)} \mathcal{D}(P_{\tilde{U}\tilde{Y}} || P_{\bar{U}\bar{Y}}) && \text{(2.38b)} \\ &= \sup_{U \in \mathscr{S}(R)} \mathcal{D}(P_{UY} || P_{\bar{U}\bar{Y}}) && \text{(2.38c)} \\ &= \sup_{U \in \mathscr{S}(R)} \mathcal{D}(P_{UY} || P_U P_Y) = \sup_{U \in \mathscr{S}(R)} I(U;Y) && \text{(2.38d)}\end{aligned}$$

Here, (2.38b) is due to the chain rule for KL divergence, (2.38c) is due to the definition of the set $\mathscr{L}(U)$, and (2.38d) stems from the assumption of testing against independence. The fact that the *optimal* result for testing against independence can be achieved through the approach of Lemma 14 gives hope that this approach may lead to good performance in general, even if not necessarily optimal in the general case. [37] briefly proposes *random binning* as an approach to improve on the result of Lemma 14. This approach was never thoroughly investigated, to the best of our knowledge. We do so in Chapter 3.

Very few works are available in the literature for the case of bidirectional (interactive) communication. One such work is [20], where testing against independence is considered, with one round of communication. In this scenario, it is assumed that node A sends a message, which is then answered with a message from node B. At the end of this process a decision must be taken. The achievability result proposed in [20] is brought forth in the following lemma:

Lemma 15. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be pairs independently drawn from $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Consider the hypothesis test*

$$\begin{cases} H_0: & P = P_{XY} \ , \\ H_1: & P = P_X P_Y \ , \end{cases} \tag{2.39}$$

over a distributed system with one-round communication. Assume that the communication from nodes A and B is restricted by rate-constraints R_A and R_B, respectively. Let

$$\theta_2(R_A, R_B, \epsilon) = \lim_{n\to\infty} -\frac{1}{n} \log \beta_n^\star(R_A, R_B, \epsilon) \tag{2.40}$$

be the optimal error-exponent of Type II under the rate-constraints as defined above and the constraint $\alpha_n \leq \epsilon$ for the error of Type I. Then

$$\theta_2(R_A, R_B, \epsilon) \geq \max_{\substack{P_{U|X}, P_{V|UY} \\ R_A \geq I(U;X) \\ R_B \geq I(V;Y|U)}} I(U;Y) + I(V;X|U) \ . \tag{2.41}$$

Proof. Refer to reference [20]. □

Remark 5. *In the same paper, it is also claimed that the expression for $\theta_2(R_A, R_B, \epsilon)$ above constitutes a* weak converse *to the performance. Unfortunately, an error has fallen in the proof of this claim. We revisit it in Chapter 4.*

Remark 6. *In this work, whenever considering interactive communication, we will focus on constraints on the* sum-rate*, and assume that the participants are allowed to divide the rate as they see fit in order to benefit performance. It is easy to see that in such a case, the expression for θ_2 changes to:*

$$\theta_2(R, \epsilon) \geq \max_{\substack{P_{U|X}, P_{V|UY} \\ R \geq I(U;X) + I(V;Y|U)}} I(U;Y) + I(V;X|U) \ . \tag{2.42}$$

The achievable result of [20] was extended by the same authors in [51] to include the case of interactive communication over multiple rounds.

Chapter 3

Joint Detection and Estimation with Unidirectional Communication

3.1 Overview

In this chapter, based on the work published in [87, 89, 90, 92], we focus on the case of unidirectional communications, as seen in Figure 3.1. We start by considering the joint problem of testing against independence and estimation in Section 3.2. Here, node B, which can be referred to as the decoder in the case of unidirectional communication, is required to estimate the vector of realizations seen by node A, $\mathbf{x}^n$. It does so with average distortion lower than some threshold D, and only under the condition that the detection phase concluded H_0 is the true hypothesis. Note that the distortion is measured under the assumption that *the right decision has been made*, as the "penalty" for erroneous detection is already embodied in the resulting detection error exponent.

In Section 3.3 we consider a similar scenario of joint detection and estimation, for the case of general hypotheses. Here, estimation is done irrespective of the decision taken during the detection phase. However, the distortion allowed under each decision may be different. Thus, the rate-exponent-distortion region, which is represented by a triplet (R, E, D) in the case of testing against independence, is represented by a quadruplet

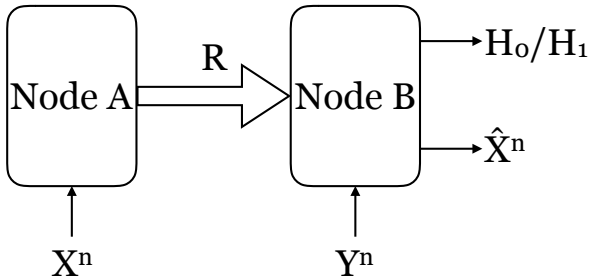


Figure 3.1: Joint detection and estimation model, with unidirectional communication.

(R, E, D_0, D_1) in the case of general hypotheses. Nevertheless, the distortion in this case is still calculated under the assumption of correct detection.

Finally, the case of general hypotheses is revisited for the case where no estimation is required. A different strategy is proposed for this case in Section 3.4. It is shown that significant gains can be achieved in some cases of testing with general hypotheses, when the requirement for source estimation is relaxed. This stands in contrast to the separation principal when testing against independence, as brought forth in Section 3.2.

3.2 Joint Detection and Estimation - Against Independence

3.2.1 System Model

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. Nodes A and B observe sequences of random variables $(X_i)_{i\in\mathbb{N}^\star}$ and $(Y_i)_{i\in\mathbb{N}^\star}$ respectively, which take values on $\mathcal{X}$ and $\mathcal{Y}$, resp. For each $i \in \mathbb{N}^\star$, random samples (x_i, y_i) are distributed according to one of two possible joint distributions:

$$\begin{cases} H_0 : & p_0(x,y) = P_{XY}(x,y) \ , \\ H_1 : & p_1(x,y) = P_{\bar{X}\bar{Y}}(x,y) = P_X(x)P_Y(y) \ . \end{cases} \tag{3.1}$$

on $\mathcal{X}\times\mathcal{Y}$, and $P_X(x) = \sum_{y\in\mathcal{Y}} P_{XY}(x,y)$ is the marginal distribution of X (and similarly for Y), according to hypothesis 0. Moreover, they are independent across time i.

Let $d : \mathcal{X}\times\hat{\mathcal{X}} \to [0\,; d_{\max}]$ be a finite distortion measure *i.e.*, such that $0 \le d_{\max} < \infty$. We also denote by d the component-wise mean distortion on $\mathcal{X}^n \times \hat{\mathcal{X}}^n$, *i.e.*, for each $(\mathbf{x}^n, \hat{\mathbf{x}}^n) \in \mathcal{X}^n \times \hat{\mathcal{X}}^n$, $d(\mathbf{x}^n, \hat{\mathbf{x}}^n) \triangleq \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$. 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 (user authentication) between the two possible hypotheses. After having decided between the two hypotheses, node B attempts to reconstruct the sequence $\mathbf{x}$, with minimum distortion, for some additive distortion measure, in case H_0 was concluded to be the correct hypothesis. While recovering the sequence seen by node A under hypothesis H_1 may still be possible, it becomes less relevant, as in this case the sequence seen by node B is completely independent and does not constitute as side information. Furthermore, it is very likely that in realistic cases where testing against independence arises, deciding H_1 implies that the information seen by node A is irrelevant to node B. Thus, for the case of testing against independence, we assume node B attempts to decode only if it has decided H_0.

Definition 9 (Code)**.** *An (n, R)-code (also referred to as "strategy" throughout this thesis) for testing against independence in this setup is defined by*

- *An encoding function at node A denoted by $f_n : \mathcal{X}^n \to \{1, \dots, \|f_n\|\}$;*

- *A decision region* $\mathcal{A}_n \subset \{1, \dots, \|f_n\|\} \times \mathcal{Y}^n$, *such that if* $(f_n(\mathbf{x}^n), \mathbf{y}^n) \in \mathcal{A}_n$ *the decoder declares* H_0 *and otherwise* H_1 *;*
- *A reconstruction function at node B denoted by* $g_n : \{1, \dots, \|f_n\|\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$.

Definition 10 (Rate-exponent-distortion region)**.** *A tuple* $(R, E, D, \epsilon) \in \mathbb{R}_+^4$ *is said to be* achievable *if, for any* $\delta > 0$ *and for* n *large enough, there exists an* $(n, R+\delta)$*-code* $(f_n, \mathcal{A}_n, g_n)$ *such that:*

$$\begin{aligned} n^{-1} \log \|f_n\| &\leq R + \delta \ , \\ \mathbb{E}_0\big[d\big(\mathbf{X}^n, g_n(f_n(\mathbf{X}^n), \mathbf{Y}_0^n)\big)\big] &\leq D + \delta \ , \\ -\frac{1}{n} \log \beta_n(\mathcal{A}_n) &\geq E - \delta \ , \\ \alpha_n(\mathcal{A}_n) &\leq \epsilon \ , \end{aligned} \tag{3.2}$$

where $\beta_n(\mathcal{A}_n) = \Pr\big(\mathcal{A}_n | XY \sim p_1(x,y)\big)$ *and* $\alpha_n(\mathcal{A}_n) = \Pr\big(\mathcal{A}_n^c | XY \sim p_0(x,y)\big)$*, and distortion is measured under the condition that node B correctly decides* H_0*. The set of all such achievable tuples is denoted by* $\mathcal{R}^\star$ *and is referred to as the* rate-exponent-distortion region*.*

In [8] and later on in [9], 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 $\mathbf{Y}^n$ and the encoded descriptions $f_n(\mathbf{X}^n)$. More specifically, by optimizing over all decision regions $\mathcal{A}_n \subset \{1, \dots, \|f_n\|\} \times \mathcal{Y}^n$, the smallest probability of error of the second type β_n asymptotically behaves as: $\beta_n \approx \exp\left(-nE(R)\right)$ with n large enough, for a fixed constraint on the error probability of the first type $\alpha_n \leq \epsilon$, and the exponent $E(R)$ satisfies [8, Lemma 1.a]:

$$E(R) = \sup_{n \geq 1} E_n(R) \ , \tag{3.3}$$

where

$$E_n(R) = \sup_{f_n} \left\{ \frac{1}{n} I\left(f_n(\mathbf{X}^n); \mathbf{Y}^n\right) \ \middle| \ \log \|f_n\| \leq nR \right\} . \tag{3.4}$$

This asymptotic equivalence implies a strong converse property that, much like in the single-node HT setup, the optimal exponential decay of β_n is not dependent upon the chosen constraint $0 < \epsilon < 1$ on the error probability of the first type α_n (e.g. see [48] for a proof based on image sets). Exploiting this equivalence the optimal rate-error-distortion region of the system depicted in Fig. 3.1 can be expressed through the following *multi-letter characterization.*

Lemma 16 (Multi-letter characterization [8])**.** *The rate-error-distortion region* $\mathcal{R}^\star$ *when testing against independence is described by the set of tuples* $(R, E, D) \in \mathbb{R}_+^3$ *satisfying:*

$$\limsup_{n \to \infty} \frac{1}{n} \log \|f_n\| \leq R \ , \tag{3.5a}$$

$$\liminf_{n \to \infty} \frac{1}{n} I\left(f_n(\mathbf{X}^n); \mathbf{Y}_0^n\right) \geq E \ , \tag{3.5b}$$

$$\limsup_{n \to \infty} \mathbb{E}_0 \left[d\big(\mathbf{X}^n, \hat{\mathbf{X}}^n = g_n(f_n(\mathbf{X}^n), \mathbf{Y}_0^n)\big) \right] \leq D \ , \tag{3.5c}$$

for some sequence of encoding and decoding mappings (f_n, g_n).

Remark 7. *Region* $\mathcal{R}^\star$ *is closed and convex.*

3.2.2 Single-Letter Rate-Error-Distortion-Region

We now state the optimal rate-error-distortion region for testing against independence, which provides a single-letter expression for that in Lemma 16:

Theorem 1 (Rate-error-distortion region)**.** *A tuple* $(R, E, D) \in \mathbb{R}_+^3$ *is achievable for the two-node detection and reconstruction problem when testing against independence, as defined in Definition 10, if and only if two random variables* $U \in \mathcal{U}$ *and* $V \in \mathcal{V}$*, as well as a reconstruction mapping* $g : \mathcal{U} \times \mathcal{V} \times \mathcal{Y} \to \hat{\mathcal{X}}$*, can be found, such that*

$$I(U;X) + I(V;X|UY) \leq R \ , \tag{3.6a}$$
$$I(U;Y) \geq E \ , \tag{3.6b}$$
$$\mathbb{E}_0\left[d\big(X, g(UVY)\big)\right] \leq D \ , \tag{3.6c}$$

with (U, V) *being two random variables satisfying* $U - V - X - Y$ *form a Markov chain with* $(X, Y) \sim p_0(x, y)$*, and* $\|\mathcal{U}\| \leq \|\mathcal{X}\| + 2$*,* $\|\mathcal{V}\| \leq \|\mathcal{X}\|\|\mathcal{U}\| + 1$.

Proof. The proof of Proposition 1 is given in Appendix B.1. □

Remark 8. *Observe that on one hand, the expression for the rate can be evaluated as follows:*

$$\begin{aligned} R &\geq I(U;X) + I(V;X|U) - I(V;Y|U) \\ &= I(U;Y) + \left[I(V;X) - I(V;Y)\right] \ , \end{aligned} \tag{3.7}$$

where the final equality stems from the Markov chain formed by the RVs, and on the other hand, from the fact that $U - V - X - Y$ *form a Markov chain, it is easy to see that*

$$\mathbb{E}_0\left[d\big(X, g'(VY)\big)\right] \leq \mathbb{E}_0\left[d\big(X, g(UVY)\big)\right] \leq D \ , \tag{3.8}$$

for some mapping g' *and any* g*. Note that the rate can now be seen as comprised of two different parts. The first part of the resulting expression in* (3.7) *is dedicated to detection since it only affects the error exponent, and is in fact identical to the expression of the error exponent given in* (3.6b) *in agreement with previous results [8, 9]. The second part of the rate is dedicated only to source reconstruction and therefore, the rate-error-distortion region can be seen as being equivalent to two uncoupled problems that share a common rate. In the following sections, we will see that this is not the case when general hypotheses are considered.*

Remark 9. *Note that while the assumption that distortion is only measured in case the detection of hypothesis* H_0 *is convenient, it is not necessary. As we assume that the*

distortion measure is bounded from above, the distortion under the decision H_0 (which may or may not be correct) may be expressed as follows:

$$\begin{aligned}
&\mathbb{E}_0\left[d\big(X, g(UVY)\big) | \text{“no assumption”}\right] && \text{(3.9a)}\\
&= \mathbb{E}_0\left[d\big(X, g(UVY)\big), \text{“correct detection”}\right] \Pr\{\text{“correct detection”}\} \\
&+ \mathbb{E}_0\left[d\big(X, g(UVY)\big), \text{“incorrect detection”}\right] \Pr\{\text{“incorrect detection”}\} && \text{(3.9b)}\\
&\leq \mathbb{E}_0\left[d\big(X, g(UVY)\big), \text{“correct detection”}\right] + \beta_n d_{max} \ , && \text{(3.9c)}
\end{aligned}$$

where d_{max} is assumed to be that maximal value that the distortion function $d(\cdot,\cdot)$ takes. As $\beta_n d_{max} \to 0$ when $n \to \infty$ the relaxation of the assumption that the distortion is only measured under correct detection does not change the optimal rate-error-distortion region. Note that the assumption that estimation is only done under the decision H_0 was not relaxed, only the fact that distortion is not measured under incorrect detection.

3.2.3 Binary Symmetric Source

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

$$X \sim \text{Bern}\left(\frac{1}{2}\right), \qquad \begin{cases} H_0: & Y = X + Z, \quad Z \sim \text{Bern}(p) \\ H_1: & Y \sim \text{Bern}\left(\frac{1}{2}\right) \perp X \ , \end{cases} \tag{3.10}$$

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 equal. Thus, a decision (or user identification) can be reached only through cooperation between the nodes. In the next theorem, the rate-error-distortion region for this problem is characterized by optimizing over all involved random variables in Theorem 1.

Theorem 2 (Rate-Error-Distortion region for Binary Symmetric Sources)**.** *The rate-error-distortion region for binary symmetric sources (BSS) and testing against independence is given by*

$$\begin{aligned}
R &\geq 1 - H_2(\alpha \star \beta \star p) + \theta\left[H_2(\alpha \star p) - H_2(\alpha)\right] \ , && \text{(3.11a)}\\
E &\leq 1 - H_2(\alpha \star \beta \star p) \ , && \text{(3.11b)}\\
D &\geq \theta\alpha - (1-\theta)\, p \ , && \text{(3.11c)}
\end{aligned}$$

for any $0 \leq \alpha, \beta \leq \frac{1}{2}$, $0 \leq \theta \leq 1$. Here $H_2(\cdot)$ is the binary entropy function.

Proof. The proof is given in Appendix B.2. □

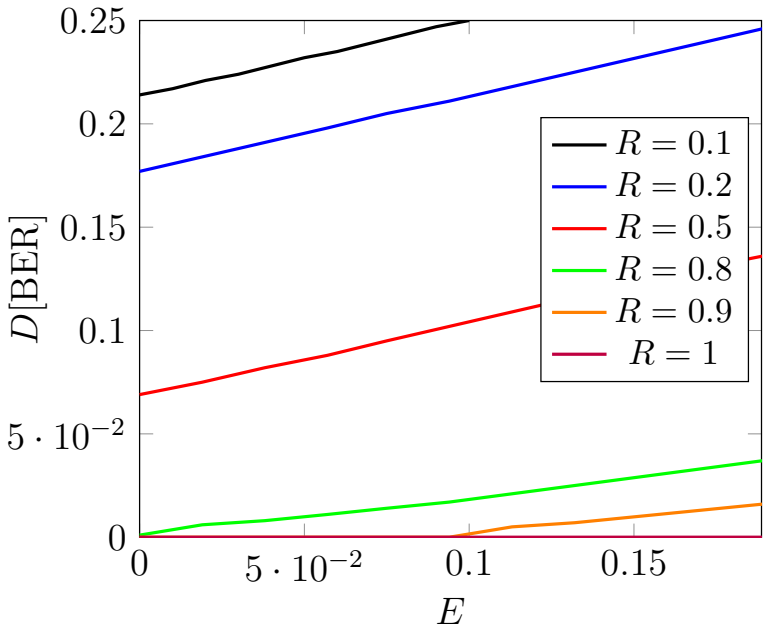


Figure 3.2: Numerical results of the optimal average distortion as a function of the desired error exponent of the second type, for different amounts of available rate and for $p = 0.25$, and testing against independence.

3.2.4 Numerical Results

We now present numerical results for the Binary Symmetric Source (BSS) case of testing against independence. Fig. 3.2 shows six curves, each representing the trade-off between user authentification and source reconstruction, expressed by the desired error exponent (second type) and the resulting average distortion of the source estimation, for a fixed value of available rate and for $p = 0.25$. Unsurprisingly, all curves are non-decreasing, meaning that when the probability of error is exponentially smaller, the amount of rate left for source reconstruction is smaller, resulting in a more crude estimation.

Assuming that both sources $\mathbf{X}^n$ and $\mathbf{Y}^n$ are available at a single location, Stein's Lemma yields an error exponent $E_{\max} = I(X;Y) = 1 - H_2(p) \approx 0.1887$. Obviously, this value constitutes an upper bound –uniform over the rate– on the achievable exponent in the distributed setup presented here. It can be seen that when $R < E_{\max}$, the average distortion reaches its maximal value $D_{\max} = p = 0.25$ for some $E < E_{\max}$. 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.

Note especially the curves for the rate values: $R = 0.9$ and $R = 1$ for which the rates comply with $R > H_2(p)$. According to Slepian-Wolf coding (see e.g. [2]), 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 source reconstruction, for any achievable choice of error exponent.

3.3 Joint Detection and Estimation - General Hypotheses

We now focus on the general case, where both hypotheses can be general distributions of two variables. Note that now, unlike the case of testing against independence, the performance of the system is measured by four quantities, namely the rate, the error exponent and two distortions, as source reconstruction is attempted under both hypotheses. Nevertheless, distortion is still measured under the assumption that the detection step was completed successfully. Unlike the case of testing against independence, optimality results for general distributed HT remain allusive. An achievable region [9] was inspired by the approach taken for testing against independence. We propose here an achievable region for the general hypothesis testing problem with source reconstruction constraints that makes use of binning for both purposes. The proposed region, while not necessarily optimal in general, aims at improving on known results for the testing part while also adding the reconstruction of the source.

3.3.1 System Model

As before, we suppose that the statistician observes $\mathbf{Y}^n$ samples directly and can be informed about $\mathbf{X}^n$ samples indirectly, via an encoding function $f_n : \mathcal{X}^n \to \{1, \dots, \|f_n\|\}$ of rate $n^{-1} \log \|f_n\| \leq R$. The code definition remains the same as in Definition 9 with two reconstruction functions $g_{n,i} : \{1, \dots, \|f_n\|\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}_i^n$. However, for each $i \in \mathbb{N}^\star$, random samples (x_i, y_i) are distributed according to one of two general joint distributions:

$$\begin{cases} H_0 : & p_0(x,y) = P_{XY}(x,y) \ , \\ H_1 : & p_1(x,y) = P_{\bar{X}\bar{Y}}(x,y) \ , \end{cases} \tag{3.12}$$

on $\mathcal{X} \times \mathcal{Y}$. Moreover, these samples are independent across time $i = \{1, \dots, n\}$, and we assume throughout this chapter that $P_X(x) = P_{\bar{X}}(x)$ and $P_Y(y) = P_{\bar{Y}}(y)$, $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}$.

Definition 11 (Rate-exponent-distortion region)**.** *A tuple* $(R, E, D_0, D_1, \epsilon) \in \mathbb{R}_+^5$ *is said to be* achievable *if, for any* $\delta > 0$*, there exists an* $(n, R+\delta)$*-code* $(f_n, \mathcal{A}_n, g_{n,0}, g_{n,1})$ *such that:*

$$\begin{aligned} n^{-1} \log \|f_n\| &\leq R \ , \\ \mathbb{E}_i\big[d_i\big(\mathbf{X}^n, g_{n,i}(f_n(\mathbf{X}^n), \mathbf{Y}^n)\big)\big] &\leq D_i + \delta \ , \ i = 0, 1 \\ -\frac{1}{n} \log \beta_n(\mathcal{A}_n) &\geq E - \delta \ , \\ \alpha_n(\mathcal{A}_n) &\leq \epsilon \ , \end{aligned} \tag{3.13}$$

where $\beta_n(\mathcal{A}_n) = \Pr\big(\mathcal{A}_n|XY \sim p_1(x,y)\big)$ *and* $\alpha_n(\mathcal{A}_n) = \Pr\big(\mathcal{A}_n^c|XY \sim p_0(x,y)\big)$*, and distortion is measured under the condition that node* B *correctly detects the correct hypothesis. The set of all such achievable tuples is denoted by* $\mathcal{R}^\star$ *and is referred to as the* rate-exponent-distortion region*.*

Remark 10. *Note the slight abuse of notation in the distortion argument of Definition 11: Conditioning on the hypothesis, along with the fact that we assume the distortion is measured only in case the detection phase was completed correctly, means that for each distortion argument the "correct" RVs are assumed to be used. Thus,* $\mathbb{E}_0\big[d_0\big(\mathbf{X}^n, g_{n,0}(f_n(\mathbf{X}^n), \mathbf{Y}^n)\big)\big] \leq D_0 + \delta$ *is the correct expression for the distortion under* H_0*, while* $\mathbb{E}_1\big[d_1\big(\bar{\mathbf{X}}^n, g_{n,1}(f_n(\bar{\mathbf{X}}^n), \bar{\mathbf{Y}}^n)\big)\big] \leq D_1 + \delta$ *is the corresponding expression under hypothesis* 1*.*

3.3.2 Achievable Rate-Error-Distortion Region

We now state our main result for the general joint distributed detection and reconstruction problem, which is a new achievable rate-error-distortion region. This region is inspired by the one offered for the special case of testing against independence. In a similar manner to the approach taken in Theorem 1, we derive an achievable region based on the separation of two distinguishable steps, namely user authentication and source reconstruction. The statistician first decodes the description needed to perform testing, and then reconstructs the samples sent by the encoder. However, the decision step requires two phases, as summarized in the corresponding constraints present in the error exponent of the next proposition.

Proposition 1 (Achievable rate-error-distortion region)**.** *A tuple* $(R, E, D_0, D_1) \in \mathbb{R}_+^4$*, is achievable for the distributed joint detection and reconstruction problem with general hypotheses, if there exists a positive rate* R' *satisfying:*

$$R \geq R' + I\big(P_{X|UY}; P_{V_0|XUY}|P_{UY}\big) + I\big(P_{\bar{X}|\bar{U}\bar{Y}}; P_{V_1|\bar{X}\bar{U}\bar{Y}}|P_{\bar{U}\bar{Y}}\big)\ , \tag{3.14a}$$

$$E \leq \inf_{Q_X \in \mathcal{P}(\mathcal{X})} \sup_{Q^\star_{U|X}(Q_X) \in \mathcal{P}(\mathcal{U})} \inf_{Q_Y \in \mathcal{P}(\mathcal{Y})} \inf_{\substack{Q_{UXY} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}) \\ Q_{U|X} = Q^\star_{U|X}}} \Big\{ \min\big[G(Q_{UXY}, Q_X, Q_Y, R'), \min_{\tilde{U}\tilde{X}\tilde{Y} \in \mathcal{L}(Q^\star_{UX}, Q^\star_{UY})} \mathcal{D}\big(P_{\tilde{U}\tilde{X}\tilde{Y}} \| P_{\bar{U}\bar{X}\bar{Y}}\big)\big]\Big\} \tag{3.14b}$$

$$D_0 \geq \mathbb{E}_0\Big[d_0\big(X, \hat{X}_0(UYV_0)\big)\Big]\ , \tag{3.14c}$$

$$D_1 \geq \mathbb{E}_1\Big[d_1\big(\bar{X}, \hat{X}_1(\bar{U}\bar{Y}V_1)\big)\Big]\ . \tag{3.14d}$$

Here, U *and* $\bar{U}$ *are auxiliary RVs such that* $Q_{U|X}(u|x) = Q_{\bar{U}|\bar{X}}(u|x)\ , \forall (u,x) \in \mathcal{U}\times\mathcal{X}$*,* V_0 *and* V_1 *are auxiliary random variables verifying the Markov chains* $U - V_0 - X - Y$ *and* $\bar{U} - V_1 - \bar{X} - \bar{Y}$ *(along with* U *and* $\bar{U}$ *respectively);* $\mathcal{L}(Q^\star_{UX}, Q^\star_{UY})$ *is the following set of*

random variables:

$$\mathcal{L}(Q^\star_{UX}, Q^\star_{UY}) = \Big\{ P_{\tilde{U}\tilde{X}\tilde{Y}} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}) \,\Big|\, P_{\tilde{U}\tilde{X}}(u,x) = Q^\star_{UX}(u,x), \\ P_{\tilde{U}\tilde{Y}}(u,y) = Q^\star_{UY}(u,y),\, \forall (u,x,y) \Big\} \,, \tag{3.15}$$

where $Q^\star_{UX}, Q^\star_{UY}$ *are joint distributions implied by* Q_X *and the chosen maximizer* $Q^\star_{U|X}$*, and*

$$G(Q_{UXY}, Q_X, Q_Y, R') = \\ \begin{cases} \min\limits_{i=\{0,1\}} \mathcal{D}\big(Q_{UXY} || P_{UXY_i}\big) + \big[R' - I\big(Q_X; Q_{U|X}\big) + I\big(Q_Y; Q_{U|Y}\big)\big]^+ & I\big(Q_X; Q_{U|X}\big) > R' \\ +\infty & \textit{else}\,, \end{cases} \tag{3.16}$$

with P_{UXY_i} *defined to be* $P_{UXY_0} \triangleq P_{UXY} = P_{XY}Q_{U|X}$ *in the case of hypothesis* 0 *and* $P_{UXY_1} \triangleq P_{\bar{U}\bar{X}\bar{Y}} = P_{\bar{X}\bar{Y}}Q_{\bar{U}|\bar{X}}$ *in the case of hypothesis* 1*.*

Proof. The proof is relegated to Appendix B.3. □

We emphasize that when a binning approach is taken, the expression (3.14b) for the error exponent E encapsulates the innate tension between two error events: decoding the description and testing based on it. Provided that a good representation $\mathbf{u}^n$ of the observed samples $\mathbf{x}^n$ at node A is reliably decoded at node B, the statistician is able to perform detection with a very large probability of success. However, such a good representation would also induce a very large size for the codebook which, for a given rate, would cause each bin to be very large in order to satisfy the rate constraint, making likely errors will appear during the decoding process of the right sequence from the specific bin. On the other hand, when a crude description is chosen, the codebook is smaller and thus so is each bin –if binning is at all necessary. The binning process is therefore not likely to significantly hurt performance, whereas the retrieved representation is much less valuable for the sake of performing the test because of the crude nature such description supplies about the samples $\mathbf{x}^n$.

In order to ensure the achievability of the error exponent introduced in Proposition 1, we take a "worst-case" approach. The minimization and maximization operators in the expression for E can thus be read as follows: For every possible type of vector $\mathbf{x}^n$, the encoder is allowed to choose its strategy of transmission (this is achieved by taking the supremum over $Q^\star_{U|X}$). Having chosen the distribution to generate the codebook, the proposed approach should apply for any type of observed vector $\mathbf{y}^n$, as well as for any joint type $(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n)$, as long as $Q^\star_{U|X}$ is respected. Much like the case of testing against independence, achievability is proven by dividing the problem into two distinct parts: hypothesis testing and source reconstruction. First, a common message –designed to allow detection– is communicated from node A to node B and is then used regardless of the probability distribution in effect which is still unknown at this stage. In order to do so, we choose a decoder based on the empirical entropy, similar to the *Empirical*

Mutual Information (MMI) decoder used in compound models (e.g. see [39] and references therein). Two private messages are then transposed upon this common message, each intended to be used (together with the common message) under each of the possible hypotheses. It should be emphasized that dividing the communication in two different phases may well be a suboptimal choice. However, we will see that even under such a choice, gains in the error exponent can be had.

Remark 11. *Much like in the case of testing against independence (see Remark 9), the assumption that distortion is only measured when correct detection has occurred is convenient but not necessary for the achievability of the region proposed in Proposition 1.*

3.3.3 Binary Symmetric Source

Having proposed a new approach for hypothesis testing with general hypotheses, based on binning, it is still not clear if this approach offers strict benefits in performance, when compared to the non-binning approach of [9]. As was demonstrated in Section 3.2, binning for testing is not necessary to achieve optimality in the case of testing against independence. One may further argue that as binning introduces additional error events, it is not clear weather or not it would be beneficial at all in the case of general hypotheses.

In the following, we investigate the benefits of binning through a Binary Symmetric Source (BSS). For the sake of simplicity, we consider the following lower bound over the performance, throughout the following numerical analysis [9]:

$$\min_{\tilde{U}\tilde{X}\tilde{Y}\in\mathcal{L}(Q^{\star}_{UX},Q_{UY})} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}}\|P_{\bar{U}\bar{X}\bar{Y}}) \geq \mathcal{D}(P_{UY}\|P_{\bar{U}\bar{Y}}) \ . \tag{3.17}$$

Consider the following statistical model:

$$X \sim \text{Bern}\left(\frac{1}{2}\right), \qquad \begin{cases} H_0: & Y = X + Z_0, \quad Z_0 \sim \text{Bern}(p) \\ H_1: & Y = X + Z_1, \quad Z_1 \sim \text{Bern}(q) \ , \end{cases} \tag{3.18}$$

where $\frac{1}{2} > q > p > 0$. Note that while H_1 does not imply independence between X and Y, the marginal distribution of Y is equal for both hypotheses, making a decision without cooperation impossible. This model was studied first in Wyner-Ziv [43] for source reconstruction. The optimal rate-distortion region (asymptotic regime) was shown to be

$$\begin{cases} R(D) = \inf\limits_{\theta,\delta} \left[\theta \left(H_2(p * \delta) - H_2(\delta)\right)\right] \ , \\ D = \theta\delta + (1-\theta)p \ , \end{cases} \tag{3.19}$$

where p is the crossover probability between the source X and the side information Y, and $p \star \delta$ is the binary convolution of p and δ. The parameters satisfy $0 \leq \theta \leq 1$ and $0 \leq \delta \leq \frac{1}{2}$. The achievability of this region was shown by using time-sharing between two strategies – in the first the auxiliary RV U is the result of passing X through a BSC with transition probability δ, while in the second U is degenerate.

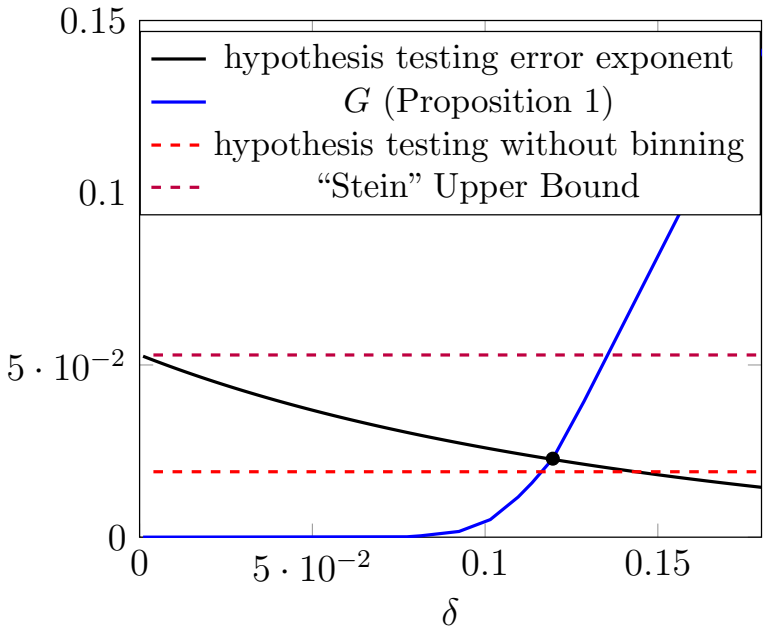


Figure 3.3: Error exponents for both error events in the BSC case with $p = 0.1$, $q = 0.2$, $R = 0.4$, under the strategy implied by Proposition 1. The resulting error exponent for each δ is the minimum between the two. Performance with a non-binned codebook is represented by a dashed line.

We now apply Proposition 1 to this setup, we choose to consider only distributions in which Q_X is a BSS, and U is the result of passing X through a BSC with crossover probability δ. While this is not necessarily an optimal choice, it can be justified as an optimal approach for the asymptotic regime, at least. To evaluate the resulting error exponent, we need to calculate two values. The first is given by:

$$\inf_{Q_Y} \inf_{\substack{Q_{UXY} \\ Q_{U|X}=Q^\star_{U|X}}} G(Q_{UXY}, R) \ , \tag{3.20}$$

as a function of $Q^\star_{U|X}$ (which, under our assumptions, boils down to be a function of δ). This expression encapsulates the error exponent of the event where the wrong sequence is chosen from the bin. The second quantity to calculate is given by:

$$\min_{\tilde{U}\tilde{X}\tilde{Y}\in\mathcal{L}(U)} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} \ | P_{\bar{U}\bar{X}\bar{Y}}) \geq \mathcal{D}(P_{UY} \| P_{\bar{U}\bar{Y}}) \ , \tag{3.21}$$

also as a function of $Q^\star_{U|X}$. This expression represents the error exponent of the event where, while using the right sequence, an error occurs during the detection process. Having calculated these two functions, we can pick $Q^\star_{U|X}$ such that the *minimum* between the two is *maximized*.

A visualization of the results achieved by the approach of Proposition 1, for the above discussed statistical model, is depicted in Figure 3.3. We choose to consider only distributions in which Q_X is a BSS and $Q^\star_{U|X}$ represents a BSC with transition probability δ,

as explained above. The "hypothesis testing" curve represents the error exponent of the probability of the event where a mistake is made in detection, when the correct sequence is used from the bin. The blue curve represents the event where a wrong sequence was erroneously selected from the bin (function G). The performance achieved by the optimal choice of δ, under the assumptions of the approach of Proposition 1 and the ones detailed above, is marked with a black dot.

The interesting tension that exists between the two error events is represented by the worst case (minimum) between those curves. when δ is very small, a sequence $\mathbf{u}^n$ can be found with high probability, such that $\mathbf{x}^n$ is very well described, and the codebook contains many sequences $\mathbf{u}^n$. Thus, given the right sequence $\mathbf{u}^n$, the error event during the test is not likely, and the error exponent of the event where the test fails is high. However, since the rate of communication is fixed, each bin has to contain many sequences in case δ is small, increasing the error probability in decoding the right sequence. When δ grows, the accuracy of the description of $\mathbf{x}^n$ by $\mathbf{u}^n$ is lower, making the probability of error of the test, while using the correct sequence, higher. The codebook, however, is smaller, making the task of choosing the right sequence in the bin easier. Note that the error exponent for choosing the sequence from within the bin has a threshold, under which it is zero. This threshold in this case is roughly $\delta \approx 0.08$, which is the value implied by [43] as the minimal value for the binning approach, in the asymptotic regime.

In addition, a lower bound can be found in Fig. 3.3. We emphasize that this bound is not drawn as a function of δ but rather depicts the best possible performance under the assumptions detailed above, when binning is not performed, as was done in [9]. Thus, δ is chosen to be the smallest possible, such that the size of the codebook would not exceed the available rate of communication. A trivial upper bound is also drawn by providing $\mathbf{x}^n$ to node B and then applying Stein's Lemma.

3.4 Revisiting the Detection of General Hypotheses

In this section, we focus on the detection part of the problem only, while still assuming general hypotheses. Although it was shown that gains in performance can be obtained by introducing binning as suggested in Proposition 1, we next show that the performance of detection can be further improved if source reconstruction is not required by the statistician. We start with the following proposition that uses a different approach for testing without source reconstruction.

Proposition 2 (Improved error exponent for general hypotheses)**.** *A pair (R, E) is an achievable rate and exponent pair for general hypothesis testing, without source reconstruction, provided that:*

$$E \leq \sup_{Q^{\star}_{U|X} \in \mathcal{P}(\mathcal{U})} \left\{ \min \left\{ \hat{G}(Q_{UXY}, R) \,, \min_{\tilde{U}\tilde{X}\tilde{Y} \in \mathcal{L}(Q^{\star}_{UX}\,, Q^{\star}_{UY})} \mathcal{D}\big(P_{\tilde{U}\tilde{X}\tilde{Y}} \| P_{\bar{U}\bar{X}\bar{Y}}\big) \right\} \right\} , \tag{3.22}$$

where

$$\hat{G}(Q_{UXY}, R) = R - \left[I\big(P_X; Q^{\star}_{U|X}\big) - I\big(P_Y; Q^{\star}_{U|Y}\big) \right] \tag{3.23}$$

and the set $\mathcal{L}(Q^{\star}_{UX}, Q^{\star}_{UY})$ *is defined by* (3.15)*. It is worth emphasizing that* $I\big(P_Y; Q^{\star}_{U|Y}\big)$ *in* (3.22) *is a direct consequence of the choice* $Q^{\star}_{U|X}$*. Moreover, the probability distribution* $Q^{\star}_{UY}$ *is derived from* $Q^{\star}_{U|X}$ *and* P_{XY}*.*

Proof. The proof of this proposition is relegated to Appendix B.4. □

The proof is very similar to that of Proposition 1. We basically derive the probability of error for a specific triplet of sequences $(\mathbf{x}^n, \mathbf{y}^n, \mathbf{u}^n)$, and then calculate the total probability of error by summing over all possible types and corresponding sequences included within each type. The main difference is that now source reconstruction is not required. Thus, instead of first selecting a sequence from within the bin and only then performing the test, we let node B operate over the entirety of the bin. The chosen strategy consists of going over all sequences within the bin. For each sequence $\mathbf{u}^n_i$ in the bin, we assume it is the correct one and perform the test by checking the typicality of the pair $(\mathbf{u}^n_i, \mathbf{y}^n)$ with relation to the hypothesis H_0. If a sequence is found in a bin such that $(\mathbf{u}^n_i, \mathbf{y}^n) \in T^n_{[UY]\delta}$, the decoder declares H_0. Otherwise, if no such sequence is found it declares H_1.

As was the case in Proposition 1, Proposition 2 implies that the resulting error exponent is the output of a trade-off between the exponents of the probabilities of two error events. In this case, the trade-off that controls $\beta_n \approx \exp(-nE)$ is between: the probability of erroneous detection while using the right sequence; and the probability of having a different sequence in the bin that is jointly typical with $\mathbf{y}^n$ and thus would make the decoder declare H_0. It turns out, that this trade-off is much preferable to the one offered by Proposition 1, as we can bound the set of sequences that might "confuse" the decoder in a manner that is not dependent on the type of $\mathbf{y}^n$. For instance, the minimizations over Q_X, Q_Y and Q_{UXY} (as seen in Proposition 1) are not longer necessary. This issue has a positive effect on behavior of the error exponent. As a matter of fact, the fact that the original sequence sent by the encoder is not retrieved implies that this strategy is not adapted for the joint problem of detection and source reconstruction.

Remark 12. *Another advantage of this strategy over the one given in Proposition 1 is that while knowledge over the probability distribution implied by* $P_{\bar{X}\bar{Y}}$ *is required in order to analyze performance, such knowledge is not needed in order to perform the test. This stems from the fact that here, the system only tests if* H_0 *is true or not rather than testing* H_0 *against* H_1*. In addition, we do not need to assume that* $P_X = P_{\bar{X}}$ *nor* $P_Y = P_{\bar{Y}}$*.*

3.4.1 Binary Symmetric Source

Having proposed two new approaches for distributed testing with general hypotheses, one that allows source reconstruction (Proposition 1) and the other that does not (Proposition 2), it is interesting to compare the performance in detection achieved under each of the approaches. In the following, we use the BSS example, presented in Section 3.3.3, in order to compare the two approaches.

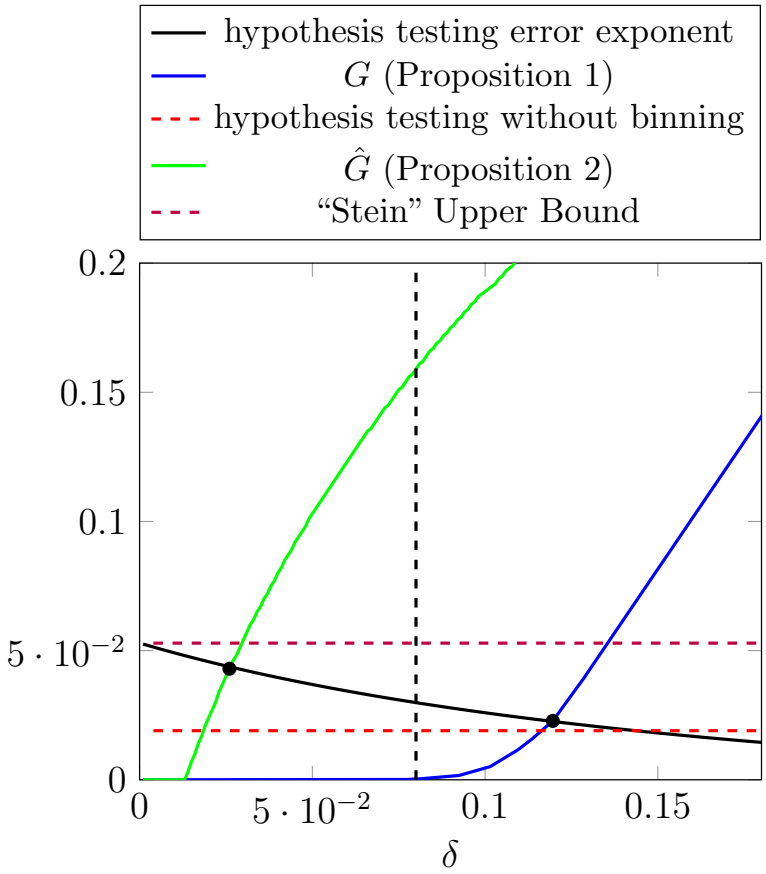


Figure 3.4: Error exponents for both error events in the BSC case with $p = 0.1$, $q = 0.2$, $R = 0.4$, under the strategies implied by Propositions 1 and 2. The resulting error exponent for each δ is the minimum between the two error events. Performance with a non-binned codebook is represented by a dashed line.

The results implied by Proposition 2 can be calculated in a very similar fashion to the calculation performed above for the performance under Proposition 1. In this case, the trade-off is between the curve representing the error while using the correct sequence as was mentioned in (3.21), and the curve implied by $\hat{G}$, representing the event of an error caused through the testing of a different sequence. A visualization of the performance achieved by each of the proposed methods for general hypotheses in the case of a BSS is plotted in Fig. 3.4. As before, we choose to consider only distributions in which Q_X is a BSS and $Q^{\star}_{U|X}$ represents a BSC with transition probability δ.

The trade-off between the two error events represented by Proposition 2 is apparent through the curve of the error exponent related to the testing error while testing the correct sequence, along with the "binning error exponent" denoted by the curve $\hat{G}$. Now, the additional error event –other than committing an error while using the correct sequence which turns out to be the same as before– is the event where a different sequence in the bin "confuses" the decoder by being jointly typical with $\mathbf{y}^n$. While this curve is lower bounded by the curve representing G for all cases, it can be seen that in the present

case this approach is largely superior. As under both approaches we are allowed to select the strategy $Q^{\star}_{U|X}$ (in this specific case δ) freely, the optimal approach under each of the propositions would be to choose the corresponding intersection point between the curve representing G or $\hat{G}$ and the curve entitled "Hypothesis Testing Error Exponent" in Fig. 3.4. These two points are marked with black dots.

3.4.2 Assessing the Gain in Performance

In this section we show that the performance gain shown for the specific example of binary symmetric sources is in fact general for many cases. In order to do so, we choose to examine a "cross-section" of the performance gain, at the point where $R = I(\bar{U};\bar{X}|\bar{Y})$. This cross-section is illustrated for the BSC example in Fig. 3.4 by a black dashed line. We choose this point because of its importance to problems where both detection and source-estimation are required at the receiver, as seen in previous sections. When $R < I(\bar{U};\bar{X}|\bar{Y})$ joint detection and estimation cannot be assured for both hypotheses, under any of the approaches presented in this work. It can be seen in Fig. 3.4, that it is at this point that the curve of $G[Q_{UXY}, R]$ leaves 0. This is in fact general for all cases, and is implied by the decoding approach of Proposition 1, where a single sequence *must* first be chosen, before detection is performed.

In the example presented above, at the same point, the curve for $\hat{G}[Q_{UXY}, R]$ is above the one representing the 'hypothesis testing error exponent' while using the intended sequence (seen in black in Fig. 3.4). This implies that at this point, performance is *not limited* by the binning approach. We now check if this observation is true in general:

$$\left[\hat{G}[Q_{UXY}, R] - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}})\right]\Big|_{R=I(\bar{U};\bar{X}|\bar{Y})} \tag{3.24a}$$

$$= [R - I(U;X) + I(U;Y) - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}})]|_{R=I(\bar{U};\bar{X}|\bar{Y})} \tag{3.24b}$$

$$= [R - I(U;X|Y) - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}})]|_{R=I(\bar{U};\bar{X}|\bar{Y})} \tag{3.24c}$$

$$= I(\bar{U};\bar{X}|\bar{Y}) - I(U;X|Y) - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}}) \tag{3.24d}$$

$$= I(\bar{U};\bar{X}\bar{Y}) - I(\bar{U};\bar{Y}) - I(U;XY) + I(U;Y) - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}}) \tag{3.24e}$$

$$= I(U;Y) - I(\bar{U};\bar{Y}) - \mathcal{D}(P_{UY}||P_{\bar{U}\bar{Y}}) \tag{3.24f}$$

$$\triangleq (*) \ . \tag{3.24g}$$

Here, (3.24c) stems from the Markov chain $U - X - Y$, while (3.24f) stems from the same Markov chain, as well as $\bar{U} - \bar{X} - \bar{Y}$. In addition, in this equality it is assumed that $P_X = P_{\bar{X}}$ (this is not a supplementary assumption, as this needs to be assumed at least for the sake of Proposition 1, as explained above). Through the chain rule for KL divergence we get that:

$$(*) = H(\bar{Y}|\bar{U}) - H(Y|U) - \mathcal{D}(P_{Y|U}||P_{\bar{Y}|\bar{U}}|P_U) \ , \tag{3.25}$$

where $\mathcal{D}(P_{Y|U}||P_{\bar{Y}|\bar{U}}|P_U)$ is the *conditional* KL-divergence, defined as:

$$\mathcal{D}(P_{Y|U}||P_{\bar{Y}|\bar{U}}|P_U) = \sum_{\substack{y\in\mathcal{Y}\\u\in\mathcal{U}}} P_{UY}(u,y)\log\frac{P_{Y|U}(y|u)}{P_{\bar{Y}|\bar{U}}(y|u)} \ . \tag{3.26}$$

In order to check if the performance at this point is not limited by the binning approach of Proposition 2, we would like to check if this expression is positive, or equivalently if

$$H(\bar{Y}|\bar{U}) - H(Y|U) \geq \mathcal{D}(P_{Y|U}||P_{\bar{Y}|\bar{U}}|P_U) \ . \tag{3.27}$$

This is a conditional version of Theorem 3 in [95]. A sufficient (but not necessary) condition for this inequality to hold is thus that $\bar{Y}$ is *majorized* by Y, for any choice of U:

Definition 12 ([95])**.** *Consider discrete probability distributions $P = \{p_i\}$ and $Q = \{q_i\}$ defined on the positive integers labeled in decreasing probabilities, i.e.,*

$$\begin{aligned} p_i &\geq p_{i+1} \ , \\ q_i &\geq q_{i+1} \ . \end{aligned} \tag{3.28}$$

Q is majorized by P if for all $k = 1, 2, \ldots$

$$\sum_{i=1}^{k} q_i \leq \sum_{i=1}^{k} p_i \ . \tag{3.29}$$

Lemma 17 ([95])**.** *If Q is majorized by P, then*

$$H(Q) - H(P) \geq D(P||Q) \ . \tag{3.30}$$

Proof. Refer to reference [95]. □

When considering the conditional case, as we are required to do here, it is enough to verify the majorization condition in Lemma 17 for the average of Y (respectively, $\bar{Y}$) over U. Nevertheless, we will restrict ourselves further by demanding that $(\bar{Y}|\bar{U} = u)$ is majorized by $(Y|U = u)$ for each $u \in \mathcal{U}$. A sufficient (but not necessary) condition for this constraint to be met is that $(\bar{Y}|\bar{X} = x)$ is majorized by $(Y|X = x)$ for any $x \in \mathcal{X}$. In such a case, there will always be a strategy $Q^{\star}_{U|X}$ (not necessarily unique) that achieves the maximum in Proposition 2, and such that the majorization constraint holds. Thus, performance is not limited by our proposed binning approach of Proposition 2, for any setting that complies with this condition, at our chosen reference point $R = I(\bar{U};\bar{X}|\bar{Y})$. Comparing this to the approach in Proposition 1, where at the same reference point binning reduces the error exponent of interest to zero, the benefits of the approach presented in Proposition 2 are clear.

Remark 13. *While at a first glance enforcing the majorization condition for each $x \in \mathcal{X}$ might seem unnecessarily strict, in fact it still includes many interesting problems, including settings in which H_0 and H_1 imply the same channel from X to Y, with the difference that the channel implied by H_1 is noisier. This is in fact the case of the BSS example above.*

3.5 Closing Remarks

In this chapter, the problem of joint detection and source estimation over a unidirectional link was studied. This scenario may arise, for example, when an authentication system aims to prevent the unauthorized injection of messages into a public channel, assuring the receiver of a message of the legitimacy of its sender. In this setup a user (referred to as node A) is required to communicate a lossy description of a memoryless source to a statistician (referred to as node B) whose task is to verify that the encoding user is the individual he claims to be and then according to its identity to reconstruct the message based on the adequate distortion measure, much like in [46, 47]. However, in the setup considered here the receiver is unaware of the value of its information as well, which leads to a two-step approach where first a decision has to be made about the identity of node A before source reconstruction can take place.

When testing against independence, this two-step approach turns out to be optimal. In this case, detection can be performed optimally as in [8], while source reconstruction is performed à la Wyner-Ziv [43], and the two-step approach does not induce performance degradation. When testing with general hypotheses, a similar, albeit more involved, approach produced a new achievable rate-error-distortion region. Here, optimality may be hard to reach, as optimality results stay allusive even in the case where the receiver is aware of the value of the side information (see [42] and references therein). Nevertheless, we showed that the two-step approach, which was optimal in the case of testing against independence, induces in the general case a significant loss in performance. It was shown that when source reconstruction is not required, valuable information for testing can be compressed much further than in the opposite case, improving significantly the performance of detection.

Chapter 4

Interactive Distributed Hypothesis Testing

4.1 Overview

In this chapter, based on work published in [88, 91], we broaden our view of the binary distributed hypothesis testing problem to include interactive communication. While still considering a two-node system, the link connecting the nodes is assumed to be bidirectional (see Figure 4.1 for a visual depiction of the system) and constrained by a *sum-rate* constraint, which the participants are free to allocate in any way to improve performance. In addition, it is assumed that the location in which the decision is made is unimportant. Note that this assumption is not very stringent, as sharing a binary decision is a zero-rate task.

We start this chapter by considering the case where only one "round" of communication is permitted (i.e., without loss of generality, we assume that node A sends the first message, which is then replied with a message from node B. A decision must be reached after these two messages). We propose an achievable error-exponent for the error of the second type in this case, under a fixed constraint over the error of the first type. This exponent is inspired by the one proposed in [9] for the case of unidirectional communication. We then

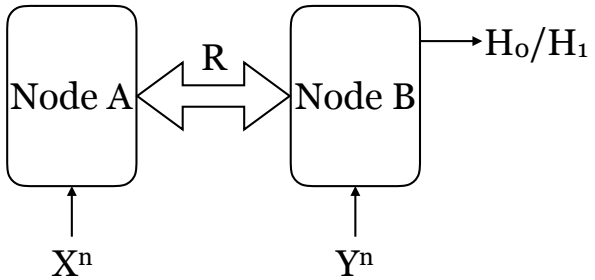


Figure 4.1: Cooperative Hypothesis Testing model, with interactive communication.

use similar methods in order to extend this result to any *finite* number of communication rounds between the participants.

The case of *testing against independence* is then revisited. When only one round of communication is allowed, we show that a known achievable error-exponent [20] can indeed be attained through our general error-exponent, when testing against independence is assumed. We proceed to show that this exponent is in fact optimal by proving a converse theorem. While an extended version of this exponent is shown to be achievable when more than one round of communication is allowed, we discuss the reasons that this exponent is no longer optimal.

Finally, we discuss the case of interactive hypothesis testing with zero rate. We show that the asymptotic performance in this case is equivalent to the performance under a much stricter *one-bit communication* constraint.

4.2 System Model

In the bidirectional communication scenario two statisticians are assumed to observe the i.i.d. realizations of two RVs, X and Y respectively, as depicted in Figure 4.1. The two RVs are jointly distributed in one of two ways, as was the case in the previous chapter:

$$\begin{cases} H_0: & p_0(x,y) = P_{XY}(x,y) \ , \\ H_1: & p_1(x,y) = P_{\bar{X}\bar{Y}}(x,y) \ . \end{cases} \tag{4.1}$$

Communication between the two statisticians is assumed to be done in rounds, with node A starting the interaction. These interactions are limited, however, by a total (exponential) rate R bits per symbol. That is, if each of the nodes sees n realizations, the total amount of bits allowed to exchange data between the nodes before the decision is made is $\exp(nR)$. The data exchange is assumed to be *perfect*, meaning that within the rate limit no errors are introduced by the communication. It is assumed that the total rate can be distributed by the two statisticians in any way that is beneficial to performance. Moreover, we assume that it does not matter *where* the decision is finally made, as its transmission can be done at no cost.

The definition of the two error events, of Type I and Type II, and their respective probabilities, stays the same as before. The task of the statisticians remains to declare the true probability distribution out of the two options while minimizing the probability of error. In a similar fashion to the unidirectional case, as analyzed in Chapter 3, the goal is to find the exponential rate: $-\frac{1}{n}\log\beta_n$ (n being the number of samples) s.t. $\beta_n \to 0$ as $n \to \infty$, while fixed constraints are enforced on α_n and the total exchange rate R. The participants employ a K-round strategy, which is defined as follows:

Definition 13 (K-round collaborative HT)**.** *A K-round decision code for the two node collaborative hypothesis testing system, when each of the statisticians is allowed to observe*

$\mathbf{X}^n$ *and* $\mathbf{Y}^n$ *realizations of* X *and* Y*, respectively, is defined by a sequence of encoders and a decision mapping:*

$$f_{[k]} : \mathcal{X}^n \times \prod_{i=1}^{k-1} \{1, \ldots, |g_{[i]}|\} \longrightarrow \{1, \ldots, |f_{[k]}|\} \ , \ k = [1 : K] \tag{4.2}$$

$$g_{[k]} : \mathcal{Y}^n \times \prod_{i=1}^{k} \{1, \ldots, |f_{[i]}|\} \longrightarrow \{1, \ldots, |g_{[k]}|\} \ , \ k = [1 : K] \tag{4.3}$$

$$\phi : \mathcal{X}^n \times \prod_{i=1}^{K} \{1, \ldots, |g_{[i]}|\} \longrightarrow \{0, 1\} \ , \tag{4.4}$$

where $f_{[k]}$ *and* $g_{[k]}$ *are encoder mappings with image sizes satisfying* $\log |f_{[i]}| \equiv \mathcal{O}(n)$ *and* $\log |g_{[i]}| \equiv \mathcal{O}(n)$*, respectively, while* ϕ *is the decision mapping. The corresponding Type I and II error probabilities are given by*

$$\alpha_n(R\,|K) \coloneqq \Pr\left[\phi\big(X^n, g_{[1:K]}\big) = 1 \,|\, \mathbf{X}^n\mathbf{Y}^n \sim P_{XY}\right] \ , \tag{4.5}$$

$$\beta_n(R\,|K) \coloneqq \Pr\left[\phi\big(X^n, g_{[1:K]}\big) = 0 \,|\, \mathbf{X}^n\mathbf{Y}^n \sim P_{\bar{X}\bar{Y}}\right] \ . \tag{4.6}$$

An exponent E *to the error probability of Type II, constrained to an error probability of Type I to be below* $\epsilon > 0$ *and a total exchange rate* R*, is said to be feasible, if for any* $\varepsilon > 0$ *there exists a code satisfying:*

$$-\frac{1}{n} \log \beta_n(R, \epsilon\,|K) \geq E - \varepsilon \ , \tag{4.7}$$

$$\frac{1}{n} \sum_{k=1}^{K} \log\left(|g_{[k]}||f_{[k]}|\right) \leq R + \varepsilon \ , \ \alpha_n(R\,|K) \leq \epsilon \ , \tag{4.8}$$

provided that n *is large enough. The supremum of all feasible exponents for given* (R, ϵ) *is defined to be the optimal error exponent.*

4.3 Collaborative Hypothesis Testing with One Round

In this section, we present a *feasible error exponent* $-\frac{1}{n} \log \beta_n(R, \epsilon\,|K = 1)$ to the error probability of Type II, under any fixed constraint $\epsilon > 0$ on the error probability of Type I for a total exchange rate R. Here, we only consider one round of exchange whereby each of the nodes exchanges one statistics (or message) before a decision is made. The extension to the case with multiple exchanging rounds is relegated to the next section.

Proposition 3 (Sufficient conditions for one round of interaction)**.** *Let* $\mathscr{S}(R) \subset \mathcal{P}(\mathcal{U} \times \mathcal{V})$ *and* $\mathscr{L}(U, V) \subset \mathcal{P}(\mathcal{U} \times \mathcal{V} \times \mathcal{X} \times \mathcal{Y})$ *denote the sets of probability measures defined in terms of corresponding RVs:*

$$\begin{aligned} \mathscr{S}(R) \coloneqq \big\{UV : \ & I(U;X) + I(V;Y|U) \leq R \\ & U - X - Y \ , \ V - (U, Y) - X \ , \ |\mathcal{U}|, |\mathcal{V}| < +\infty\big\} \ , \end{aligned} \tag{4.9}$$

$$\mathscr{L}(U, V) \coloneqq \left\{\tilde{U}\tilde{V}\tilde{X}\tilde{Y} \ : \ P_{\tilde{U}\tilde{V}\tilde{X}} = P_{UVX} \ , \ P_{\tilde{U}\tilde{V}\tilde{Y}} = P_{UVY}\right\} \ . \tag{4.10}$$

A feasible error exponent to the error probability of Type II, when the total exchange rate is R (bits per sample), is given by

$$\lim_{\epsilon\to 0}\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R,\epsilon\,|K=1) \geq \max_{UV\in\mathscr{S}(R)}\ \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}\in\mathscr{L}(U,V)} \mathcal{D}\big(P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}||P_{\bar{U}\bar{V}\bar{X}\bar{Y}}\big)\ . \tag{4.11}$$

Proof. The proof of Proposition 3 is given in Appendix C.1. □

The proposed region of Proposition 3 constitutes an extension of the region given in [9], for the case of testing general hypotheses with unidirectional communication. By setting the resources allocated to the message of node B to zero ($V\equiv 0$), the region of [9] is retrieved immediately. Note that the equivalent strategy of unidirectional communication from node B to node A is also contained in this result, by setting $U\equiv 0$.

Remark 14. *The error-exponent proposed for the error of type II by Proposition 3, under a fixed constraint over the error probability of type I can be further improved by using binning, as discussed in Chapter 3, for unidirectional communication.*

4.4 Collaborative Hypothesis Testing with Multiple Rounds

We now allow the statisticians to exchange data over an arbitrary but *finite* number of exchange rounds, and investigate the extension of Proposition 3 to this more general case. The corresponding result is stated below.

Proposition 4 (Sufficient conditions for K-rounds of interaction)**.** *Let $\mathscr{S}(R)$ and $\mathscr{L}\big(U_{[1:K]},V_{[1:K]}\big)$ denote the sets of probability measures defined in terms of corresponding RVs:*

$$\begin{aligned}\mathscr{S}(R) \coloneqq \Big\{U_{[1:K]}V_{[1:K]} : R \geq \sum_{k=1}^{K}\big[&I(X;U_{[k]}|U_{[1:k-1]}V_{[1:k-1]}) \\ &+ I(Y;V_{[k]}|U_{[1:k-1]}V_{[1:k-2]})\big]\ , \\ U_{[k]} - \big(X,U_{[1:k-1]},V_{[1:k-1]}\big) - Y\ &,\ |\mathcal{U}_{[k]}|<+\infty\ , \\ V_{[k]} - \big(Y,U_{[1:k]},V_{[1:k-1]}\big) - X\ &,\ |\mathcal{V}_{[k]}|<+\infty\ ,\forall\, k\in[1:K]\Big\}\ ,\end{aligned} \tag{4.12}$$

$$\begin{aligned}\mathscr{L}\big(U_{[1:K]},V_{[1:K]}\big) \coloneqq \Big\{&\tilde{U}_{[1:K]}\tilde{V}_{[1:K]}\tilde{X}\tilde{Y} : \\ &P_{\tilde{U}_{[1:K]}\tilde{V}_{[1:K]}\tilde{X}} = P_{U_{[1:K]}V_{[1:K]}X}\ ,\ P_{\tilde{U}_{[1:K]}\tilde{V}_{[1:K]}\tilde{Y}} = P_{U_{[1:K]}V_{[1:K]}Y}\Big\}\ ,\end{aligned} \tag{4.13}$$

where $U_{[1:k]}\coloneqq(U_{[1]},\ldots,U_{[k]})$ and $V_{[1:k]}\coloneqq(V_{[1]},\ldots,V_{[k]})$ represent the exchanged data between nodes A and B until round k. A feasible error exponent to the error probability

of Type II, when the total (over K-rounds) exchange rate is R (bits per sample), is given by

$$\lim_{\epsilon\to 0}\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R,\epsilon|K) \geq \max_{\mathscr{S}(R)} \min_{\mathscr{L}\left(U_{[1:K]},V_{[1:K]}\right)} \mathcal{D}\Big(P_{\tilde{U}_{[1:K]}\tilde{V}_{[1:K]}\tilde{X}\tilde{Y}}\Big|\Big|P_{\bar{U}_{[1:K]}\bar{V}_{[1:K]}\bar{X}\bar{Y}}\Big)\ . \tag{4.14}$$

Proof. The proof of Proposition 4 is given in Appendix C.2. □

This proposition is very clearly an extension of Proposition 3 to allow multiple rounds of interaction. The implication of this result is as follows. Given a limited budget of rate R for data exchange, which the nodes can divide as they choose into any finite number of K exchange rounds, the gain of interaction attained through the different characteristics of the underlying Markov process between the RVs comes at no cost in terms of the form of the expression for the error exponent.

Remark 15. *For reasons of brevity and clarity, we chose in this work to concentrate on scenarios where the interaction begins and ends at node A. However, it is easy to see that this does not necessarily need to be the case. The process could start or end at node B, implying that the final round of exchange is in fact only half of a round, without any significant changes to the theory or our proofs.*

4.5 Collaborative Testing Against Independence

We now concentrate on the special problem of testing against independence, where it is assumed that under H_1 the n observed samples of the RVs (X,Y) defined on $(\mathcal{X}\times\mathcal{Y},\mathcal{B}_{\mathcal{X}\times\mathcal{Y}})$ are distributed according to a product measure:

$$\begin{cases} H_0: & P_{XY}(x,y)\ ,\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}\ , \\ H_1: & P_{\bar{X}\bar{Y}}(x,y)=P_X(x)P_Y(y)\ ,\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}\ , \end{cases} \tag{4.15}$$

where $P_X(x)$ and $P_Y(y)$ are the marginal probability measures implied by $P_{XY}(x,y)$. Testing against independence in a cooperative scenario was first studied in [20], for the case of a single round of interaction. It was shown that a feasible error exponent to the error probability of Type II is given by

$$\lim_{\epsilon\to 0}\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R,\epsilon|K=1) \geq E(R) \tag{4.16}$$

subject to a total available exchange rate R, where:

$$E(R) \coloneqq \max_{\substack{P_{U|X}:\mathcal{X}\mapsto\mathcal{P}(\mathcal{U}) \\ P_{V|UY}:\mathcal{U}\times\mathcal{Y}\mapsto\mathcal{P}(\mathcal{V}) \\ \text{s.t. } I(U;X)+I(V;Y|U)\leq R}} \big[I(U;Y)+I(V;X|U)\big]\ . \tag{4.17}$$

While the proof of feasibility inspired the approach taken in Proposition 3 for general hypotheses, unfortunately, the auxiliary RVs identified in the *weak* unfeasibility proof in [20] do not match the required Markov chains to lead to a feasible exponent (the reader may refer to [50, 96] for further details).

In this section, we revisit the problem of characterizing the reverse inequality in (4.16). We prove a *weak unfeasibility* result, determining necessary and sufficient conditions to the optimality of the error exponent (4.17) satisfying $\alpha_n \leq \epsilon$ for *any* $0 < \epsilon < 1$ (i.e., we prove that the exponent in (4.17) is optimal in the case where we constrain α_n to go to 0 with n). We first show that Proposition 3 implies the feasibility part, i.e., inequality (4.16), and then follow with a new proof for the unfeasibility (for ϵ arbitrarily small) of any higher exponent.

Theorem 3 (Necessary and sufficient conditions for testing against independence with $K = 1$)**.** *The* optimal *error exponent to the error probability of Type II for testing against independence is given by*

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n(R, \epsilon | K = 1) \coloneqq E(R) \ , \quad \forall \ 0 < \epsilon < 1 \ , \tag{4.18}$$

where $E(R)$ is defined in (4.17)*, and R denotes the available rate of interaction between the statisticians and ϵ is the error probability of Type I.*

Proof. In order to show the feasibility to the exponent (4.17) through the general result stated in Proposition 3, it is convenient to use an intermediary form of the exponent for general hypotheses, which appears in the final steps of the proof (see (C.17g)):

$$\begin{aligned} \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n(R, \epsilon | K = 1) \geq \\ \max_{UV \in \mathscr{S}(R)} \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y} \in \mathscr{L}(U,V)} \Big[\mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} || P_{\bar{U}\bar{X}\bar{Y}}) + I(\tilde{X}; \tilde{V} | \tilde{U}\tilde{Y}) \Big] \ . \end{aligned} \tag{4.19}$$

We analyze each of these components separately:

$$\begin{aligned} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} || P_{\bar{U}\bar{X}\bar{Y}}) &= \mathcal{D}(P_{\tilde{U}\tilde{Y}} || P_{\bar{U}\bar{Y}}) + \mathcal{D}(P_{\tilde{X}|\tilde{U}\tilde{Y}} || P_{\bar{X}|\bar{U}\bar{Y}} | P_{\tilde{U}\tilde{Y}}) && \text{(4.20a)} \\ &= I(U;Y) + \mathcal{D}(P_{\tilde{X}|\tilde{U}\tilde{Y}} || P_{\bar{X}|\bar{U}} | P_{\tilde{U}\tilde{Y}}) && \text{(4.20b)} \\ &= I(U;Y) + \mathcal{D}(P_{\tilde{X}|\tilde{U}\tilde{Y}} || P_{\tilde{X}|\tilde{U}} | P_{\tilde{U}\tilde{Y}}) + \mathcal{D}(P_{\tilde{X}|\tilde{U}} || P_{\bar{X}|\bar{U}} | P_{\tilde{U}}) && \text{(4.20c)} \\ &\geq I(U;Y) + \mathcal{D}(P_{\tilde{X}|\tilde{U}\tilde{Y}} || P_{\tilde{X}|\tilde{U}} | P_{\tilde{U}\tilde{Y}}) \ , && \text{(4.20d)} \end{aligned}$$

where (4.20a) is due to the chain rule and $\mathcal{D}(P_{\tilde{X}|\tilde{U}\tilde{Y}} || P_{\bar{X}|\bar{U}\bar{Y}} | P_{\tilde{U}\tilde{Y}})$ is the conditional KL-divergence; 4.20b stems from the assumption of testing against independence, as well as the Markov chain $\bar{U} - \bar{X} - \bar{Y}$ and the fact that $P_{\tilde{U}\tilde{Y}} = P_{UY}$; and (4.20d) is due to the fact

that the KL-divergence is non-negative. To conclude the analysis, we note that:

$$\begin{aligned}
&\mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}}||P_{\bar{U}\bar{X}\bar{Y}}) \geq \\
&\quad I(U;Y) + \sum_{(u,x,y)\in\mathcal{U}\times\mathcal{X}\times\mathcal{Y}} P_{\tilde{U}\tilde{X}\tilde{Y}}(u,x,y)\log\left(\frac{P_{\tilde{X}|\tilde{U}\tilde{Y}}(x|u,y)}{P_{\tilde{X}|\tilde{U}}(x|u)}\right) \\
&= I(U;Y) + \sum_{(u,x,y)\in\mathcal{U}\times\mathcal{X}\times\mathcal{Y}} P_{\tilde{U}\tilde{X}\tilde{Y}}(u,x,y)\log\left(\frac{P_{\tilde{X}\tilde{Y}|\tilde{U}}(x,y|u)}{P_{\tilde{X}|\tilde{U}}(x|u)P_{\tilde{Y}|\tilde{U}}(y|u)}\right) \\
&= I(U;Y) + I(\tilde{X};\tilde{Y}|\tilde{U}) \ .
\end{aligned} \tag{4.21}$$

As for the second term in (4.19), we express it as follows:

$$I(\tilde{V};\tilde{X}|\tilde{U}\tilde{Y}) = I(\tilde{V}\tilde{Y};\tilde{X}|\tilde{U}) - I(\tilde{X};\tilde{Y}|\tilde{U}) \geq I(\tilde{V};\tilde{X}|\tilde{U}) - I(\tilde{X};\tilde{Y}|\tilde{U}) \ . \tag{4.22}$$

This allows us to conclude through (4.19) that

$$\begin{aligned}
\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R,\epsilon\,|K=1) &\geq \max_{UV\in\mathscr{S}(R)} \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}\in\mathscr{L}(U,V)} \Big[I(U;Y) + I(\tilde{V};\tilde{X}|\tilde{U})\Big] \\
&= \max_{UV\in\mathscr{S}(R)} \left[I(U;Y) + I(V;X|U)\right] \ ,
\end{aligned} \tag{4.23}$$

which completes the proof of feasibility through Proposition 3. The proof of converse is given in Appendix C.3 □

Remark 16. *In a similar manner to Theorem 3, a* feasible *error exponent to the error probability of Type II with K rounds is given by*

$$\begin{aligned}
&\lim_{\epsilon\to 0}\ \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R,\epsilon|K) \geq \\
&\max_{U_{[1:K]}V_{[1:K]}\in\mathscr{S}(R)} \sum_{k=1}^{K} \left[I\big(U_{[k]};Y|U_{[1:k-1]}V_{[1:k-1]}\big) + I\big(V_{[k]};X|U_{[1:k]}V_{[1:k-1]}\big)\right] \ .
\end{aligned} \tag{4.24}$$

The proof of the feasibility of (4.24) *follows largely the same path as the one for the feasibility part provided for Theorem 3. However, for $K > 1$ our unfeasibility proof does not hold and this feasible exponent result may not longer be optimal. The reasons for this are explained in Appendix C.4.*

4.6 Collaborative Hypothesis Testing with Zero Rate

We now consider another special case of Proposition 4, whereby testing is done over two general hypotheses, but the total exchange rate is zero. It is worth mentioning that zero-rate does not mean that *no information exchange* is possible, but rather that the size of the codebook grows slower than exponentially with the blocklength n, as stated in the following theorem.

Theorem 4 (Necessary and sufficient conditions under zero-rate)**.** *Let P_{XY} and $P_{\bar{X}\bar{Y}}$ be any probability measures such that $supp(P_{\bar{X}\bar{Y}}) = supp(P_{XY}) = \mathcal{X} \times \mathcal{Y}$. Assume the total exchange rate $R = 0$, that is:*

$$\sum_{k=1}^{K} \log |f_{[k]}| + \sum_{k=1}^{K} \log |g_{[k]}| \equiv o(n) \ , \tag{4.25}$$

the optimal error exponent to the probability of Type II is given by

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n(R = 0, \epsilon \,|K) = \min_{\tilde{X}\tilde{Y} \in \mathscr{L}_0(X,Y)} \mathcal{D}(P_{\tilde{X}\tilde{Y}} \| P_{\bar{X}\bar{Y}}) \coloneqq E(R = 0) \ , \ \ \forall\, 0 < \epsilon < 1 \ , \tag{4.26}$$

where $\mathscr{L}_0(X,Y) \coloneqq \left\{ \tilde{X}\tilde{Y} : P_{\tilde{X}} = P_X \ , \ P_{\tilde{Y}} = P_Y \right\}$.

Proof. The proof is given in Appendix C.5. □

It is worth mentioning that the same expression (4.26) was proven in [9] to be feasible based on *unidirectional one bit exchange*, i.e., $|f_{[1]}| = 2$, $|g_{[1]}| = 0$. This observation implies that when zero-rate is enforced, not only does interactive data exchanges not help, but only one bit of unidirectional exchange is enough. In addition, note that this is a *strong unfeasability* result, as the optimal exponent for β_n is not dependent on the constraint ϵ over the error probability of Type I.

4.7 Closing Remarks

In this chapter we focused on the problem of binary hypothesis testing over a bidirectional link. Much like the case of a unidirectional link discussed in the previous chapter, problems of this type can arise in cases user authentication is necessary, for example. In fact, the bidirectional link scenario adds difficulty for a potential imposter, as it would have to continue to come up with messages that correspond to jointly typical sequences at every round of communication. Moreover, this scenario may constitute a theoretical base for cases where automatic decision-making is needed in real-time and with few resources, as may be the case is IoT or smart-home applications.

When considering the general case, the approach proposed by [9] was extended to include cooperative communication, first over one round, and then to any finite number of communication rounds. The result achieved through this extension showed that in fact, by allowing cooperative communication we do not change the basic form of the error exponent, but at the same time allow ourselves to "play" with the Markovian relations between the auxiliary RVs, in a way that may improve performance.

The special case of testing against independence was revisited in this chapter. Although this case was already investigated in [20,51], some holes in the theory remained to

be filled. This was accomplished in this chapter by first reestablishing the achievability result through our result for general hypotheses, and then proving a *weak converse* for this special case when one round of communication is permitted. Unfortunately, while the achievability result can be extended to multiple rounds of communication, it was shown that extending the converse is a non-trivial task. This is unsurprising in some sense, as after the first round of communication, the *total amount of information* that is present at both sides is *no longer independent*, even under H_1.

Finally, a *strong converse* property was demonstrated for the case of interactive communication with zero rate. Note that while assuming zero rate prevents the codebooks to grow exponentially with the number of realizations n, it *does not mean* that no communication is allowed between the nodes. This strong converse turned out to be compatible with an achievable scheme that only allows unidirectional communication of one bit, which implies that under a zero-rate constraint only one bit is necessary in order to achieve the optimal result.

Chapter 5

Conclusions and Outlook

5.1 Concluding Remarks

In this work, the performance of distributed systems was studied, under different communication constraints, for different tasks in hypothesis testing. Focusing on a two-node distributed model with binary HT, a model which was used in literature extensively (see e.g., [8,9,20] and more), the communication constraints imposed on the system, as well as the required task, differed throughout the thesis. While different methods were used for each case specifically, some tools rose as universally indispensable in this subject. These were mainly tools from the method of types [21], which allow a type-by-type analysis of performance, and may prove to be more precise in some cases then methods of typicality.

The first problem we focused on, in Chapter 3, consisted of the joint problem of distributed detection and lossy compression with side information. This scenario arises when an authentication system prevents the unauthorized injection of messages into a public channel, assuring the receiver of a message of the legitimacy of its sender. In this setup a user (referred to as node A) is required to communicate a lossy description of a memoryless source to a statistician (referred to as node B) whose task is to verify that the encoding user is the individual he claims to be and then, according to its identity, to reconstruct the message based on the adequate distortion measure, much like in [46, 47]. However, in the setup considered here the receiver is unaware of the value of its information as well, which leads to a two-step approach where first a decision has to be made about the identity of node A, before source reconstruction can take place.

When testing against independence, this two-step approach turns out to be optimal. In this case, detection can be performed optimally as in [8], while source remonstration is performed à la Wyner-Ziv [43], and the two-step approach does not induce performance degradation. An application example to a binary symmetric source was also shown for which the optimal region was explicitly derived, emphasizing an interesting tension between the error exponent corresponding to the Type II error probability and the average distortion measure.

When testing with general hypotheses, a similar, albeit more involved, approach produced a new achievable rate-error-distortion region. Here, optimality may be hard to reach, as optimality results remain allusive even in the case where the receiver is aware of the value of the side information (see [42] and references therein). Nevertheless, we showed that the two-step approach, which was optimal in the case of testing against independence, induces in the general case a significant loss in performance. It was shown that when source reconstruction is not required, valuable information for testing can be compressed much further than in the opposite case, improving significantly the performance of detection.

The second problem, presented and analyzed in Chapter 4, focuses on tasks in HT only, while allowing interactive communication between the statisticians of the system. This scenario was studied for the special case of testing against independence in [20, 51], although many questions still remained unanswered, even for this special case. The more general case, where both hypotheses can induce any common distribution of the source RVs, was never analyzed, to the best of our knowledge. The interest in this scenario goes beyond problems in source authentication (although these too may benefit from interactive communication) to problems involving automated decision-making. This is especially true when the communication is assumed to be made over an open channel, into which each of the participants can tap in order to broadcast to the other participants. One field in which the theoretic work done in this thesis can constitute a background to interesting practical scenarios is the IoT (see e.g., [22, 23]) –a field that has seen increasing academic interest in recent years.

An achievable error-exponent for the Type II error event, constrained by a fixed constraint for the Type I error probability, was given in Chapter 4, based on methods developed for a unidirectional communication link in [9]. Interestingly, it seems that the cooperative approach allows for gains in terms of degrees-of-freedom (by allowing the users to divide resources over different RVs, which have different characteristics) while not changing the basic form of the expression being minimized. This fact gives hope that interaction can lead to considerable gains in performance in some practical scenarios. The same observation was shown to stay true in a subsequent section of the work, when *any finite* number of communication rounds between the nodes was assumed to be allowed.

Revisiting the special case of testing against independence, the previously developed achievable error exponent of [20] was shown to also be achievable through our result for general hypotheses. This distinction becomes important as we were able to show that this error exponent was in fact optimal, *at least in a weak sense*, when testing against independence is considered over one round of communication. Optimality could not be shown for testing against independence over multiple rounds. The reason for this was explained in this thesis. The fact that the totality of the information available at each node after the first round is not independent between the nodes, even under hypothesis H_1, makes the fact that optimality is hard to achieve in this case very conceivable.

Finally, the special case of interactive hypothesis testing with zero-rate communication was also investigated. A specific error-exponent, showed in [9] to be achievable with

unidirectional one-bit communication was shown to be *strongly optimal*, even with codebooks that may grow (sub-exponentially) with the number of realizations n, as well as bidirectional communication.

Weaknesses: Some basic assumptions of the research presented in this thesis raise questions about its possible applicability to real-world systems. First and foremost, it seems unlikely that the probability distributions implied by both hypotheses would be available to the statisticians. It is far more likely that these distributions would have to be observed and assessed out of observations, or that information about them would be missing. This angle of the problem can conceivably be attacked through methods related to universal source coding (see e.g., [97] and references therein).

Another problematic aspect of the assumptions made in this thesis may be the fact that it is assumed that the probability distribution controlling the RVs does not change throughout the detection process, even when the number of observations n may be very large. Recently, the interesting work in [98] addressed the need to detect possible *changes* in the correct hypothesis, that may occur during the process of collecting the observations. Likewise, [99, 100] consider the case of transient changes, where it is assumed that the system starts and finishes at the same state, and only transits through another one at some point.

The rest of this chapter is dedicated to open problems and prospects of future research, related to the progress made in this thesis.

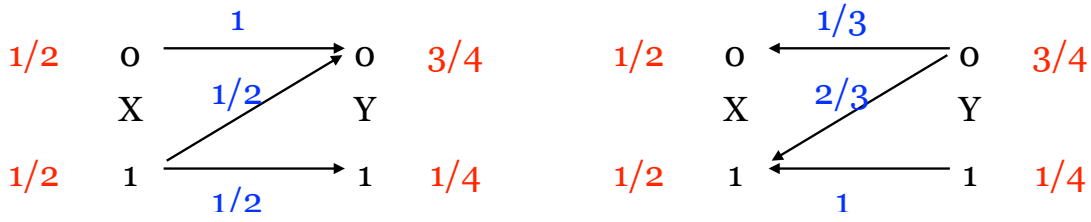
5.2 Outlook

The final remarks of this work are dedicated to a brief outlook of possible future directions of research in the subject of distributed HT with communication constraints.

5.2.1 The Benefit of Interaction when Testing Against Independence

In Chapter 4, a (weakly) optimal error-exponent was given for the error probability of Type II, under a fixed constraint over the error probability of Type I, for testing against independence over one round of interaction. It was shown that the error-exponent that was shown to be achievable in [20] is also achievable through the general error-exponent of Proposition 3. In addition, Theorem 3 proved the (weak) optimality of this result. We now focus on this case and pose the following question –when testing against independence over one round, is interaction beneficial?

In [20], an example is given to the possible benefit in performance, when comparing interactive communication to *unidirectional communication from node A to node B*. This example consists of a common distribution P_{XY} that assumes that X and Y are connected



(a) Z-channel from X to Y (b) Reverse Z-channel from Y to X

Figure 5.1: Two equivalent representations of the Z-channel controlling X and Y under hypothesis 0, in the example of Xiang and Kim. Probabilities are marked in red and transition probabilities are marked in blue.

through a Z-channel, with the following joint probabilities:

$$P_{XY}(0,0) = \frac{1}{2}\ , P_{XY}(0,1) = 0\ , \tag{5.1a}$$

$$P_{XY}(1,0) = \frac{1}{4}\ , P_{XY}(1,1) = \frac{1}{4}\ . \tag{5.1b}$$

A visual description of this channel can be seen in Figure 5.1, both as a Z-channel from X to Y as well as the backwards Z-channel from Y to X. Probabilities are in red and conditional probabilities (or transition probabilities) are in blue. Naturally, it is assumed that hypothesis 1 implies the same marginal probabilities for X and Y, while keeping them independent. The following was shown to be true:

Lemma 18. *Given the test against independence defined above, where a Z-channel is assumed between the sources X and Y under hypothesis 0, implying the joint probability in (5.1), full interaction leads to a gain in performance, in terms of error-exponent of Type II under a fixed constraint over the error probability of Type I, when compared to a unidirectional scenario, from node A to node B. In other words, when computed for the example given above,*

$$\max_{\substack{P_{U|X}:\mathcal{X}\mapsto\mathcal{P}(\mathcal{U}) \\ P_{V|UY}:\mathcal{U}\times\mathcal{Y}\mapsto\mathcal{P}(\mathcal{V}) \\ s.t.\ I(U;X)+I(V;Y|U)\leq R}} \left[I(U;Y)+I(V;X|U)\right] \geq \max_{\substack{P_{W|X}:\mathcal{X}\mapsto\mathcal{P}(\mathcal{W}) \\ s.t.\ I(W;X)\leq R}} \left[I(W;Y)\right]\ . \tag{5.2}$$

Proof. Refer to reference [20]. □

While Lemma 18 is indeed proved in [20], only the expression for unidirectional communication is evaluated. As calculating the expression for bidirectional communication is complicated, the authors choose to use a lower bound, consisting of allocating all resources to the communication of node B. While indeed this shows that bidirectional communication beats the unidirectional option, when node A acts as the transmitter, the question

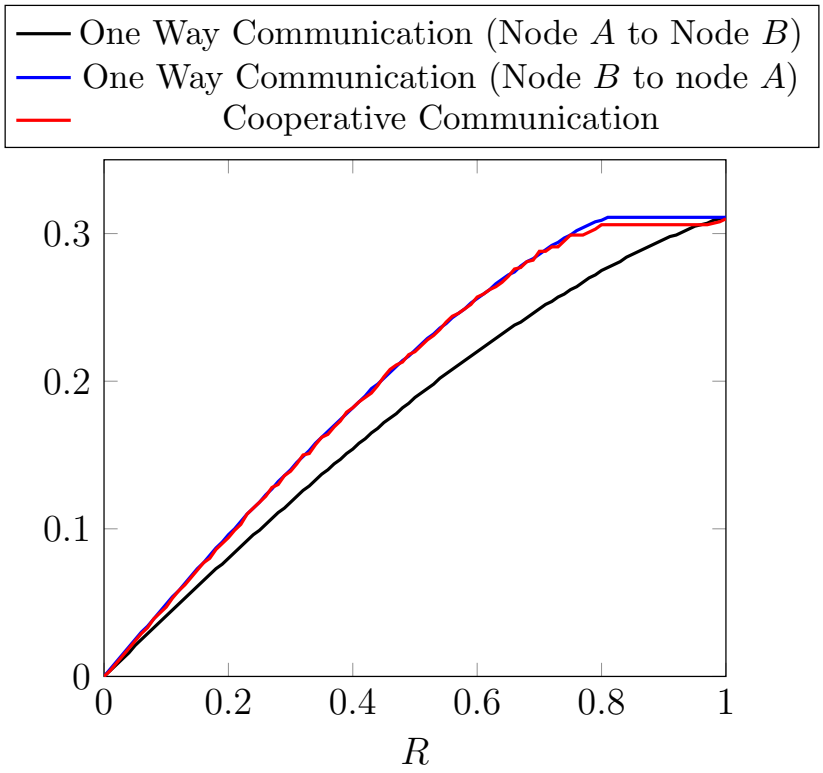


Figure 5.2: Numerical results for the Z-channel example presented by Xiang and Kim, compared with cooperative communication.

that arises is, are there cases in which real cooperative communication (i.e., when some resources are allocated to both sides) beats both unidirectional options?

Taking the same example and calculating the error exponent achieved by cooperative communication numerically, we found that it does not beat, in this case, unidirectional communication from node B to node A. These results can be seen in Figure 5.2. While the results of the numerical calculations are not smooth, as the larger alphabets, which become necessary, offer substantial numerical complexity, it is clear that the performance when "real" interaction is allowed matches the one of unidirectional communication from node B in this case. In other words, when the channel from X to Y is not identical to the reverse channel from Y to X, the benefit in interaction may be the result of the superiority of one of these channels, for the sake of HT against independence, with relation to the other. We feel that this could in fact be a general phenomena, as summarized in the following conjecture:

Conjecture 1. *When testing against independence over a bidirectional channel, cooperative communication does not lead to a gain in performance, in terms of error exponent of Type II, when the error of Type I is constrained to diminish with the number of realizations n ($\alpha_n \to 0$), when compared to a choice between the two possible options of unidirectional*

communication:

$$\begin{aligned} &\max \quad \big[I(U;Y) + I(V;X|U)\big] \\ &P_{U|X} : \mathcal{X} \mapsto \mathcal{P}(\mathcal{U}) \\ &P_{V|UY} : \mathcal{U} \times \mathcal{Y} \mapsto \mathcal{P}(\mathcal{V}) \\ &s.t.\ I(U;X) + I(V;Y|U) \leq R \\ &= \max\left\{ \begin{array}{ll} \max \ \big[I(W;Y)\big], & \max \ \big[I(Z;X)\big] \\ P_{W|X} : \mathcal{X} \mapsto \mathcal{P}(\mathcal{W}) & P_{Z|Y} : \mathcal{Y} \mapsto \mathcal{P}(\mathcal{Z}) \\ s.t.\ I(W;X) \leq R & s.t.\ I(Z;Y) \leq R \end{array} \right\}. \end{aligned} \tag{5.3}$$

Note that the constraint $\alpha_n \to 0$ is to compare both expressions in the setting of a weak converse, as we don't know that the expression for interactive testing against independence is optimal in the strong sense. Intuitively, this conjecture can be explained as follows: It is conceivable that one side may be more apt to sending relevant information to the other, for the sake of testing against independence. In case such directionality exists, there could never be a "reason" to allocate some of the resources to the other side. In other words, the benefit of creating "real" interaction cannot compensate for the choice to allocate some resources to the inferior side. In the case where both sides are equivalent, there is still no gain to be had from dividing the resources between the nodes. This can be seen for the BSS example, as was treated in the case of testing against independence and unidirectional communication in Section 3.2.3, in Figure 5.3.

We propose this conjecture for proof (or a counter-example) in future work on the subject. Such a conclusion could lead to great simplification in real-life systems.

5.2.2 Strong Converse for Interactive Hypothesis Testing

In [8], a strong property is proven for distributed hypothesis testing over a unidirectional link, for any two hypotheses. Thus, while we do not have a single-letter expression for the optimal error-exponent for the error event of Type II, we do know that such an optimal exponent cannot be dependent on the constraint enforced upon the error probability of Type I. We propose a similar approach as a direction for future research. While the interest in such a result is naturally quite clear, we try to explain in this section why proving a strong property for testing with interactive communication is a formidable task.

The proof of [8] relies heavily on the blow-up lemma (Lemma 25, see Appendix C.5). The main idea is that given a strategy such that $\beta_n \leq \exp\{-nE\}$ and $\alpha_n \leq \epsilon$ is achieved for n large enough, a new strategy can be devised such that $\beta_n \leq \exp\{-nE\}$ and $\alpha_n \to 0$. Thus, while the actual expression for the optimal error exponent is not known, it is clear it cannot depend on the constraint ϵ.

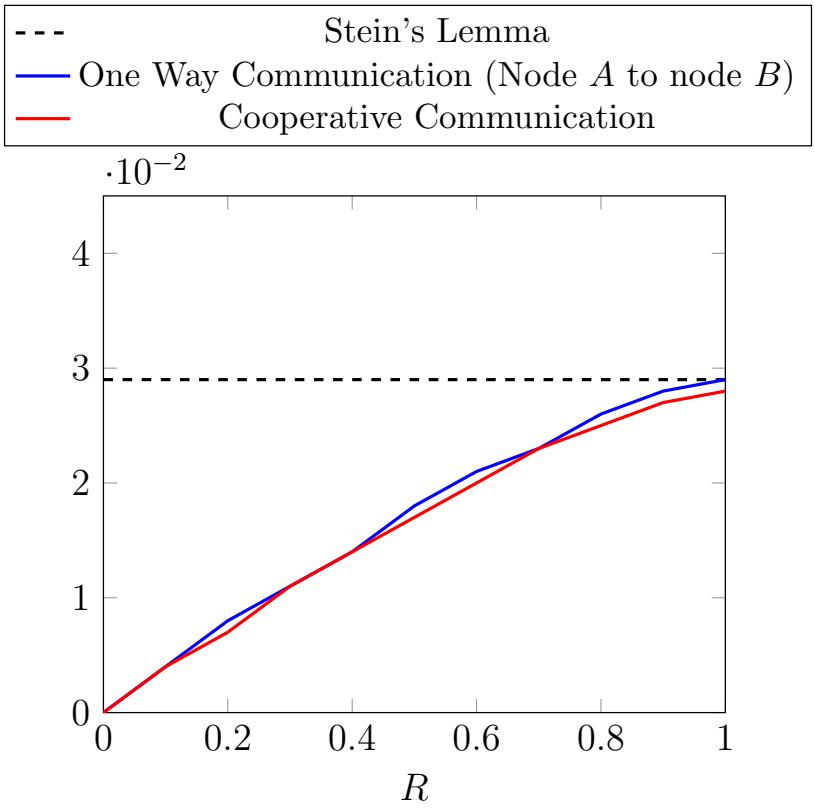


Figure 5.3: Numerical results for the BSC example with transition probability $p = 0.4$, comparing unidirectional and cooperative communication.

Trying to implement a similar approach for a bidirectional channel (even when testing over one round of communication) raises some major difficulties. Here, three types of sets need to be defined in the following manner:

$$\mathcal{A}_i^{(n)} = \{\mathbf{x}^n \in \mathcal{X}^n : f(\mathbf{x}) = i\} \ , \tag{5.4a}$$

$$\mathcal{B}_{ij}^{(n)} = \{\mathbf{y}^n \in \mathcal{Y}^n : g(i, \mathbf{y}) = j\} \ , \tag{5.4b}$$

$$\mathcal{C}_j^{(n)} = \{\mathbf{x}^n \in \mathcal{X}^n : \phi(j, \mathbf{x}) = 0\} \ . \tag{5.4c}$$

Here, $f(\cdot) : \mathcal{X}^n \to \{1, 2, \ldots, \exp\{nR_A\}\}$ is the encoding function of node A, taking the vector $\mathbf{x}$ to message i, $g(\cdot, \cdot) : \{1, 2, \ldots, \exp\{nR_A\}\} \times \mathcal{Y}^n \to \{1, 2, \ldots, \exp\{nR_B\}\}$ is the encoding function at node B, taking the received message i and the vector $\mathbf{y}$ to a message j, and $\phi(\cdot, \cdot) : \{1, 2, \ldots, \exp\{nR_B\}\} \times \mathcal{X}^n \to H_0/H_1$ is the decoding function, taking the received message and the vector $\mathbf{x}$ to a final decision. The desired property now could be referred to as a "conditional" blow-up lemma. Unfortunately, this property does not exist, as summarized in the following lemma:

Lemma 19. *Let $P_{X|Y} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ be any given conditional probability distribution on finite sets $\mathcal{X}$ and $\mathcal{Y}$. Assume that*

$$\Pr(\mathcal{C}_n | X^n \in \mathcal{A}_n, Y^n = \mathbf{y}) \geq \exp(-n\epsilon_n) \ , \tag{5.5}$$

for each n, and some sets $\mathcal{A}_n, \mathcal{C}_n \subseteq \mathcal{X}^n$ and for some sequence $\epsilon_n \to 0$. It cannot be said in general *that*

$$\Pr(\Gamma^{l_n}\mathcal{C}_n | X^n \in \mathcal{A}_n, Y^n = \mathbf{y}) \geq \eta_n \ , \tag{5.6}$$

for some pair of sequences such that $l_n = o(n)$ and $\eta_n \to 1$ as $n \to \infty$, and the operator Γ^{l_n} as defined in Lemma 25.

Proof. We need to show a counter-example. Neglecting $\mathbf{y}$, as it is of no consequence to the result, let $\mathcal{X} = \{0, 1\}$ with $P_X(0) = P_X(1) = \frac{1}{2}$. Let the sets $\mathcal{A}_n$ and $\mathcal{C}_n$ be defined as follows:

$$\begin{aligned} \mathcal{A}_n &= \left\{x^n : x_1 = x_2 = \cdots = x_{\lfloor \frac{n}{2} \rfloor} = 0 \text{ or } x_1 = x_2 = \cdots = x_{\lfloor \frac{n}{2} \rfloor} = 1\right\} \ , \\ \mathcal{C}_n &= \left\{x^n : x_1 = x_2 = \cdots = x_{\lfloor \frac{n}{2} \rfloor} = 0\right\} \ . \end{aligned} \tag{5.7}$$

That is, $\mathcal{A}_n$ is the set of all the vectors such that the first $\lfloor \frac{n}{2} \rfloor$ entries are identical (either 0 or 1), while $\mathcal{C}_n$ is the set of all vectors such that the first $\lfloor \frac{n}{2} \rfloor$ entries are equal to 0. Clearly, while both $P_X^n(\mathcal{A}_n) = \left(\frac{1}{2}\right)^{\lfloor \frac{n}{2} \rfloor - 1}$ and $P_X^n(\mathcal{C}_n) = \left(\frac{1}{2}\right)^{\lfloor \frac{n}{2} \rfloor}$ approach 0 exponentially, $P_X^n(\mathcal{C}_n | \mathcal{A}_n) = \frac{1}{2}$ does not.

Considering the set $\Gamma^{l_n}\mathcal{C}$, it can be defined as follows:

$$\Gamma^{l_n}\mathcal{C}_n = \left\{x^n : x_1 = x_2 = \cdots = x_{\lfloor \frac{n}{2} \rfloor} = 0 \text{ except for up to } l_n \text{ times}\right\} \ . \tag{5.8}$$

In order for a sequence $\gamma_n \to 0$ to exist, such that $P_X^n(\Gamma^{l_n}\mathcal{C}_n | \mathcal{A}_n) \geq 1 - \gamma_n$, l_n must at least approach $\frac{n}{2}$, thus not complying with $\frac{l_n}{n} \to 0$. □

While this example shows that the property above is not always true, the number of examples that still turn out to fulfill it is quite amazing. It is straight-forward to show that the desired property is true whenever the probability of the set $\mathcal{A}_n$ can be bounded from below (there exist a constant $\eta > 0$ such that $P_X^n(\mathcal{A}_n) \geq \eta, \forall n$), for example. Other examples also turn out to fulfill this property quite often.

While the existence of a strong property for hypothesis testing with cooperative communication is not *dependent* on the existence of a "conditional blow-up lemma", we do feel that in some sense the blow-up lemma captures the essence of the strong property, and the lack of a conditional version puts this property in doubt for the cooperative case. An interesting direction for future research could be to investigate weather such a property exists in general for HT with interactive communication. In case it does not exist in general, what are the mathematical constraints on the sets $\mathcal{A}_n, \mathcal{C}_n$ such that a "conditional" blow up lemma exists? Are those constraints enough to also prove a strong property in HT?

5.2.3 Other directions

Except for the two directions proposed above, research in distributed HT can advance in many other directions. Clearly, the question of optimality stays open in the general case,

for any type of communication constraint. Even the relatively simple case of unidirectional communication is open –not only in terms of optimality, but for any non-trivial upper bound. As the work in this thesis shows that finally, the expressions in the bidirectional case are similar to the ones of the unidirectional case (at least in terms of an achievable error exponent), this gives hope that any method that would work for one case could be adapted to the other.

Another direction, proposed briefly in Chapter 4, is the exploration of random binning with the goal of improving the Type II error exponent in the bidirectional case. This direction seems to be quite clearly one that could improve the proposed error exponent of Proposition 3. However, as the binning approach leads to a trade-off between two error events, and this for each step of the cooperative communication, the resulting expression risks being highly non-compact and thus contributing little to the general understanding of the problem. Additionally, as the binning approach takes advantage of the randomness of the side information, available at the decoder, in order to improve performance, we believe that the benefit of applying it would decrease with every additional round of communication, as each side "learns" about the realizations seen by the other side.

Finally, finding ways to calculate the formulas discussed throughout this thesis would be very interesting. As in most cases the achievable error exponent is the result of a *minimum operator* taken over some set, this task could pose quite a challenge. Specifically –how to evaluate the correct error exponent while making sure it is not surpassed? Possible approaches to this question could potentially be found in works on iterative algorithms such as the Blahut-Arimoto algorithm ([101, 102], see also the connection with the information bottleneck problem in the introduction).

Bibliography

[1] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, ser. Springer Texts in Statistics. Springer, 2005.

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

[3] V. Y. F. Tan, "Stein's lemma," in *Information Theory for Communication Systems, EE5139R Lecture 12*, November 2015. [Online]. Available: https://www.ece.nus.edu.sg/stfpage/vtan/ee5139/

[4] R. Tenney and N. R. Sandell, "Detection with distributed sensors," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-17, no. 4, pp. 501–510, July 1981.

[5] T. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.

[6] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[7] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.

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

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

[10] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall, 1971.

[11] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin, "Wireless information-theoretic security," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2515–2534, June 2008.

[12] U. M. Maurer, "Protocols for secret key agreement by public discussion based on common information," in *Annual International Cryptology Conference*. Springer, 1992, pp. 461–470.

[13] ——, "Secret key agreement by public discussion from common information," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 733–742, May 1993.

[14] C. H. Bennett, G. Brassard, and J.-M. Robert, "Privacy amplification by public discussion," *SIAM journal on Computing*, vol. 17, no. 2, pp. 210–229, 1988.

[15] W. Diffie and M. Hellman, "New directions in cryptography," *Information Theory, IEEE Transactions on*, vol. 22, no. 6, pp. 644–654, Nov 1976.

[16] Y.-C. Lin, D. Varodayan, and B. Girod, "Image authentication using distributed source coding," *Image Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 273–283, Jan 2012.

[17] C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing jpeg compression from malicious manipulation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 2, pp. 153–168, Feb 2001.

[18] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in ac state estimation," *Smart Grid, IEEE Transactions on*, vol. 6, no. 5, pp. 2476–2483, Sept 2015.

[19] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks," *Smart Grid, IEEE Transactions on*, vol. 4, no. 3, pp. 1244–1253, Sept 2013.

[20] Y. Xiang and Y.-H. Kim, "Interactive hypothesis testing with communication constraints," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, Oct 2012, pp. 1065–1072.

[21] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct 1998.

[22] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[23] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[24] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future internet: The internet of things architecture, possible applications and key challenges," in *Frontiers of Information Technology (FIT), 2012 10th International Conference on*, Dec 2012, pp. 257–260.

[25] N. Li, M. Sun, Z. Bi, Z. Su, and C. Wang, "A new methodology to support group decision-making for iot-based emergency response systems," *Information Systems Frontiers*, vol. 16, no. 5, pp. 953–977, 2014.

[26] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.

[27] P. Bonnet, J. Gehrke, and P. Seshadri, "Querying the physical world," *IEEE Personal Communications*, vol. 7, no. 5, pp. 10–15, Oct 2000.

[28] A. Chandrakasan, R. Amirtharajah, S. Cho, J. Goodman, G. Konduri, J. Kulik, W. Rabiner, and A. Wang, "Design considerations for distributed microsensor systems," in *Custom Integrated Circuits, 1999. Proceedings of the IEEE 1999*, 1999, pp. 279–286.

[29] I. A. Essa, "Ubiquitous sensing for smart and aware environments," *IEEE personal communications*, vol. 7, no. 5, pp. 47–49, 2000.

[30] T. M. Cover, "Hypothesis testing with finite statistics," *The Annals of Mathematical Statistics*, vol. 40, no. 3, pp. 828–835, 1969.

[31] M. E. Hellman and T. M. Cover, "Learning with finite memory," *The Annals of Mathematical Statistics*, vol. 41, no. 3, pp. 765–782, 1970.

[32] S. Yakowitz, "Multiple hypothesis testing by finite memory algorithms," *The Annals of Statistics*, vol. 2, no. 2, pp. 323–336, 1974.

[33] J. Bucklew and P. Ney, "Asymptotically optimal hypothesis testing with memory constraints," *The Annals of Statistics*, vol. 18, no. 2, pp. 982–998, 1991.

[34] T. Chiyonobu, "Hypothesis testing for signal detection problem and large deviations," *Nagoya Mathematical Journal*, vol. 162, pp. 187–203, 2001.

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

[36] R. Ahlswede and M. Burnashev, "On minimax estimation in the presence of side information about remote data," *The Annals of Statistics*, vol. 18, no. 1, pp. 141–171, 1990.

[37] H. Shimokawa, T. Han, and S.-I. Amari, "Error bound of hypothesis testing with data compression," in *Inf. Theory, 1994 IEEE International Symposium on (ISIT)*, Jun 1994, p. 114.

[38] 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.

[39] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2148–2177, Oct 1998.

[40] H. Shalaby and A. Papamarcou, "Multiterminal detection with zero-rate data compression," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 254–267, March 1992.

[41] W. Zhao and L. Lai, "Distributed testing with zero-rate compression," in *Inf. Theory, 2015 IEEE International Symposium on (ISIT)*, June 2015, pp. 2792–2796.

[42] Y. Steinberg and N. Merhav, "On successive refinement for the wyner-ziv problem," *Information Theory, IEEE Transactions on*, vol. 50, no. 8, pp. 1636–1654, Aug 2004.

[43] 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.

[44] F.-W. Fu and R. W. Yeung, "On the rate-distortion region for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 2012–2021, Jul 2002.

[45] R. Timo, T. Chan, and A. Grant, "Rate distortion with side-information at many decoders," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5240 –5257, Aug. 2011.

[46] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 727–734, Nov 1985.

[47] A. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, pp. 2031–2034, Nov 1994.

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

[49] ——, "Remote vector gaussian source coding with decoder side information under mutual information and distortion constraints," *Information Theory, IEEE Transactions on*, vol. 55, no. 10, pp. 4676–4680, Oct 2009.

[50] A. Kaspi, "Two-way source coding with a fidelity criterion," *Information Theory, IEEE Transactions on*, vol. 31, no. 6, pp. 735–740, Nov 1985.

[51] Y. Xiang and Y.-H. Kim, "Interactive hypothesis testing against independence," in *Inf. Theory, 2013 IEEE International Symposium on (ISIT)*, July 2013, pp. 2840–2844.

[52] S. Bayram and S. Gezici, "Noise-enhanced m-ary hypothesis-testing in the minimax framework," in *Signal Processing and Communication Systems, 2009. ICSPCS 2009. 3rd International Conference on*, Sept 2009, pp. 1–6.

[53] M. Naghshvar and T. Javidi, "Active m-ary sequential hypothesis testing," in *2010 IEEE International Symposium on Information Theory*, June 2010, pp. 1623–1627.

[54] Z. B. Tang, K. R. Pattipati, and D. L. Kleinman, "A distributed m-ary hypothesis testing problem with correlated observations," in *Decision and Control, 1989., Proceedings of the 28th IEEE Conference on*, Dec 1989, pp. 562–568 vol.1.

[55] X. Zhu, Y. Yuan, C. Rorres, and M. Kam, "Distributed m-ary hypothesis testing with binary local decisions," *Information Fusion*, vol. 5, no. 3, pp. 157–167, 2004.

[56] G. Vazquez-Vilar, A. T. Campo, A. G. i Fàbregas, and A. Martinez, "Bayesian m -ary hypothesis testing: The meta-converse and verdù-han bounds are tight," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2324–2333, May 2016.

[57] P. Moulin, "Asymptotically achievable error probabilities for multiple hypothesis testing," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1541–1545.

[58] C. C. Leang and D. H. Johnson, "On the asymptotics of m-hypothesis bayesian detection," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 280–282, Jan 1997.

[59] E. Tuncel, "On error exponents in hypothesis testing," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2945–2950, Aug 2005.

[60] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *The Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, 2013.

[61] M. Nussbaum and A. Szkoła, "The chernoff lower bound for symmetric quantum hypothesis testing," *The Annals of Statistics*, vol. 37, no. 2, pp. 1040–1057, 2009.

[62] K. Audenaert, M. Nussbaum, A. Szkoła, and F. Verstraete, "Asymptotic error rates in quantum hypothesis testing," *Communications in Mathematical Physics*, vol. 279, no. 1, pp. 251–283, 2008.

[63] T. Ogawa and H. Nagaoka, "Strong converse and stein's lemma in quantum hypothesis testing," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2428–2433, 2000.

[64] S. Zhu and B. Chen, "Distributed detection over connected networks via one-bit quantizer," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1526–1530.

[65] 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.

[66] 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.

[67] 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.

[68] J. Goldberger, H. Greenspan, and S. Gordon, "Unsupervised image clustering using the information bottleneck method," in *Joint Pattern Recognition Symposium*. Springer, 2002, pp. 158–165.

[69] A. Bardera, J. Rigau, I. Boada, M. Feixas, and M. Sbert, "Image segmentation using information bottleneck method," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1601–1612, July 2009.

[70] S. Chiappino, L. Marcenaro, and C. S. Regazzoni, "Information bottleneck-based relevant knowledge representation in large-scale video surveillance systems," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4364–4368.

[71] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research*, vol. 1, 2001, p. 200.

[72] ——, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 208–215.

[73] M. Wang, Y. He, and M. Jiang, "Text categorization of enron email corpus based on information bottleneck and maximal entropy," in *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, Oct 2010, pp. 2472–2475.

[74] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Information Theory Workshop (ITW), 2015 IEEE*, April 2015, pp. 1–5.

[75] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[76] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 129–136.

[77] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.* IEEE, 2007, pp. 250–255.

[78] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 34, no. 2, pp. 157–177, 1976.

[79] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[80] R. Fano, "Class notes for transmission of information, course 6.574," *Massachusetts Institute of Technology, Tech. Rep*, 1952.

[81] J. Wolfowitz *et al.*, "The coding of messages subject to chance errors," *Illinois Journal of Mathematics*, vol. 1, no. 4, pp. 591–606, 1957.

[82] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[83] J. C. Kieffer, "Strong converses in source coding relative to a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 257–262, 1991.

[84] ——, "Sample converses in source coding theory," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 263–268, 1991.

[85] G. Dueck, "The strong converse to the coding theorem for the multiple–access channel," *J. Comb. Inform. Syst. Sci*, vol. 6, no. 3, pp. 187–196, 1981.

[86] B. Kelly and A. Wagner, "Reliability in source coding with side information," *Information Theory, IEEE Transactions on*, vol. 58, no. 8, pp. 5086–5111, Aug 2012.

[87] G. Katz, P. Piantanida, and M. Debbah, "Distributed Binary Detection with Lossy Data Compression," *ArXiv e-prints*, Jan. 2016, Submitted to Information Theory, IEEE Trans. on. [Online]. Available: http://arxiv.org/abs/1601.01152

[88] ——, "Collaborative distributed hypothesis testing," *ArXiv e-prints*, Apr. 2016, submitted to The Annals of Statistics. [Online]. Available: http://arxiv.org/abs/1604.01292

[89] G. Katz, P. Piantanida, R. Couillet, and M. Debbah, "Joint estimation and detection against independence," in *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, 2014.

[90] ——, "On the necessity of binning for the distributed hypothesis testing problem," in *Inf. Theory, 2015 IEEE International Symposium on (ISIT)*, June 2015, pp. 2797–2801.

[91] G. Katz, P. Piantanida, and M. Debbah, "Collaborative distributed hypothesis testing with general hypotheses," in *Inf. Theory, 2016 IEEE International Symposium on (ISIT)*, July 2016.

[92] ——, "A new approach to distributed hypothesis testing," in *50th Asilomar Conf. on Signals, Systems and Computers*, November 2016, to be presented.

[93] P. Piantanida, L. Rey Vega, and A. Hero, "A proof of the generalized markov lemma with countable infinite sources," in *Information Theory Proceedings (ISIT), 2014 IEEE International Symposium on*, July 2014.

[94] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629–637, Nov 1975.

[95] S.-W. Ho and S. Verdu, "On the interplay between conditional entropy and error probability," *Information Theory, IEEE Transactions on*, vol. 56, no. 12, pp. 5930–5942, Dec 2010.

[96] L. R. Vega, P. Piantanida, and A. O. Hero, "The three-terminal interactive lossy source coding problem," *Information Theory, IEEE Trans. on*, 2015, (revised). [Online]. Available: http://arxiv.org/abs/1502.01359

[97] T. Linder, G. Lugosi, and K. Zeger, "Fixed-rate universal lossy source coding and rates of convergence for memoryless sources," *Information Theory, IEEE Trans. on*, vol. 41, no. 3, pp. 665–676, May 1995.

[98] G. V. Moustakides and V. V. Veeravalli, "Sequentially detecting transitory changes," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 26–30.

[99] B. K. Guépié, L. Fillatre, and I. Nikiforov, "Sequential detection of transient changes," *Sequential Analysis*, vol. 31, no. 4, pp. 528–547, 2012.

[100] E. Ebrahimzadeh and A. Tchamkerten, "Sequential detection of transient changes in stochastic systems under a sampling constraint," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 156–160.

[101] P. O. Vontobel, "A generalized blahut-arimoto algorithm," in *Information Theory, 2003. Proceedings. IEEE International Symposium on.* IEEE, 2003, p. 53.

[102] H. H. Permuter and I. Naiss, "Extension of the blahut-arimoto algorithm for maximizing directed information," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on.* IEEE, 2010, pp. 1442–1449.

[103] I. Sanov, *On the probability of large deviations of random variables.* United States Air Force, Office of Scientific Research, 1958.

[104] 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.

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

[106] A. Schrijver, *Theory of linear and integer programming.* John Wiley & Sons, 1998.

[107] G. A. Margulis, "Probabilistic characteristics of graphs with large connectivity," *Problemy Peredači Informacii*, vol. 10, no. 2, pp. 101–108, 1974.

Appendix A

Useful Results

A.1 Proof of Lemma 11 (Stein's Lemma)

While many different proofs of Stein's Lemma exist in literature (for example in [1,2]), the proof we choose to present here has the advantage of demonstrating the *strong property* of Stein's Lemma. This proof, while common in tutorials on the subject (see e.g., [3]), is hard to come by in an official publication, to the best of our knowledge.

A.1.1 Proof of Achievability

Consider the decision regions

$$\mathcal{B}_n = \{\mathbf{x} : \mathcal{D}(Q_{\mathbf{x}}||P_0) \leq \delta_n\} \ , \tag{A.1}$$

where $Q_{\mathbf{X}}$ is the type of the vector $\mathbf{x}$ and $\delta_n = \frac{1}{\sqrt{n}}$. In this case we can calculate

$$\alpha_n(\mathcal{B}_n) = P_0^n(\mathcal{B}_n^c) = P_0^n(\{\mathbf{x} : \mathcal{D}(Q_{\mathbf{x}}||P_0) > \delta_n\}) = \sum_{\mathbf{x}:\mathcal{D}(Q_{\mathbf{x}}||P_0)>\delta_n} P_0^n(\mathbf{x}) \tag{A.2a}$$

$$= \sum_{\hat{P}\in\mathcal{P}_n(\mathcal{X}):\mathcal{D}(\hat{P}||P_0)>\delta_n} \sum_{\mathbf{x}\in\mathcal{T}_{[\hat{P}]}} P_0^n(\mathbf{x}) = \sum_{\hat{P}\in\mathcal{P}_n(\mathcal{X}):\mathcal{D}(\hat{P}||P_0)>\delta_n} P_0^n(\mathcal{T}_{[\hat{P}]}) \tag{A.2b}$$

$$\leq \sum_{\hat{P}\in\mathcal{P}_n(\mathcal{X}):\mathcal{D}(\hat{P}||P_0)>\delta_n} \exp\{-n\mathcal{D}(\hat{P}||P_0)\} \leq \sum_{\hat{P}\in\mathcal{P}_n(\mathcal{X}):\mathcal{D}(\hat{P}||P_0)>\delta_n} \exp\{-n\delta_n\} \tag{A.2c}$$

$$= \sum_{\hat{P}\in\mathcal{P}_n(\mathcal{X}):\mathcal{D}(\hat{P}||P_0)>\delta_n} \exp\{-\sqrt{n}\} \leq (n+1)^{|\mathcal{X}|}\exp\{-\sqrt{n}\} \leq \epsilon \ . \tag{A.2d}$$

Here, (A.2b) is the result of changing the order of summation to be over types first and over sequences within each type second. (A.2c) is due to Lemma 6 and then to the definition of the chosen acceptance sets $\mathcal{B}_n$. (A.2d) stems from the choice of δ_n and the bound over the number of possible types of length n (see Lemma 4). Finally, the resulting expression is lower than ϵ when n is large enough.

Estimating the error of Type II, we have

$$\beta_n(\mathcal{B}_n) = P_1^n(\mathcal{B}_n) \leq (n+1)^{|\mathcal{X}|} \exp\{-n\mathcal{D}(P^\star||P_1)\} \tag{A.3}$$

by Sanov's Theorem (see e.g., [2, 103]), with

$$P^\star = \underset{\hat{P}:\mathcal{D}(\hat{P}||P_0)\leq\delta_n}{\operatorname{argmin}} \mathcal{D}(\hat{P}||P_1) \ . \tag{A.4}$$

As $\delta_n \to 0$ and $\mathcal{D}(\hat{P}||P_0) = 0$ if and only if (iff) $\hat{P} = P_0$, the optimizer $P^\star$ converges to P_0, i.e.,

$$P^\star(a) \to P_0(a) \ , \forall a \in \mathcal{X} \ . \tag{A.5}$$

Hence,

$$\beta_n(\mathcal{B}_n) \leq (n+1)^{|\mathcal{X}|} \exp\{-n(\mathcal{D}(P_0||P_1) + o(1))\} \ , \tag{A.6}$$

which completes the proof of achievability.

A.1.2 Proof of Converse

We now prove that for any choice of strategy,

$$\beta_n(\epsilon) \geq \exp\{-n\mathcal{D}(P_0||P_1)\} \ . \tag{A.7}$$

Note that this exponent does not depend on the constraint over the Type I error probability, ϵ. To do so, we first establish the following lemma:

Lemma 20. *Let $\mathcal{D}_n \subset \mathcal{X}^n$ be a subset satisfying*

$$P_0^n(\mathcal{D}_n) > 1 - \epsilon \ , \tag{A.8}$$

where $\epsilon \in (0,1)$. Then for any $0 < \delta < 1 - \epsilon$ and n large enough, we have

$$P_1^n(\mathcal{D}_n) > (1 - \epsilon - \delta) \exp\{-n(\mathcal{D}(P_0||P_1) + \delta)\} \ . \tag{A.9}$$

Proof. Fix $\delta \in (0, 1-\epsilon)$, Define the relative entropy typical set:

$$\mathcal{E}_n = \left\{\mathbf{x} : -\delta \leq \frac{1}{n} \log \frac{P_0^n(\mathbf{x})}{P_1^n(\mathbf{x})} - \mathcal{D}(P_0||P_1) \leq \delta\right\} \ . \tag{A.10}$$

By the weak law of large numbers, $P_0^n(\mathcal{E}_n) > 1 - \delta$ for all n large enough. Furthermore, by the union bound

$$P_0^n(\mathcal{E}_n^c \cup \mathcal{D}_n^c) \leq P_0^n(\mathcal{E}_n^c) + P_0^n(\mathcal{D}_n^c) \leq \delta + \epsilon \ , \tag{A.11}$$

so

$$P_0^n(\mathcal{E}_n \cap \mathcal{D}_n) \geq 1 - (\delta + \epsilon) \ . \tag{A.12}$$

Now consider

$$P_1^n(\mathcal{D}_n) \geq P_1^n(\mathcal{D}_n \cap \mathcal{E}_n) \geq \sum_{\mathbf{x} \in \mathcal{D}_n \cap \mathcal{E}_n} P_1^n(\mathbf{x}) \tag{A.13a}$$

$$\geq \sum_{\mathbf{x} \in \mathcal{D}_n \cap \mathcal{E}_n} P_0^n(\mathbf{x}) \exp\{-n(\mathcal{D}(P_0||P_1) + \delta)\} \tag{A.13b}$$

$$= \exp\{-n(\mathcal{D}(P_0||P_1) + \delta)\} P_0^n(\mathcal{D}_n \cap \mathcal{E}_n) \tag{A.13c}$$

$$\geq (1 - \delta - \epsilon) \exp\{-n(\mathcal{D}(P_0||P_1) + \delta)\} \ , \tag{A.13d}$$

where (A.13b) uses the definition of the relative entropy typical set $\mathcal{E}_n$, and (A.13d) stems from (A.12). □

Having proven Lemma 20, the proof of a strong converse to Stein's Lemma becomes quite simple: For any "legal" acceptance region $\mathcal{A}_n$, we have the property $P_0^n \mathcal{A}_n \leq \epsilon$. Thus, using $\mathcal{D}_n = \mathcal{A}_n^c$ in the premises of Lemma 20, we get immediately that

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n(\epsilon) \geq -\mathcal{D}(P_0||P_1) - \delta + \lim_{n \to \infty} \log(1 - \epsilon - \delta) \ . \tag{A.14}$$

As the final term obviously vanishes to 0, along with the fact that this claim is true for any $\delta \in (0, 1 - \epsilon)$, this proves that

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n(\epsilon) \geq -\mathcal{D}(P_0||P_1) \tag{A.15}$$

as desired.

Appendix B

Hypothesis Testing with Unidirectional Communication

B.1 Proof of Theorem 1

In this appendix, we prove the achievability and converse to Theorem 1.

B.1.1 Proof of Achievability

In order to prove achievability, we propose an encoding and decoding strategy, based on a two-tier approach. We then analyze the probabilities of error in detection, as well as the average distortion, in order to guarantee that the performance implied by Theorem 1 is achieved.

Codebook generation: Fix a conditional probability distribution $Q_{VU|XY} = Q_{V|UX}Q_{U|X}P_{XY}$ such that $U - V - X - Y$ form a Markov chain. Let $Q_U(u) = \sum_{x\in\mathcal{X}} P_X(x)Q_{U|X}(u|x)$ and $Q_{V|U}(v|u) = \sum_{x\in\mathcal{X}} Q_{V|UX}(v|u,x)$. Let the total available rate of communication R be divided into two, such that the parts are dedicated to U and V, which represent the different parts of the message. Denote the rate dedicated to the transmission of U by $\hat{R}$, while the rate dedicated to the transmission of V is denoted by R'. Randomly and independently generate $\exp(n\hat{R})$ sequences $\mathbf{u}$ through the i.i.d. pmf $Q_U(u)$, with replacement, such that $\mathbf{u}(s_1) \in \mathcal{T}_{[U]\delta}$, $\forall s_1$, with $s_1 \in [1 : \exp(n\hat{R})]$. For each codeword $\mathbf{u}(s_1)$, randomly and independently generate $\exp(nS_2)$ sequences denoted by $\mathbf{v}^n(s_1, s_2)$ and indexed with $s_2 \in [1 : \exp(nS_2)]$ by using the conditional pmf $Q_{V|U}(\cdot|\mathbf{u}(s_1))$, with replacement, such that $\mathbf{v}(s_1, s_2) \in \mathcal{T}_{[V|U]\delta}(\mathbf{u}(s_1))$. Divide theses sequences into $\exp[nR']$ bins, such that each bin contains roughly $\exp[n(S_2 - R')]$ sequences.

Encoding: Assuming that the source sequence $\mathbf{x}^n$ is produced from X, look for the first codeword in U's codebook such that $(\mathbf{u}^n(s_1), \mathbf{x}^n) \in \mathcal{T}^n_{[UX]\delta}$. Then, look for the first codeword $\mathbf{v}^n(s_1, s_2)$ s.t. $(\mathbf{v}^n(s_1, s_2), \mathbf{x}^n) \in \mathcal{T}^n_{[VX|U]\delta}(\mathbf{u}(s_1))$. 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 $(\mathbf{u}^n(s_1), \mathbf{y}^n) \in \mathcal{T}^n_{[UY]\delta}$. If so, it declares H_0 and otherwise it declares H_1. If the decoder decides H_0, it then attempts to decode the message (with average distortion D) based on $\mathbf{v}(s_1, s_2)$. This codeword is first recovered by looking in the bin b for the unique codeword such that $\mathbf{v}^n(s_1, s_2) \in \mathcal{T}^n_{[V|UY]\delta}(\mathbf{u}(s_1), \mathbf{y}^n)$. Then, a per-letter function $g(\cdot)$ is applied over the entire available information (U, V and Y) in order to produce a reconstruction of the source.

Error events and constraints: We start with the HT part, and the relation between the expression $I(U;X)$ and the achievable error exponent. Denoting by $\mathcal{B}_0$ the event "an error occurred during encoding" (of the HT part U), we expend its probability as $\Pr(\mathcal{B}_0) \leq \Pr(\mathcal{B}_1) + \Pr(\mathcal{B}_2)$ with:

$$\begin{aligned} \Pr(\mathcal{B}_1) &\triangleq \Pr\{\mathbf{X}^n \notin \mathcal{T}^n_{[X]\delta}\} \ , \\ \Pr(\mathcal{B}_2) &\triangleq \Pr\{\nexists s_1 \text{ s.t. } (\mathbf{u}(s_1), \mathbf{X}^n) \in \mathcal{T}^n_{[UX]\delta} | \mathbf{X}^n \in T^n_{[X]\delta}\} \ , \end{aligned} \tag{B.1}$$

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), $\Pr(\mathcal{B}_1) \leq \eta^{(1)}_n \underset{n\to\infty}{\longrightarrow} 0$. As for $\Pr(\mathcal{B}_2)$:

$$\Pr(\mathcal{B}_2) = \left(\Pr\{(\mathbf{U}^n, \mathbf{X}^n) \notin \mathcal{T}^n_{[UX]\delta} | \mathbf{U}^n \in \mathcal{T}^n_{[U]\delta}, \mathbf{X}^n \in \mathcal{T}^n_{[X]\delta}\}\right)^{\exp(n\hat{R})} \tag{B.2a}$$

$$= \left(1 - \Pr\{(\mathbf{U}^n, \mathbf{X}^n) \in \mathcal{T}^n_{[UX]\delta} | \mathbf{U}^n \in \mathcal{T}^n_{[U]\delta}, \mathbf{X}^n \in \mathcal{T}^n_{[X]\delta}\}\right)^{\exp(n\hat{R})} \tag{B.2b}$$

$$\leq \exp[-\exp(n\hat{R})\Pr\{(\mathbf{U}^n, \mathbf{X}^n) \in \mathcal{T}^n_{[UX]\delta} | \mathbf{U}^n \in \mathcal{T}^n_{[U]\delta}, \mathbf{X}^n \in T^n_{[X]\delta}\}] \tag{B.2c}$$

$$\leq \exp[-\exp(n\hat{R})\exp^{-n\left(I(U;X)+\eta^{(2)}_n\right)}] \tag{B.2d}$$

$$= \exp\{-\exp[-n\left(I(U;X) - \hat{R} + \eta^{(2)}_n\right)]\} \ . \tag{B.2e}$$

Here, inequality (B.2c) is due to the inequality $(1-a)^n \leq \exp(an)$ [2]. Since $\eta^{(2)}_n \underset{n\to\infty}{\longrightarrow} 0$, $\Pr(\mathcal{B}_2) \to 0$ if $\hat{R} > I(U;X)$.

Analysis of α_n: Calculating the probability of error of the first type, α_n, boils down to the following:

$$\alpha_n = \Pr(H_1 | XY \sim P_{XY}) \tag{B.3a}$$

$$\leq \Pr(\mathcal{B}_0) + \Pr\{(\mathbf{U}^n, \mathbf{Y}^n) \notin \mathcal{T}^n_{[UY]\delta} | \mathbf{U}^n \in \mathcal{T}^n_{[U]\delta}, (\mathbf{U}^n, \mathbf{X}^n) \in \mathcal{T}^n_{[UX]\delta}, XY \sim P_{XY}\} \tag{B.3b}$$

$$\leq \Pr(\mathcal{B}_0) + \eta^{(3)} \ . \tag{B.3c}$$

Here, (B.3b) is due to the fact that when calculating the probability of error of Type I, we may assume that the true distribution controlling the RVs is the one implied by hypothesis 0. (B.3c), with $\eta^{(3)} \to 0$, is due to the Generalized Markov Lemma (see Lemma 8 in Chapter 2). Thus, it may be concluded that $\alpha_n \to 0$ when $n \to \infty$, and thus $\alpha_n \leq \epsilon$ for any constraint $\epsilon > 0$ and n large enough.

Analysis of β_n: Next, we look at the achievable error exponent of Type II with the proposed encoding scheme, by following steps similar to [104, Lemma 6]:

$$\frac{1}{n} I\left(f(\mathbf{X}^n); \mathbf{Y}^n | \mathcal{C}\right) = \frac{1}{n}\left[H(\mathbf{Y}^n|\mathcal{C}) - H(\mathbf{Y}^n|f(\mathbf{X}^n), \mathcal{C})\right] = H(Y) - \frac{1}{n} H(\mathbf{Y}^n|f(\mathbf{X}^n), \mathcal{C}) \ . \tag{B.4}$$

Here, $\mathcal{C}$ denotes the chosen codebook, which is known to all parties. The second term here can be evaluated by defining the RV

$$\hat{\mathbf{Y}}^n = \begin{cases} \mathbf{Y}^n & \text{if } (\mathbf{u}^n(s_1), \mathbf{Y}^n) \in \mathcal{T}^n_{[UY]\delta} \\ \emptyset & \text{else} \end{cases} \ , \tag{B.5}$$

and writing

$$\begin{aligned}
\frac{1}{n} H(\mathbf{Y}^n|f(\mathbf{X}^n), \mathcal{C}) &\leq \frac{1}{n} H(\mathbf{Y}^n|s_1) && \text{(B.6a)} \\
&= \frac{1}{n} \sum_{j=1}^{\exp(ns_1)} H(\mathbf{Y}^n|s_1 = j)\Pr(s_1 = j) && \text{(B.6b)} \\
&= \frac{1}{n} \sum_{j=1}^{\exp(ns_1)} H(\mathbf{Y}^n\hat{\mathbf{Y}}^n|s_1 = j)\Pr(s_1 = j) && \text{(B.6c)} \\
&= \frac{1}{n} \sum_{j=1}^{\exp(ns_1)} \left(\underbrace{H(\hat{\mathbf{Y}}^n|s_1 = j)}_{(*)} + \underbrace{H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n, s_1 = j)}_{(**)} \right) \Pr(s_1 = j) \ . && \text{(B.6d)}
\end{aligned}$$

Here, the inequality in (B.6a) 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. (B.6c) stems from the fact that $\hat{\mathbf{Y}}$ is a function of $\mathbf{Y}$, and (B.6d) is due to the chain rule. We bound this expression further by treating each part separately:

$$\begin{aligned}
(*) &= \frac{1}{n} \sum_{j=1}^{\exp(nS_1)} H(\hat{\mathbf{Y}}^n|s_1 = j)\Pr(s_1 = j) && \text{(B.7a)} \\
&\leq \frac{1}{n} \sum_{j=1}^{\exp(ns_1)} \log\left(||\mathcal{T}^n_{[Y|U]\delta}(\mathbf{u}^n(j))|| + 1\right) \Pr(s_1 = j) && \text{(B.7b)} \\
&\leq \sum_{j=1}^{\exp(ns_1)} \left(H(Y|U) + \eta_n^{(3)}\right) \Pr(s_1 = j) = H(Y|U) + \eta_n^{(4)} \ , && \text{(B.7c)}
\end{aligned}$$

where (B.7b) is due to the fact that uniform distribution maximizes entropy and (B.7c) stems from bounding the size of the typical set $\mathcal{T}^n_{[Y|U]\delta}(\mathbf{u}^n(j))$, as can be found in Sec-

tion 2.3.

$$(**) = \frac{1}{n}\sum_{j=1}^{\exp(n\hat{R})} H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n, s_1 = j)\Pr(s_1 = j) \tag{B.8a}$$

$$\leq \frac{1}{n}\sum_{j=1}^{\exp(n\hat{R})} \left(1 + \Pr\{\mathbf{Y}^n \neq \hat{\mathbf{Y}}^n|s_1 = j\}\log|\mathcal{Y}|^n\right)\Pr(s_1 = j) \tag{B.8b}$$

$$\leq \frac{1}{n} + \sum_{j=1}^{\exp(n\hat{R})} \Pr\{(\mathbf{u}^n(s_1), \mathbf{Y}^n) \notin \mathcal{T}^n_{[UY]\delta}|s_1 = j\}\log|\mathcal{Y}|\Pr(s_1 = j) \tag{B.8c}$$

$$\leq \frac{1}{n} + (\Pr(\mathcal{B}_1) + \Pr(\mathcal{B}_2))\log|\mathcal{Y}| \ . \tag{B.8d}$$

Here, (B.8b) stems from Fano's inequality [2]. As was already shown, if $\hat{R} > I(U;X)$ both $\Pr(\mathcal{B}_1)$ and $\Pr(\mathcal{B}_2)$ go to 0 when $n \to \infty$. Thus

$$H(\mathbf{Y}^n|\hat{\mathbf{Y}}^n, s_1 = j)\Pr(s_1 = j) \leq \eta_n^{(5)} \underset{n\to\infty}{\longrightarrow} 0 \ . \tag{B.9}$$

All in all:

$$\frac{1}{n}H(\mathbf{Y}^n|s_1) \leq H(Y|U) + \eta_n^{(4)} + \eta_n^{(5)} \ , \tag{B.10}$$

and

$$\frac{1}{n}I(f(\mathbf{X}^n); \mathbf{Y}^n) \geq H(Y) - H(Y|U) - \eta_n^{(4)} - \eta_n^{(5)} = I(U;Y) - \eta_n^{(4)} - \eta_n^{(5)} \ . \tag{B.11}$$

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

Analysis of the Estimation Phase: Finally, we show that given a (correct) decision H_0, the RV V can be used to decode $\mathbf{X}^n$ with the desired distortion: Denoting by $\mathcal{B}_3$ the event "an error occurred during encoding or decoding" (of V), we expend its probability as follows $\Pr(\mathcal{B}_3) \leq \Pr(\mathcal{B}_4) + \Pr(\mathcal{B}_5)$, with $\Pr(\mathcal{B}_4)$ 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 $\Pr(\mathcal{B}_5)$ being the probability that a different codeword in the same bin b is compatible with $\mathbf{y}^n$ and $\mathbf{u}(s_1)$.

$$\begin{aligned}
\Pr(\mathcal{B}_4) &\triangleq \Pr\{\nexists s_2 \text{ s.t. } (\mathbf{v}^n(s_1, s_2), \mathbf{x}^n) \in \mathcal{T}^n_{[VX|U]\delta}(\mathbf{u}^n(s_1))\} \\
&= \left[\Pr\{(\mathbf{V}^n, \mathbf{X}^n) \notin \mathcal{T}^n_{[VX|U]\delta}(\mathbf{u}(s_1))|V^n \in \mathcal{T}^n_{[V|U]\delta}(\mathbf{u}(s_1)), \mathbf{X}^n \in \mathcal{T}^n_{[X]\delta}(\mathbf{u}(s_1))\}\right]^{\exp(nS_2)} \\
&\leq \exp\left\{-\exp(nS_2)\exp[-n\left(I(V;X|U) + \eta_n^{(6)}\right)]\right\} \\
&= \exp\left\{-\exp[-n\left(I(V;X|U) - S_2 + \eta_n^{(6)}\right)]\right\} \ .
\end{aligned} \tag{B.12}$$

Thus, $\Pr(\mathcal{B}_4) \underset{n\to\infty}{\longrightarrow} 0$ if $S_2 > I(V;X|U)$. Finally,

$$\Pr(\mathcal{B}_5) \triangleq \Pr\{\exists s_2' \in b \text{ s.t. } \mathbf{v}^n(s_1, s_2') \in \mathcal{T}^n_{[V|UY]\delta}(\mathbf{u}^n(s_1), \mathbf{y}^n)\} \ , \tag{B.13}$$

with b being the bin sent to node B.

$$\begin{aligned}\Pr(\mathcal{B}_5) &\leq \exp[n\,(S_2 - R' + \epsilon)]\Pr\{\mathbf{V}^n \in \mathcal{T}^n_{[V|UY]\delta}(\mathbf{u}^n(s_1), \mathbf{y}^n) | V^n \in \mathcal{T}^n_{[V|U]\delta}(\mathbf{u}^n(s_1))\} \\ &\leq \exp[n\,(S_2 - R' + \epsilon)]\exp[-n\left(I(V;Y|U) + \eta_n^{(7)}\right)] \\ &= \exp\left[-n\left(I(V;Y|U) - (S_2 - R') + \eta_n^{(7)} - \epsilon\right)\right] .\end{aligned} \tag{B.14}$$

Thus, $\Pr(\mathcal{B}_5) \underset{n\to\infty}{\longrightarrow} 0$ if $S_2 - R' < I(V;Y|U)$, or equivalently

$$R' > S_2 - I(V;Y|U) > I(V;X|U) - I(V;Y|U) \tag{B.15a}$$
$$= I(V;XY|U) - I(V;Y|U) = I(V;X|UY) , \tag{B.15b}$$

where equality (B.15b) 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]

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 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:

$$d\big(\mathbf{x}^n, \hat{\mathbf{x}}^n(\mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n)\big) = \frac{1}{n}\sum_{i=1}^{n} d\big(x_i, \hat{x}(u_i, v_i, y_i)\big) \tag{B.16a}$$
$$= \sum_{\forall(x,u,v,y)} d\big(x, \hat{x}(u,v,y)\big) Q_{x^n u^n v^n y^n}(x,u,v,y) \tag{B.16b}$$
$$\leq \mathbb{E}\left[d(X, \hat{X}(UVY))\right] + \sum_{\forall(x,u,v,y)} |Q_{x^n u^n v^n y^n}(x,u,v,y) - p(x,u,v,y)| \tag{B.16c}$$
$$\leq \mathbb{E}_0\left[d(X, \hat{X}(UVY))\right] + d_{\max}|\mathcal{X}||\mathcal{U}||\mathcal{V}||\mathcal{Y}|\delta_n , \tag{B.16d}$$

where the summation in (B.16b) and (B.16c) 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 (B.16d) holds since $(\mathbf{x}^n, \mathbf{u}^n, \mathbf{v}^n, \mathbf{y}^n) \in \mathcal{T}^n_{[XUVY]\delta}$. Since $\delta_n \underset{n\to\infty}{\longrightarrow} 0$, the condition $D > \mathbb{E}_0\left[d\big(X, \hat{X}(UVY)\big)\right]$ is sufficient to achieve distortion $D + \epsilon$ at node B. This concludes the proof of achievability.

[1] We explicitly ignored an additional error event, which is that $\mathbf{y}^n$ is not typical. The probability of this event goes to 0 much like $\Pr(\mathcal{B}_1)$, thanks to the AEP.

B.1.2 Proof of Converse

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

$$
\begin{aligned}
nR &\geq I(W;\mathbf{X}^n) && \text{(B.17a)}\\
&= I(W;\mathbf{X}^n,\mathbf{Y}^n) = I(W;\mathbf{Y}^n) + I(W;\mathbf{X}^n|\mathbf{Y}^n) && \text{(B.17b)}\\
&= \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) + \sum_{i=1}^{n} I(W;X_i|\mathbf{Y}^n,\mathbf{X}^{i-1}) && \text{(B.17c)}\\
&= \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) + \sum_{i=1}^{n} I(W;X_i|Y_i,\mathbf{Y}_{i+1}^n,\mathbf{Y}^{i-1},\mathbf{X}^{i-1}) && \text{(B.17d)}\\
&= \sum_{i=1}^{n} \left[I(W,\mathbf{Y}^{i-1};Y_i) + I(W,\mathbf{Y}_{i+1}^n,\mathbf{Y}^{i-1},\mathbf{X}^{i-1};X_i|Y_i)\right] && \text{(B.17e)}\\
&= \sum_{i=1}^{n} \left[I(W,\mathbf{Y}^{i-1};Y_i) + I(W,\mathbf{Y}^{i-1};X_i|Y_i) + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\right] && \text{(B.17f)}\\
&= \sum_{i=1}^{n} \left[I(W,\mathbf{Y}^{i-1};Y_i,X_i) + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\right] && \text{(B.17g)}\\
&= \sum_{i=1}^{n} \left[I(W,\mathbf{Y}^{i-1};X_i) + I(\mathbf{Y}_{i+1}^n,\mathbf{X}^{i-1};X_i|Y_i,\mathbf{Y}^{i-1},W)\right] \ . && \text{(B.17h)}
\end{aligned}
$$

Here, (B.17b) and (B.17h) are due to the Markov chains $W - \mathbf{X}^n - \mathbf{Y}^n$ and $W - X_i - Y_i$, respectively. (B.17e) 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

$$
R \geq \frac{1}{n}\sum_{i=1}^{n}\left[I(U_i;X_i) + I(V_i;X_i|U_i,Y_i)\right] = I(U;X) + I(V;X|UY) \ , \tag{B.18}
$$

with U and V defined through time-sharing as is subsequently shown in (B.21).

The error exponent can now be expressed as follows:

$$
I(W;\mathbf{Y}^n) = \sum_{i=1}^{n} I(W,\mathbf{Y}^{i-1};Y_i) = \sum_{i=1}^{n} I(U_i;Y_i) = nI(U;Y) \ , \tag{B.19}
$$

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 bounded as follows. Define the function $\hat{X}_i$ as the i-th coordinate of the estimate in node B:

$$
\hat{X}_i(U_i,V_i,Y_i) \triangleq g_i(W,\mathbf{Y}^{i-1},Y_i,\mathbf{Y}_{i+1}^n) \ . \tag{B.20}
$$

The component-wise mean distortion thus verifies

$$\begin{aligned} D + \epsilon \geq \mathbb{E}_0\left[d\big(\mathbf{X}^n, g(W, \mathbf{Y}^n)\big)\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_0\left[d\big(X_Q, \hat{X}_Q(U_Q, V_Q, Y_Q)\big)|Q=i\right] \\ &= \mathbb{E}_0\left[d\big(X_Q, \hat{X}_Q(U_Q, V_Q, Y_Q)\big)\right] = \mathbb{E}_0\left[d\big(X, \hat{X}(U, V, Y)\big)\right] \ . \end{aligned} \tag{B.21}$$

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 proof in Proposition 1.

B.1.3 Cardinality bounds

It remains to establish that the cardinality bounds specified by the conditions in Theorem 1 do not affect the minimization. Toward that end we invoke the support lemma [82, p. 310] in order to deduce that $\mathcal{U}$ must have $\|\mathcal{X}\| - 1$ letters in order to ensure preservation of $p(x|u)$ plus three more to preserve the constraints on D, $I(U;X)$ and $I(U;Y)$, so $\|\mathcal{U}\| \leq \|\mathcal{X}\| + 2$ suffices. Similarly, $\mathcal{V}$ must have $\|\mathcal{X}\|\|\mathcal{U}\| - 1$ letters in order to ensure preservation of $p(x,u|v)$ plus two more to preserve D, and $I(X;V|UY)$. Thus, it suffices to have $\|\mathcal{V}\| \leq \|\mathcal{X}\|\|\mathcal{U}\| + 1$.

B.2 Proof of Theorem 2

B.2.1 Proof of Achievability

In order to achieve the region proposed in Theorem 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:

$$\begin{aligned} V &= X + W_1, \quad W_1 \sim \text{Bern}\,(\alpha) \ , \\ U &= V + W_2, \quad W_2 \sim \text{Bern}\,(\beta) \ . \end{aligned} \tag{B.22}$$

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

$$I(U;Y) = H(U) - H(U|Y) = 1 - H_2(\alpha \star \beta \star p) \ . \tag{B.23}$$

This complies with the expression proposed in Theorem 2. 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, inspired by the approach taken in [43], for the case of source estimation with side information, jointly distributed according to a BSC (without uncertainty in the probability distribution of the sources):

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 $\alpha \star p$. Thus (3.7) results in

$$R_a = I(U;Y) + [I(V;X) - I(V;Y)] = 1 - H_2(\alpha \star \beta \star p) + [H_2(\alpha \star p) - H_2(\alpha)] \ . \tag{B.24}$$

b) In this part, we let V be degenerate and $\hat{X} = g(Y,V) = Y$. 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

$$R_b = I(U;Y) = 1 - H_2(\alpha \star \beta \star p) \ . \tag{B.25}$$

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),

$$\begin{aligned} R(E,D) &= R(\theta\alpha + (1-\theta)p) \leq \theta R(\alpha) + (1-\theta)R(p) \\ &= \theta R_a + (1-\theta)R_b \leq 1 - H_2(\alpha \star \beta \star p) + \theta\left[H_2(\alpha \star p) - H_2(\alpha)\right] \ . \end{aligned} \tag{B.26}$$

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

B.2.2 Proof of Converse

Theorem 1, along with the development in (3.7), 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 Theorem 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, Theorem 1 implies that

$$E \leq I(U;Y) = H(Y) - H(Y|U) = 1 - A \ , \tag{B.27}$$

while

$$R \geq 1 - A + \theta\left[I(V;X) - I(V;Y)\right] \ , \tag{B.28}$$

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,

$$A \geq H_2\left(H_2^{-1}\left(H(X|U)\right) \star p\right) \ , \tag{B.29}$$

which stems from Ms. Gerber's Lemma (see e.g. [105]). In order to allow the exploration of the entire region defined by the bounds over A, we define $\gamma \triangleq H_2^{-1}\left(H(X|U)\right)$. Thus, the trade-off between rate and error exponent becomes

$$\begin{aligned} E &\leq 1 - H_2(\gamma \star p) \ , \\ R &\geq 1 - H_2(\gamma \star p) + \theta \left[I(V;X) - I(V;Y)\right] \ . \end{aligned} \tag{B.30}$$

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 Theorem 2. This proof has already been given in [43] and is thus omitted from this work. 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 [43], and keeping in mind the Markov chain implied by Theorem 1, it is clear that $\gamma = H^{-1}\left(H(X|U)\right) \geq \alpha$. Thus, γ can be expressed as $\gamma = \alpha \star \beta$ for some $0 \leq \beta \leq \frac{1}{2}$, which completes the proof.

B.3 Proof of Proposition 1

We now prove the achievability of the region offered in Proposition 1 for the joint detection and lossy compression problem, with general hypotheses. We start by describing the codebook, as well as encoding and decoding strategies, and follow by an analysis of error events under the proposed strategy.

B.3.1 Encoding and decoding strategy

Codebook Construction: For a given block-length n we operate on a type-by-type basis. For each type $Q_X \in \mathcal{P}_n(\mathcal{X})$, fix a conditional type $Q^{\star}_{U|X}(Q_X) \in \mathcal{P}_n(\mathcal{U})$. Randomly and uniformly choose a set of codewords denoted by $\mathcal{C}_U^n(Q_X)$, from the resulting marginal type class $\mathcal{T}^n_{Q^{\star}_U}(Q_X)$ which is induced by Q_X and $Q^{\star}_{U|X}(Q_X)$. The size of $\mathcal{C}_U^n(Q_X)$ is an integer satisfying:

$$\begin{aligned} \exp\left[nI\big(Q_X; Q^{\star}_{U|X}(Q_X)\big)\right] + (|\mathcal{U}||\mathcal{X}| + 2)\log(n+1) \\ \leq |\mathcal{C}_U^n(Q_X)| \leq \\ \exp\left[nI\big(Q_X; Q^{\star}_{U|X}(X)\big)\right] + (|\mathcal{U}||\mathcal{X}| + 4)\log(n+1) \ , \end{aligned} \tag{B.31}$$

where $\mathcal{C}_U^n(Q_X)$ is the codebook of the common message for source type Q_X. Define $f_U : \mathcal{T}^n_{Q_X} \to \mathcal{C}_U^n(Q_X)$, i.e., a function $f_U(x^n)$ that determines the codeword sent by the encoder (node A) to the decoder (node B), as subsequently explained. We define $\mathbf{U}^n \triangleq f_U(\mathbf{X}^n)$. In addition, assign an index: $k(Q_X) : \mathcal{P}_n(\mathcal{X}) \to \{1, \ldots, (n+1)^{|\mathcal{X}|}\}$ to each of the possible types of vectors $\mathbf{x}^n \in \mathcal{X}^n$.

As a second step, let V_0 and V_1 be two RVs, designed to transmit a private message to the decoder, depending on the actual distribution in effect (i.e., if it is decided that H_0 is

the true hypothesis V_0 is used and otherwise V_1 is used) such that $U - V_0 - X - Y$ and $\bar{U} - V_1 - \bar{X} - \bar{Y}$.

For each codeword $\mathbf{u}^n \in \mathcal{C}_U^n$, randomly pick $\exp[nS_0]$ sequences $\mathbf{v}_0^n(s_0)$, indexed with $s_0 = [1 : \exp(nS_0)]$, and $\exp[nS_1]$ sequences $\mathbf{v}_1^n(s_1)$, indexed with $s_1 = [1 : \exp(nS_1)]$, from the conditional typical sets $\mathcal{T}_{[V_0|U]\delta}^n(\mathbf{u}^n)$ and $\mathcal{T}_{[V_1|\bar{U}]\delta}^n(\mathbf{u}^n)$, respectively. Divide them into $\exp(nR_0)$ (respectively $\exp(nR_1)$) bins, such that each bin contains roughly $\exp[n(S_0 - R_0)]$ (respectively $\exp[n(S_1 - R_1)]$) sequences. In the remainder of this proof we only treat source reconstruction in case hypothesis H_0 was chosen, as the complementary case is completely symmetric.

Encoding: Given a sequence $\mathbf{x}^n \in \mathcal{T}_{Q_X}^n$, search for a sequence $\mathbf{u}^n \in \mathcal{C}_U^n(Q_{\mathbf{x}^n})$, i.e., in the codebook that belongs to the type $Q_{\mathbf{x}^n}$, such that $(\mathbf{u}^n, \mathbf{x}^n) \in \mathcal{T}_{[UX]\delta}^n$. As a second step, look for a codeword $\mathbf{v}_0^n(s_0)$ such that $(\mathbf{v}_0^n(s_0), \mathbf{x}^n) \in \mathcal{T}_{[V_0X|U]\delta}^n(\mathbf{u}^n)$ with the typicality measured according to the distribution induced by hypothesis H_0. Let $B_0(\mathbf{v}_0^n(\mathbf{x}^n, \mathbf{u}^n))$ denote the element (or "bin") to which $\mathbf{v}_0^n$ is mapped. Perform the same steps for the case where H_1 is the chosen hypothesis.

The encoder's message then consists of four parts:

$$\begin{aligned}
\mathcal{M}_1 &= \{1, 2, \ldots, M_1 \triangleq \exp(nR')\} , \\
\mathcal{M}_2 &= \left\{1, 2, \ldots, M_2 \triangleq (n+1)^{|\mathcal{X}|}\right\} , \\
\mathcal{M}_3 &= \{1, 2, \ldots, M_3 \triangleq \exp(nR_0)\} , \\
\mathcal{M}_4 &= \{1, 2, \ldots, M_3 \triangleq \exp(nR_1)\} , \\
\mathcal{M} &= \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3 \times \mathcal{M}_4 .
\end{aligned} \tag{B.32}$$

The encoder sends the type of $\mathbf{x}^n$ which requires $|\mathcal{M}_2|$ values but with zero rate, and also $F(f_U(\mathbf{x}^n))$, as well as the respective bins for both private messages, $B_0(\mathbf{v}_0^n(\mathbf{x}^n, \mathbf{u}^n))$ and $B_1(\mathbf{v}_1^n(\mathbf{x}^n, \mathbf{u}^n))$, to be defined subsequently. There are two cases to consider:

1 $\log|\mathcal{C}_U^n(Q_{\mathbf{x}^n})| < nR'$, in which case we can map each member of $C_U^n(Q_{\mathbf{x}^n})$ to an element of $\mathcal{M}_1$ in a one-to-one manner.

2 $\log|\mathcal{C}_U^n(Q_{\mathbf{x}^n})| \geq nR'$, in which case we assign each distinct member of $C_U^n(Q_{\mathbf{x}^n})$ to $\mathcal{M}_1$ uniformly at random.

Let $F(f_U(\mathbf{x}^n))$ denote the element to which $f_U(\mathbf{x}^n)$ is mapped. The encoder can be expressed mathematically as

$$\Psi(x) = \left(F(f_U(\mathbf{x}^n)), k(Q_{\mathbf{x}^n}), B_0(\mathbf{v}_0^n(\mathbf{x}^n, \mathbf{u}^n)), B_1(\mathbf{v}_1^n(\mathbf{x}^n, \mathbf{u}^n))\right) , \tag{B.33}$$

for each $\mathbf{x}^n \in \mathcal{T}_{Q_{\mathbf{x}^n}}^n$.

Decoding: The decoder first attempts to discover the word $\mathbf{u}^n$, by using the information sent from the encoder and the observation vector $\mathbf{y}^n$:

- If $\log|\mathcal{C}_U^n(Q_{\mathbf{x}})| < nR'$ the codeword can be decoded without error;

- Otherwise $\log|\mathcal{C}_U^n(Q_{\mathbf{x}})| \geq nR'$ the decoder receives a bin index and uses side information $\mathbf{y}^n$ to pick the best $\mathbf{u}^n$ in the bin. Given the bin number, the type $Q_{\mathbf{x}^n}$ and the side information $\mathbf{y}^n$, the decoder uses a minimal empirical entropy decoding [2] that is:
$$\phi(F(f_U(\mathbf{x}^n)), Q_{\mathbf{x}^n}, \mathbf{y}^n) = \hat{\mathbf{u}}^n \ , \tag{B.34}$$
if $H(\tilde{\mathbf{u}}^n|\mathbf{y}^n) > H(\hat{\mathbf{u}}^n|\mathbf{y}^n)$ for $\hat{\mathbf{u}}^n \in F(f_U(\mathbf{x}^n))$ and all $\tilde{\mathbf{u}}^n \in F(f_U(\mathbf{x}^n))$ with $\tilde{\mathbf{u}}^n \neq \hat{\mathbf{u}}^n$, where
$$H(\hat{\mathbf{u}}^n|\mathbf{y}^n) \triangleq -\sum_{a\in\mathcal{U},b\in\mathcal{Y}} Q_{\hat{\mathbf{u}}^{\mathbf{n}}\mathbf{y}^n}(a,b)\log Q_{\hat{\mathbf{u}}^{\mathbf{n}}|\mathbf{y}^n}(a|b)$$
is the empirical entropy of the vector $\hat{\mathbf{u}}^{\mathbf{n}}$ given the vector $\mathbf{y}^n$.

As a second step, the decoder uses the private message –either $\mathbf{v}_0^n$ or $\mathbf{v}_1^n$– destined for the case of the current hypothesis in order to estimate $\mathbf{x}^n$, with distortion D_0 or D_1, respectively. Assume hypothesis H_0 is in effect, it searches for a single sequence $\hat{\mathbf{v}}_0^n \in B_0(\mathbf{v}_0^n(\mathbf{x}^n,\mathbf{u}^n))$ such that $\hat{v}_0^n(s_0) \in \mathcal{T}_{[V_0|UY]\delta}(\mathbf{u}^n\mathbf{y}^n)$. If it finds no such sequence it declares an error during the reconstruction. If it finds more than one, it chooses one sequence at random.

B.3.2 Error probability of the testing step

We now show that, for the detection part, the exponential rate of decay of the error of the second type, under a fixed constraint over the error of the first type, is not smaller than the value claimed by Proposition 1. The analysis of possible errors at the encoder's side stays identical to the one done in the proof of Theorem 1 in Appendix B.1 (note that we assume the $P_X(x) = P_{\bar{X}}(x)$, without which the analysis of the encoder's side, with an emphasis on the codebook construction, might become more involved. Such an analysis can be found in proofs related to Chapter 4, where a similar assumption is not made). Note also that when a problem does arise during encoding, our proposed scheme calls for an error message which prompts node B to declare H_1. Thus, the influence of such errors is only on the error probability of Type I, and not on the error exponent of Type II. We concentrate in this analysis on possible errors at the decoder's side. Define two error events: First, let
$$\mathcal{B}_6 \triangleq \{\mathbf{u}^n \neq F(f_U(\mathbf{x}^n))\} \tag{B.35}$$
be the event that the chosen sequence from the bin at the decoder is different from the original sequence sent by the encoder. Then, define $\mathcal{B}_7$ to be the event of erroneous detection despite using the correct sequence. We denote the probabilities of events $\mathcal{B}_6$ and $\mathcal{B}_7$ by $P_r^{(n)}$ and $P_d^{(n)}$, respectively. Using the union bound, the probability of error in detection can be bounded by
$$P_e^{(n)} \leq P_r^{(n)} + P_d^{(n)} \ . \tag{B.36}$$

[2]Note that since our chosen test is over empirical entropies, it does not matter at this stage which hypothesis is the true one, for the sake of choosing the sequence from the bin. After having retrieved a single sequence from the bin, the decoder can continue to perform HT by discarding the rest of the sequences in the bin and only using the chosen sequence.

Evaluation of $P_r^{(n)}$: We evaluate the probability that node B chooses the wrong sequence from the bin under the suggested encoding and decoding schemes. Our evaluation is reliant on the method of types [21], and is specifically inspired by the techniques used in [86, Appendix C]. We first evaluate $P_r^{(n)}$ for a finite block-length n and then use a continuity argument to show that in the limit of $n \to \infty$,

$$-\frac{1}{n}\log P_r^{(n)} \leq G(Q_{UXY}, Q_X, Q_Y, R') = \begin{cases} \min\limits_{i=\{0,1\}} \mathcal{D}\big(Q_{UXY} \| P_{UXY_i}\big) + \big[R' - I\big(Q_X; Q_{U|X}\big) + I\big(Q_Y; Q_{U|Y}\big)\big]^+ & I\big(Q_X; Q_{U|X}\big) > R' \\ \infty & \text{else.} \end{cases} \cdot \tag{B.37}$$

Since choosing the wrong sequence can only happen in case binning is used, we are only interested in the following subset of the set of all possible sequences:

$$\mathcal{A}_n = \Big\{ (\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{U}^n \times \mathcal{X}^n \times \mathcal{Y}^n \,\Big|\, \mathbf{u}^n \in T^n_{Q^\star_{U|X}}(Q_{\mathbf{x}^n}) \,,\, \log|\mathcal{C}_U^n(Q_{\mathbf{x}^n})| \geq nR \Big\} \,. \tag{B.38}$$

We first evaluate the probability of choosing the wrong sequence within the set $\mathcal{A}_n$ by using the following lemma.

Lemma 21. *Let* $(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{A}_n$ *and let* $\mathcal{B}_8$ *be the event that* $\mathbf{u}^n \neq \phi(\psi(\mathbf{x}^n), \mathbf{y}^n)$. *If* $\log|\mathcal{C}_U^n(Q_{\mathbf{x}^n})| \geq nR$ *then*

$$\Pr\left(\mathcal{B}_8 | \mathbf{U}^n = \mathbf{u}^n, \mathbf{X}^n = \mathbf{x}^n, \mathbf{Y}^n = \mathbf{y}^n\right) \leq \exp\big[-n\left(R - J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) - \delta_n\right)\big] \,, \tag{B.39}$$

with

$$J\big(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}\big) \triangleq I\big(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n})\big) - I\big(Q_{\mathbf{u}^n|\mathbf{y}^n}; Q_{\mathbf{y}^n}\big) \tag{B.40}$$

and

$$\delta_n \triangleq \frac{1}{n}\log(n+1)^{|\mathcal{U}|(1+|\mathcal{X}|+|\mathcal{Y}|)+4} \,. \tag{B.41}$$

The probability in (B.39) *is taken over the choice of the codebook in use.*

Proof. Let $\mathcal{S}(\mathbf{u}^n|\mathbf{y}^n)$ be the set that includes all sequences $\tilde{\mathbf{u}}^n$, such that $\tilde{\mathbf{u}}^n$ has the same

type as $\mathbf{u}$ and $H(\tilde{\mathbf{u}}^n|\mathbf{y}^n) \leq H(\mathbf{u}^n|\mathbf{y}^n)$. Then

$$
\begin{aligned}
&\Pr\left(\mathcal{B}_8|\mathbf{U}^n = \mathbf{u}^n, \mathbf{X}^n = \mathbf{x}^n, \mathbf{Y}^n = \mathbf{y}^n\right) \\
&\leq \sum_{\substack{\tilde{\mathbf{u}}^n \in \mathcal{S}(\mathbf{u}^n|\mathbf{y}^n) \\ \tilde{\mathbf{u}}^n \neq \mathbf{u}^n}} \Pr\big(\tilde{\mathbf{u}}^n \in \mathcal{C}_U^n(Q_{\mathbf{x}^n}), \{F(\tilde{\mathbf{u}}^n) = F(\mathbf{u}^n)\}|\mathbf{U}^n = \mathbf{u}^n, \mathbf{X}^n = \mathbf{x}, \mathbf{Y}^n = \mathbf{y}\big) && \text{(B.42a)} \\
&\leq \sum_{\substack{\tilde{\mathbf{u}}^n \in \mathcal{S}(\mathbf{u}^n|\mathbf{y}^n) \\ \tilde{\mathbf{u}}^n \neq \mathbf{u}^n}} \Pr\big(\tilde{\mathbf{u}}^n \in \mathcal{C}_U^n(Q_{\mathbf{x}^n})|\mathbf{X}^n = \mathbf{x}^n, \mathbf{Y}^n = \mathbf{y}^n\big)\Pr\big(\{F(\tilde{\mathbf{u}}^n) = F(\mathbf{u}^n)\}\big) && \text{(B.42b)} \\
&\leq \sum_{\substack{\tilde{\mathbf{u}}^n \in \mathcal{S}(\mathbf{u}^n|\mathbf{y}^n) \\ \tilde{\mathbf{u}}^n \neq \mathbf{u}^n}} (n+1)^{|\mathcal{U}|(1+|\mathcal{X}|)+4} \exp\left[n\left(I(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n})) - H(Q_{\mathbf{u}^n})\right)\right]\frac{1}{M_1} && \text{(B.42c)} \\
&\leq (n+1)^{|\mathcal{U}||\mathcal{Y}|} \exp\left[nH(Q_{\mathbf{u}^n|\mathbf{y}^n}|Q_{\mathbf{y}^n})\right]\frac{1}{M_1}(n+1)^{|\mathcal{U}|(1+|\mathcal{X}|)+4}\times && \text{(B.42d)} \\
&\qquad\qquad \times \exp\left[n\left(I(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n})) - H(Q_{\mathbf{u}^n})\right)\right] \\
&= (n+1)^{|\mathcal{U}|(1+|\mathcal{X}|+|\mathcal{Y}|)+4} \exp\left[-n\left(R - H(Q_{\mathbf{u}^n|\mathbf{y}^n}|Q_{\mathbf{y}^n}) + H(Q_{\mathbf{u}^n}) - I(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n}))\right)\right] && \text{(B.42e)} \\
&= (n+1)^{|\mathcal{U}|(1+|\mathcal{X}|+|\mathcal{Y}|)+4} \exp\left[\, -n\left(R + I(Q_{\mathbf{u}^n|\mathbf{y}^n}; Q_{\mathbf{y}^n})) - I(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n}))\right)\right] && \text{(B.42f)} \\
&\triangleq (n+1)^{|\mathcal{U}|(1+|\mathcal{X}|+|\mathcal{Y}|)+4} \exp\left[\, -n\left(R - J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n})\right)\right] && \text{(B.42g)} \\
&\leq \exp\left[\, -n\left(R - J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) - \delta_n\right)\right], && \text{(B.42h)}
\end{aligned}
$$

with δ_n as defined above. Here, the probability $\Pr\left(\tilde{\mathbf{u}}^n \in \mathcal{C}_U^n(Q_{x^n})\right)$ is over the choice of the codebook. Inequality (B.42b) stems from the codebook construction, which divides sequences into bins randomly and independently. Inequality (B.42c) is due to [86, Lemma 12], which applies here without change, and to the upper bound over the size of $\mathcal{C}_U^n(Q_{\mathbf{x}^n})$, given in (B.31). Inequality (B.42d) is due to Lemma 10. Finally, equality (B.42e) is due to the definition of M_1 and (B.42h) stems from the fact that $\Pr\left(\mathcal{B}_8|\mathbf{U}^n = \mathbf{u}^n, \mathbf{X}^n = \mathbf{x}^n, \mathbf{Y}^n = \mathbf{y}^n\right) \leq 1$ and the definition of δ_n. □

We now bound the probability of error in choosing the right sequence in the bin $P_r^{(n)}$, for a finite block-length n:

$$
\begin{aligned}
P_r^{(n)} &= \Pr\left(\{\mathbf{u}^n \neq F(f_U(\mathbf{x}^n))\}\right) && \text{(B.43a)} \\
&\leq \sum_{(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{A}_n} \Pr\left(\mathcal{B}_8|\mathbf{U}^n = \mathbf{u}, \mathbf{X}^n = \mathbf{x}, \mathbf{Y}^n = \mathbf{y}\right)\Pr\left(\mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\right) && \text{(B.43b)} \\
&\leq \sum_{(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{A}_n} \exp\left[-n\left(R - J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) - \delta_n\right)\right] P_{XY}^n(\mathbf{x}^n, \mathbf{y}^n)\frac{1}{|\mathcal{T}^n_{Q^\star_{U|X}}(Q_{\mathbf{x}^n})|}\,. && \text{(B.43c)}
\end{aligned}
$$

Here, claim (B.43c) is derived from Lemma 21. Note the slight abuse of notation here, where $P_{XY}^n(\mathbf{x}^n, \mathbf{y}^n)$ in (B.43c) refers to the *real distribution* controlling the RVs, and can thus actually be, according to the true hypothesis, wither $P_{XY}^n(\mathbf{x}^n, \mathbf{y}^n)$ or $P_{\bar{X}\bar{Y}}^n(\mathbf{x}^n, \mathbf{y}^n)$. The probability of choosing a specific sequence $\mathbf{u}^n$ given both source sequences $\mathbf{x}^n$ and $\mathbf{y}^n$

stems from averaging over the code. We can now change the expression to sum first on types and then on sequences within each type class. In order to transform our summation over a set of sequences $\mathcal{A}_n$ into a summation over a set of types (and only then over the sequences within each type) we define the following set of types:

$$\mathcal{D}(Q_X, Q_Y) = \left\{ Q_{UXY} \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X} \times \mathcal{Y}) : Q_{U|X} = Q^{\star}_{U|X}(Q_X)\,,\, \log|\mathcal{C}^n_U(Q_X)| \geq nR \right\}. \tag{B.44}$$

The probability of error in selecting the sequence can thus be bound by:

$$P_r^{(n)} \leq \sum_{Q_X, Q_Y} \left[\sum_{Q_{UXY} \in \mathcal{D}(Q_X, Q_Y)} \sum_{(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{T}^n_{Q_{UXY}}} \frac{P^n_{XY}(\mathbf{x}^n, \mathbf{y}^n)}{|\mathcal{T}^n_{Q^{\star}_{U|X}}(Q_{\mathbf{x}^n})|} \exp\left[-n\big(R - J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) - \delta_n\big)\right] \right]. \tag{B.45}$$

In the case of distributed HT, the probability of the source sequences $(\mathbf{x}^n, \mathbf{y}^n)$ is unknown, since the sequences can be created by one of two possible distributions. We thus bound the probability of the observed sources by

$$\begin{aligned} P^n_{XY}(\mathbf{x}^n, \mathbf{y}^n) &\leq \max\{P_{XY}(\mathbf{x}^n, \mathbf{y}^n), P_{\bar{X}\bar{Y}}(\mathbf{x}^n, \mathbf{y}^n)\} \\ &= \max_{i=\{0,1\}} \left\{ \exp\left[-n\left(\mathcal{D}(Q_{XY} \| P_{XY_i}) + H(Q_{XY})\right)\right] \right\} \\ &= \exp\left[-n \left(\min_{i=\{0,1\}} \mathcal{D}(Q_{XY} \| P_{XY_i}) + H(Q_{XY}) \right) \right], \end{aligned} \tag{B.46}$$

where, in accordance to the notation of Proposition 1, we use the subscript i in order to differentiate between P_{XY} and $P_{\bar{X}\bar{Y}}$. Using the following facts detailed in Lemma 5,

$$|\mathcal{T}^n_{Q_{UXY}}| \leq \exp\left[n(H(Q_{UXY}))\right] \leq \exp\left(n \log|\mathcal{U}||\mathcal{X}||\mathcal{Y}|\right), \tag{B.47a}$$

$$|\mathcal{T}^n_{Q_{U|X}}| \geq (n+1)^{-|\mathcal{U}||\mathcal{X}|} \exp\left[n\left(H(Q_{U|X}|Q_X)\right)\right], \tag{B.47b}$$

we obtain that

$$\begin{aligned} P_r^{(n)} \leq & \sum_{Q_X \in \mathcal{P}_n(\mathcal{X})} \sum_{Q_Y \in \mathcal{P}_n(\mathcal{Y})} \left[\sum_{Q_{UXY} \in \mathcal{D}(Q_X, Q_Y)} \exp\left[-n \left(\min_{i=\{0,1\}} \mathcal{D}(Q_{XY} \| P_{XY_i}) + H(Q_{XY}) \right) \right] \times \right. \\ & (n+1)^{|\mathcal{U}||\mathcal{X}|} \exp\left[nH(Q_{U|X}|Q_X)\right] \times \exp\left[nH(Q_{UXY})\right] \exp\left[-n\left(R - J(Q_{UXY}) - \delta_n\right)\right] \\ \leq & \sum_{Q_X \in \mathcal{P}_n(\mathcal{X})} \sum_{Q_Y \in \mathcal{P}_n(\mathcal{Y})} \sum_{Q_{UXY} \in \mathcal{D}(Q_X, Q_Y)} \exp\left[-n\left(\Gamma + R - J(Q_{UXY}) - \delta_n\right)\right], \end{aligned}$$

with Γ satisfying:

$$\begin{aligned}
\Gamma &= \min_{i=\{0,1\}} \mathcal{D}(Q_{XY}\|P_{XY_i}) + H(Q_{XY}) + H(Q_{U|X}|Q_X) - H(Q_{UXY}) \\
&= \min_{i=\{0,1\}} \mathcal{D}(Q_{XY}\|P_{XY_i}) + H(Q_{U|X}|Q_X) - H(Q_{U|XY}|Q_{XY}) \\
&= \min_{i=\{0,1\}} \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} Q_{XY}(x,y)\log\frac{Q_{XY}(x,y)}{P_{XY_i}(x,y)} - \sum_{\substack{u\in\mathcal{U}\\ x\in\mathcal{X}}} Q_{UX}(u,x)\log\frac{Q_{UX}(u,x)}{Q_X(x)} \\
&\qquad + \sum_{\substack{u\in\mathcal{U}\\ x\in\mathcal{X}\\ y\in\mathcal{Y}}} Q_{UXY}(u,x,y)\log\frac{Q_{UXY}(u,x,y)}{Q_{XY}(x,y)} \\
&= \min_{i\in\{0,1\}} \left\{ \sum_{\substack{u\in\mathcal{U}\\ x\in\mathcal{X}\\ y\in\mathcal{Y}}} Q_{UXY}(u,x,y)\log\frac{Q_{XY}(x,y)}{P_{XY_i}(x,y)}\frac{Q_X(x)}{Q_{UX}(u,x)}\frac{Q_{UXY}(u,x,y)}{Q_{XY}(x,y)} \right\} \\
&= \min_{i=\{0,1\}} \left\{ \sum_{\substack{u\in\mathcal{U}\\ x\in\mathcal{X}\\ y\in\mathcal{Y}}} Q_{UXY}(u,x,y)\log\frac{Q_{UXY}(u,x,y)}{P_{XY_i}(x,y)Q_{U|X}(u|x)} \right\} \\
&= \min_{i=\{0,1\}} \mathcal{D}(Q_{UXY}\|P_{XY_i}Q_{U|X}) \ .
\end{aligned} \tag{B.48}$$

The probability of error in bin decoding can thus be concluded to satisfy

$$\begin{aligned}
P_r^{(n)} \leq & \sum_{Q_X\in\mathcal{P}_n(\mathcal{X})} \sum_{Q_Y\in\mathcal{P}_n(\mathcal{Y})} \sum_{Q_{UXY}\in\mathcal{D}(Q_X,Q_Y)} \\
& \exp\left[-n\left(\min_{i=\{0,1\}} \mathcal{D}(Q_{UXY}\|P_{XY_i}Q_{U|X}) + R - J(Q_{UXY}) - \delta_n\right)\right] \ .
\end{aligned} \tag{B.49}$$

We may now upper bound the summations by maximizing over the types and optimizing over the choice of the of the test channel $Q^\star_{U|X}$. We optimize to then obtain:

$$P_r^{(n)} \leq |\mathcal{P}_n(\mathcal{X})| \max_{Q_X} \min_{Q^\star_{U|X}} |\mathcal{P}_n(\mathcal{Y})| \max_{Q_Y} |\mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})| \max_{\substack{Q_{UXY}\\ Q_{U|X}=Q^\star_{U|X}}} \exp\left\{-nG_n\left[Q_{UXY},Q_X,Q_Y,R\right]\right\} . \tag{B.50}$$

Thus,

$$\begin{aligned}
\frac{1}{n}\log P_r^{(n)} \leq & - \min_{Q_X\in\mathcal{P}_n(\mathcal{X})} \max_{Q^\star_{U|X}(Q_X)} \min_{Q_Y\in\mathcal{P}_n(\mathcal{Y})} \min_{\substack{Q_{UXY}\\ Q_{U|X}=Q^\star_{U|X}}} G_n\left[Q_{UXY},Q_X,Q_Y,R\right] \\
& \times \log\left(|\mathcal{P}_n(\mathcal{X})||\mathcal{P}_n(\mathcal{Y})||\mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})|\right)
\end{aligned}$$

with the function $G_n\left[Q_{UXY},Q_X,Q_Y,R\right]$ defined as follows:

$$G_n\left[Q_{UXY},Q_X,Q_Y,R\right] = \begin{cases} \min\limits_{i=\{0,1\}} \mathcal{D}(Q_{UXY}\|P_{XY_i}Q_{U|X}) & \\ \quad + \left[R - I(Q_X;Q^\star_{U|X}) + I(Q_Y;Q^\star_{U|Y})\right] & I(Q_X;Q^\star_{U|X}) > R \\ +\infty & \text{else .} \end{cases} \tag{B.51}$$

The cardinalities can be absorbed inside the exponent and become insignificant as $n \to \infty$. From continuity arguments under discrete alphabets, it is made clear that [86, Lemma 14]:

$$P_r^{(n)} \leq \inf_{Q_X \in \mathcal{P}(\mathcal{X})} \sup_{Q^\star_{U|X}(Q_X) \in \mathcal{P}(\mathcal{U})} \inf_{Q_Y \in \mathcal{P}(\mathcal{Y})} \inf_{\substack{Q_{UXY} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}) \\ Q_{U|X} = Q^\star_{U|X}}} G\left[Q_{UXY}, Q_X, Q_Y, R\right] \ , \tag{B.52}$$

where all the optimization steps are now being taken over *probability distributions*, and G is as defined in Proposition 1.

Evaluation of $P_d^{(n)}$: We now study the Type II error probability of detection, under the assumption that the right sequence has been correctly extracted from the bin. The probability that, given the right sequence $\mathbf{u}^n$, node B makes a wrong decision was investigated in detail in [9], using the method of types [21], as well as properties of types and typical sequences, detailed in Chapter 2 of this thesis. That result, however, is dependent on a specific codebook, conceived to allow detection with high probability. As we use a random codebook in our scheme, it is essential to adapt the method of [9]. We give here a general description of this adaptation. The complete proof of the error exponent while using the right sequence is a special case of the proof given in Appendix C.1 for hypothesis testing with cooperative communication.

We propose here a slight modification to [9]. Intuitively, since we investigate the exponential decay of β_n while only enforcing a fixed upper bound on α_n, we show that the penalty of replacing the codebook construction in [9] with random coding can be fully absorbed into α_n, leaving the error exponent result of β_n unmodified. Nevertheless, α_n can still be shown to approach 0 as n grows, which indicates that any constraint $\alpha_n \leq \epsilon$ can be fulfilled, for n large enough and $\epsilon > 0$. For the given codebook, define

$$\begin{aligned}\mathcal{L}(Q^\star_{UX}, Q^\star_{UY}) = \Big\{ P_{\tilde{U}\tilde{X}\tilde{Y}} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y}) \ : & P_{\tilde{U}\tilde{X}}(u,x) = Q^\star_{UX}(u,x), \\ & P_{\tilde{U}\tilde{Y}}(u,y) = Q^\star_{UY}(u,y), \forall\,(u,x,y) \Big\} \ ,\end{aligned} \tag{B.53}$$

to be the set of all triplets of auxiliary RVs such that the marginal distribution of each pair (U,X) and (U,Y) is maintained. Similarly to [9], it is not difficult to show that, for the codebook described above,

$$\theta_L(R) \triangleq \min_{\tilde{U}\tilde{X}\tilde{Y} \in \mathcal{L}(Q^\star_{UX}, Q^\star_{UY})} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}} \| P_{\bar{U}\bar{X}\bar{Y}}) \tag{B.54}$$

provides a lower bound to the error probability of the second type, after the correct sequence has been recovered from the bin, and under a fixed error probability of the first type.

From the construction of the codebook (specifically the size of the set $\mathcal{C}_U^n(Q_{\mathbf{x}^n})$), it can be seen that the number of sequences in the codebook *per type of X* complies with $M = \exp\left[n(I(Q_{\mathbf{x}^n}; Q^\star_{U|X}(Q_{\mathbf{x}^n})) + \eta)\right]$. Given a sequence $\mathbf{x}^n$, search for a sequence $\mathbf{u}_i$ in the codebook that belongs to the type of $\mathbf{x}^n$, such that $(\mathbf{u}_i^n, \mathbf{x}^n) \in \mathcal{T}^n_{[UX]\delta}$ and send its

index (or bin number, depending on the type of $\mathbf{x}^n$) to the receiver. As we only consider here the error event where the wrong hypothesis is chosen despite the correct sequence is used, we ignore errors in choosing the correct sequence from the bin, in case binning has occurred, for the sake of this analysis. If there is more than one such sequence choose randomly. If there is no such sequence in the codebook, send an error message. At the decoder (node B), if $(\mathbf{u}_i^n, \mathbf{y}^n) \in \mathcal{T}_{[UY]\delta}^n$ (notice that typicality here is checked only under hypothesis H_0) declare H_0. In any other case (including the case an error message was received) declare H_1. This choice allows us to "push" the penalty of not using the code proposed in [9, Lemma 4] into α_n (which, when $n \to \infty$ can still be bounded by any fixed $\epsilon > 0$), thus leaving the evaluation of β_n unchanged, as shown subsequently.

Evaluation of α_n: An error of the first type occurs if for n i.i.d. samples $(\mathbf{x}^n, \mathbf{y}^n) \sim P_{XY}(x, y)$ (hypothesis H_0 holds) the decoder declares H_1. According to the proposed coding schemes, two possible events can induce the decoder to such an error. The first is given by

$$\text{(i)} \quad \mathcal{B}_9 \triangleq \{\nexists\ i \text{ such that } (\mathbf{u}_i^n, \mathbf{x}^n) \in \mathcal{T}_{[UX]\delta}^n\} \ . \tag{B.55}$$

Assuming without loss of generality that the sequence $\mathbf{u}_1^n$ was chosen and sent from node A, the second relevant error event is:

$$\text{(ii)} \quad \mathcal{B}_{10} \triangleq \{H_0 \text{ is true and } (\mathbf{u}_1^n, \mathbf{y}^n) \notin \mathcal{T}_{[UY]\delta}^n\} \ . \tag{B.56}$$

From the union bound, it is obvious that:

$$\alpha_n \leq \Pr(\mathcal{B}_9) + \Pr(\mathcal{B}_{10} \cap \mathcal{B}_9^c) \ . \tag{B.57}$$

Through the AEP it is easy to conclude that both of these probabilities approach zero when $n \to \infty$. Thus, for n large enough one can conclude that $\alpha_n \leq \epsilon$ for any fixed $\epsilon > 0$.

Evaluation of β_n: The error of the second type can be defined by a single event:

$$\mathcal{B}_{11} \triangleq \{H_1 \text{ is true and } (\mathbf{u}_1^n, \mathbf{y}^n) \in \mathcal{T}_{[UY]\delta}^n\} \ . \tag{B.58}$$

The analysis of β_n is identical to what was done in [9]. One important difference, however, is that by defining

$$\mathscr{C}_i \triangleq \left\{\mathbf{x}^n \in \mathcal{X}^n : \quad (\mathbf{u}_i^n, \mathbf{x}^n) \in \mathcal{T}_{[UX]\delta}^n\right\} \ , \tag{B.59}$$

the sets $\mathscr{C}_i$ are not necessarily disjoint. This, however, does not change the calculations by following same steps as in [9]. Readers are invited to consult Appendix C.1 for a full analysis of the error exponent of Type II in a bidirectional distributed system through random codes, of which the unidirectional scenario is a private case.

B.3.3 Source reconstruction

As a final step, we demonstrate the achievability of the estimation part in Proposition 1, for the case where hypothesis H_0 is chosen (the case of hypothesis H_1 is symmetric). Denoting by $\mathcal{B}_{12}$ the event "an error occurred during encoding or decoding, under the

correct decision H_0", we expend its probability as follows: $\Pr(\mathcal{B}_{12}) \leq P' + P''$, with P' being the probability that no codeword $\mathbf{v}_0^n(s_0)$ could be found in the codebook for the given sequence $\mathbf{x}^n$ and the chosen sequence $\mathbf{u}^n$, and P'' being the probability that a different codeword in the same bin is compatible with $\mathbf{y}^n$ and $\mathbf{u}^n$.

Using standard arguments, both error probabilities can be bounded as follows:

$$\begin{aligned} P' &\triangleq \Pr\{\nexists\, s_0 = [1 : \exp(nS_0)] \text{ s.t. } (\mathbf{V}_0^n(s_0), \mathbf{X}^n) \in \mathcal{T}^n_{[V_0X|U]\delta}(\mathbf{u}^n)\} \\ &\leq \Pr\{(V_0^n, X^n) \notin \mathcal{T}^n_{[V_0X|U]\delta}(\mathbf{u}^n) | \, V^n \in \mathcal{T}^n_{[V_0|U]\delta}(\mathbf{u}^n), X^n \in \mathcal{T}^n_{[X|U]\delta}(\mathbf{u}^n)\}^{\exp(nS_0)} \\ &\leq \exp\big\{-\exp[nS_0]\exp\big[-n(I(P_{X|U}; P_{V_0|XU}|P_U) + \eta_n^{(1)})\big]\big\} \\ &= \exp\big\{-\exp\big[-n\big(I(P_{X|U}; P_{V_0|XU}|P_U) - S_0 + \eta_n^{(1)}\big)\big]\big\} \,. \end{aligned} \tag{B.60}$$

Thus, $P' \to 0$ provided that $S_0 > I(P_{X|U}; P_{V_0|XU}|P_U)$. Next,

$$\begin{aligned} P'' &\triangleq \Pr\big\{\exists \hat{s}_0 \in [1 : \exp(nS_0)] \text{ s.t. } \mathbf{V}_0^n(\hat{s}_0) \in \mathcal{T}^n_{[V_0|UY]\delta}(\mathbf{u}^n\mathbf{y}^n), B_0\big(\mathbf{v}_0^n(s_0)\big) = B_0\big(\mathbf{v}_0^n(\hat{s}_0)\big)\big\} \\ &\leq \exp[n(S_0 - R_0 + \epsilon)] \Pr\{(\mathbf{V}_0^n, \mathbf{Y}^n) \in \mathcal{T}_{[V_0Y|U]\delta}(\mathbf{u}^n) | \mathbf{V}_0^n \in \mathcal{T}_{[V_0|U]\delta}(\mathbf{u}^n)\,,\, \mathbf{Y}^n \in \mathcal{T}_{[Y|U]\delta}(\mathbf{u}^n)\} \\ &\leq \exp[n(S_0 - R_0 + \epsilon)] \exp\big[-n\big(I(P_{Y_0|U}; P_{V_0|Y_0U}|P_U) + \eta_n^{(2)}\big)\big] \\ &= \exp\big\{-n\big[I(P_{Y_0|U}; P_{V_0|Y_0U}|P_U) - (S_0 - R_0) + \eta_n^{(2)} - \epsilon\big]\big\} \,. \end{aligned} \tag{B.61}$$

Here, $B_0(\mathbf{v}_0^n(s_0))$ denotes the bin $\mathbf{v}_0^n(s_0)$ belongs to, as defined as part of the encoding strategy. R_0 is the rate consecrated to the estimation part, for the case that H_0 was chosen as the correct hypothesis. Defining R_1 equivalently for hypothesis H_1, the total available rate can be said to be divided, under the proposed achievable scheme, to three parts, such that $R = R' + R_0 + R_1$. Thus, $P'' \to 0$ if $S_0 - R_0 < I(P_{X|U}; P_{V_0|XU}|P_U)$, or equivalently

$$\begin{aligned} R_0 &> S_0 - I(P_{Y_0|U}; P_{V_0|Y_0U}|P_U) \\ &> I(P_{X|U}; P_{V_0|XU}|P_U) - I(P_{Y_0|U}; P_{V_0|YU}|P_U) \\ &= I(P_{XY|U}; P_{V_0|XYU}|P_U) - I(P_{Y|U}; P_{V_0|YU}|P_U) \\ &= I(P_{X|UY}; P_{V_0|UXY}|P_{UY}) \,. \end{aligned} \tag{B.62}$$

Thus, the probability of error related to source reconstruction goes to zero provided that $S_0 > I(P_{X|U}; P_{V_0|XU}|P_U)$ and $R_0 > I(P_{X|UY}; P_{V_0|XUY}|P_{UY})$. Combining this result with the symmetric case of H_1 and the result for the detection step, the required total rate of communication reads

$$R > R' + I(P_{X|UY}; P_{V_0|XUY}|P_{UY}) + I(P_{\bar{X}|\bar{U}\bar{Y}}; P_{V_1|\bar{X}\bar{U}\bar{Y}}|P_{\bar{U}\bar{Y}}) \,. \tag{B.63}$$

We now know that our scheme allows the decoding of either $\mathbf{v}_0$ and $\mathbf{v}_1$, depending on the case, with high probability, when $n \to \infty$. It remains to be shown that using the sequence $\mathbf{v}_0^n$, it is possible to recover $\mathbf{x}^n$ with distortion D_0. We choose a (possibly

suboptimal) decoder, that reconstructs $\mathbf{x}^n$ only from $(\mathbf{u}^n, \mathbf{y}^n, \mathbf{v}_0^n)$:

$$\begin{aligned} d(\mathbf{x}^n,&\hat{\mathbf{x}}^n(\mathbf{u}^n, \mathbf{y}^n, \mathbf{v}_0^n)) = \frac{1}{n}\sum_{i=1}^{n} d\big(x_i, \hat{x}_i(u^n, y^n, v_0^n)\big) \\ &= \frac{1}{n}\sum_{\forall(x,u,y,v_0)} d\big(x, \hat{x}(u,y,v_0)\big) N(x,u,y,v_0|\mathbf{x}^n\mathbf{u}^n\mathbf{y}^n\mathbf{v}_0^n) \\ &\leq \mathbb{E}_0\left[d\big(X, \hat{X}(UYV_0)\big)\right] + \sum_{\forall(x,u,y,v_0)} \left|\frac{1}{n}N(x,u,y,v_0|\mathbf{x}^n\mathbf{u}^n\mathbf{y}^n\mathbf{v}_0^n) - p(x,u,y,v_0)\right| \\ &\leq \mathbb{E}_0\left[d\big(X, \hat{X}(UYV_0)\big)\right] + d_{\max}|\mathcal{X}||\mathcal{Y}||\mathcal{U}||\mathcal{V}_0|\delta_n \ , \end{aligned} \tag{B.64}$$

where the summation is over all the possible letters in the respective alphabets of the RVs, and the final inequality holds since $(\mathbf{x}^n, \mathbf{y}^n, \mathbf{u}^n, \mathbf{v}_0^n) \in \mathcal{T}^n_{[XYUV_0]\delta}$. Since $\delta_n \to 0$ when $n \to \infty$, any distortion D_0 can be achieved, as long as $D_0 > \mathbb{E}_0\left[d\big(X, \hat{X}(UYV_0)\big)\right]$.

B.4 Proof of Proposition 2

We now prove the achievability of the error exponent offered in Proposition 2, for the case where source reconstruction is not required. As the proof is in many ways similar to the proof of Proposition 1, given in Appendix B.3, we concentrate mainly on the differences.

B.4.1 Codebook generation and encoding strategy

Both the codebook generation and the encoding strategy in this case are very similar to what was done in the proof of Proposition 1, in the part consecrated to detection. The only difference is that now we choose to only work with δ-typical sequences, for some arbitrary δ. When node A sees a non-typical sequence $\mathbf{x}$, it sends an error message. In the opposite case, encoding is done as before. Note that while we only work with δ-typical sequences, there are still different codebooks for each type *within the set* of δ-typical sequences.

B.4.2 Decoding strategy

In case an error message is received, the decoder declares H_1. This strategy implies that any probability of the error event caused by the encoder not seeing a δ-typical sequence is allocated to α_n, rather than β_n. The probability of this event, however, goes to zero when $n \to \infty$ thanks to the AEP, implying that $\alpha_n \leq \epsilon$ for any $\epsilon > 0$, for $n \geq n_0(\epsilon, \delta)$, thus satisfying the constraint over α_n.

When the encoder does not send an error message, the decoder operates on the entire bin in order to make a decision. Going over the sequences in the bin one by one, the

decoder checks for each $\mathbf{u}_i^n$ if $(\mathbf{u}_i^n, \mathbf{y}^n) \in T_{[UY]\delta}^n$. If a sequence in the bin is found, which is jointly typical with $\mathbf{y}^n$, the decoder declares H_0. If no such sequence is found, the decoder declares H_1. Note that under this strategy, the decoder does not attempt to find the original sequence sent by the encoder. Specifically, when the decoder declares H_1 it is completely oblivious to the original codeword.

B.4.3 Probability of error

The analysis of the probability of error in detection under this new strategy is very similar to the analysis given in Appendix B.3. We separately bound the corresponding error probabilities on the two possible error events.

Analysis of α_n: When analyzing $\alpha_n(\mathcal{A}_n) = \Pr\left(\mathcal{A}_n^c | XY \sim p_0(x,y)\right)$, we assume throughout that the probability measure in effect is p_0. Two scenarios can lead to an event where the decoder erroneously declares H_1:

$$\begin{aligned} \mathcal{B}_{13} &\triangleq \left\{ \nexists\, i \in \mathcal{C}_U^n(Q_{\mathbf{x}^n}) \,\middle|\, (\mathbf{x}^n, \mathbf{u}_i^n) \in \mathcal{T}_{[UX]\delta}^n \right\} , \\ \mathcal{B}_{14} &\triangleq \left\{ \nexists\, i \in F(f(\mathbf{x}^n)) \,\middle|\, (\mathbf{u}_i^n, \mathbf{y}^n) \in \mathcal{T}_{[UY]\delta}^n \right\} . \end{aligned} \tag{B.65}$$

In the first event, an error message is sent, as there is no fitting codeword within the codebook for the observed sequence $\mathbf{x}^n$. Whereas for the second event, there is no sequence in the bin that prompts the decoder to decide H_0, despite it being the true hypothesis. The probability of event $\mathcal{B}_{13}$ goes to zero with n, thanks to the AEP and the size of the codebook. As for event $\mathcal{B}_{14}$, assume without loss of generality, that the encoder intended to send the first word in the bin $\mathbf{u}_1^n$, i.e., $\mathbf{u}_1^n = f(\mathbf{x}^n)$. The probability that the decoder declares H_1 can be upper-bounded by

$$\Pr(\mathcal{B}_{14}) = \Pr\left\{ \nexists\, i \in F(f(\mathbf{X}^n)) \,\middle|\, (\mathbf{U}_i^n, \mathbf{Y}^n) \in \mathcal{T}_{[UY]\delta}^n \right\} \leq \Pr\{(\mathbf{U}_1^n, \mathbf{Y}^n) \notin \mathcal{T}_{[UY]\delta}^n\} , \tag{B.66}$$

where typicality is measured over the probability measure $p_0 = P_{XY}$. As was already discussed above, this probability tends to 0 with the number of available realizations n. This result is attributed to the AEP, by which $\mathbf{x}$ and $\mathbf{y}$ are jointly typical with high probability, and to the generalized Markov Lemma (Lemma 8). Thus, any fixed constraint over the probability of error of the first type $\alpha \leq \epsilon$ ($\epsilon > 0$), may be satisfied when n is large enough.

Analysis of β_n: As we now turn to analyzing the probability of error of the second type, we assume throughout this part that the real hypothesis is H_1. As was the case in Appendix B.3, the resulting error exponent is the result of a trade-off between two error events. While the analysis of the event where the correct sequence prompts a wrong decision (i.e. in this case is $(f(\mathbf{x}^n), \mathbf{y}^n) \in \mathcal{T}_{[UY]\delta}^n$) stays the same, the second error event is now different. We thus concentrate in this appendix on calculating the probability of the event that some sequence in the bin $\mathbf{u}^n \neq f(\mathbf{x}^n)$ prompts the decoder to declare H_0. We start by presenting the following lemma:

Lemma 22. *Let $\mathcal{A}_n$ be the set of triplets, such that a binned codebook is necessary:*

$$\mathcal{A}_n = \left\{ (\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in T^n_{Q^\star_{U|X}} \times \mathcal{X}^n \times \mathcal{Y}^n \;\middle|\; \log |\mathcal{C}^n_U(Q_{\mathbf{x}^n})| \geq nR \right\} . \tag{B.67}$$

Let $(\mathbf{u}^n, \mathbf{x}^n, \mathbf{y}^n) \in \mathcal{A}_n$ and denote by $\mathcal{B}_{15}$ the event indicating that $(\mathbf{u}^n, \mathbf{y}^n) \in \mathcal{T}^n_{[UY]\delta}$, for some $\mathbf{u}^n \neq f(\mathbf{x}^n)$ in the bin. Then,

$$\Pr\left(\mathcal{B}_{15} | \mathbf{U}^n = \mathbf{u}^n, \mathbf{X}^n = \mathbf{x}^n, \mathbf{Y}^n = \mathbf{y}^n\right) \leq \exp\left[-n \left(R - \hat{J}(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) - \delta_n \right) \right] , \tag{B.68}$$

with

$$\hat{J}(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n}) \triangleq I\big(Q_{\mathbf{x}^n}; Q^\star_{U|X}\big) - H(Q_{\mathbf{u}^n}) + H\big(Q_{U|Y}|P_Y\big) \tag{B.69}$$

and

$$\delta_n \triangleq \frac{1}{n} \log(n+1)^{|\mathcal{U}|(1+|\mathcal{X}|+|\mathcal{Y}|)+4} + \epsilon_n \tag{B.70}$$

with $\epsilon_n \to 0$ when $n \to \infty$. Moreover, the probability in (B.68) *is taken over the choice of the codebook in use.*

Proof. The proof of Lemma 22 is very similar to the one given for Lemma 21. The difference is that now the set of sequences that "confuses" the decoder is simply $\hat{\mathcal{S}}(\mathbf{y}^n) = \mathcal{T}^n_{[U|Y]\delta}(\mathbf{y}^n)$. Bounding the set of conditionally typical sequences by [105]:

$$\left|\mathcal{T}^n_{[U|Y]\delta}(\mathbf{y}^n)\right| \leq (n+1)^{|\mathcal{U}||\mathcal{Y}|} \exp\left[n(H(Q_{U|Y}|P_Y) + \epsilon_n) \right] , \tag{B.71}$$

for each $\mathbf{y}^n \in \mathcal{T}^n_{[Y]\delta}$, completes the proof. □

Remark 17. *Note that unlike $J(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n})$, the quantity $\hat{J}(Q_{\mathbf{u}^n\mathbf{x}^n\mathbf{y}^n})$ is not dependent on the observed $\mathbf{y}^n$. The quantity $H(Q_{U|Y}|P_Y)$ can be analytically calculated when the type of $\mathbf{x}^n$ and the chosen strategy $Q_{U|X}$ is known, without knowing neither the specific sent sequence $\mathbf{u}^n$ nor the observed sequence $\mathbf{y}^n$.*

Using Lemma 22 and summing over all involved types and sequences within each type as was done in Appendix B.3, the probability of the event where an unintended sequence in the bin causes an error can be bounded by

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n} \log \Pr(\mathcal{B}_{15}) \geq \\
&\min_{Q_X \in \mathcal{P}_n(\mathcal{X})} \; \max_{Q^\star_{U|X}(Q_X) \in \mathcal{P}_n(\mathcal{U})} \; \min_{Q_Y \in \mathcal{P}_n(\mathcal{Y})} \; \min_{Q_{UXY} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})} \left\{ \mathcal{D}(Q_{UXY} \| P_{\bar{U}\bar{X}\bar{Y}}) + R - \hat{J}(Q_{UXY}) \right\} \\
&= \min_{Q_X \in \mathcal{P}_n(\mathcal{X})} \; \max_{Q^\star_{U|X}(Q_X) \in \mathcal{P}_n(\mathcal{U})} \; \min_{Q_Y \in \mathcal{P}_n(\mathcal{Y})} \; \min_{Q_{UXY} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})} \Big\{ \mathcal{D}(Q_{UXY} \| P_{\bar{U}\bar{X}\bar{Y}}) + R \\
&\qquad\qquad - I(Q_X; Q^\star_{U|X}) + I(Q^\star_{U|Y}; P_Y) \Big\} .
\end{aligned}$$

As in this case we only work with δ-typical x-sequences, we may choose δ to be any value, as long as it is strictly positive. Thus, we may force Q_X to be arbitrarily close to P_X by taking $\delta \to 0^+$. The error exponent in question thus becomes

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log \Pr(\mathcal{B}_{15}) \\
&\geq \max_{Q^\star_{U|X}\in\mathcal{P}(\mathcal{U})} \Big\{ R - I(P_X; Q^\star_{U|X}) \\
&+ I(P_Y; Q^\star_{U|Y}) + \min_{Q_Y\in\mathcal{P}(\mathcal{Y})} \min_{Q_{UXY}\in\mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})} \mathcal{D}(Q_{UXY}\|P_{\bar{U}\bar{X}\bar{Y}}) \Big\} + \hat{\epsilon} \\
&= \max_{Q^\star_{U|X}\in\mathcal{P}(\mathcal{U})} \Big\{ R - I(P_X; Q^\star_{U|X}) + I(P_Y; Q^\star_{U|Y}) \Big\} + \hat{\epsilon}\ ,
\end{aligned}$$

with $\hat{\epsilon} \to 0$ as $\delta \to 0$. This, along with an analysis of the complementary error event similar to the one given for Proposition 1, completes the proof of Proposition 2.

Appendix C

Hypothesis Testing with Bidirectional Communication

C.1 Proof of Proposition 3

We start by describing the random construction of codebooks, as well as encoding and decision functions. By analyzing the asymptotic properties of such decision systems, we aim at implying a *feasibility (existence) result* of interactive functions and decision regions that satisfy, for any given $\epsilon, \varepsilon > 0$, the following inequalities:

$$\frac{1}{n}\log\big(|f_{[1]}||g_{[1]}|\big) \leq I(U;X) + I(V;Y|U) + \varepsilon \ , \ \alpha_n(R\,|K=1) \leq \epsilon \ , \tag{C.1}$$

$$-\frac{1}{n}\log\beta_n(R,\epsilon\,|K=1) \geq \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}\in\mathscr{L}(U,V)} \mathcal{D}\big(P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}||P_{\bar{U}\bar{V}\bar{X}\bar{Y}}\big) - \varepsilon \ , \tag{C.2}$$

provided that n is large enough and for any given pair of random variables $(U,V) \in \mathscr{S}(R)$, where $|f_{[1]}|$ and $|g_{[1]}|$ denote the number of codewords in the codebooks used for interaction (note that *feasibility* is defined in the information-theoretic sense which implies the *random existence* of interactive and decision functions with desired properties).

Codebook generation. Without loss of generality, we assume that node A is the first to communicate. Fix a conditional probability $P_{UV|XY}(u,v|x,y) = P_{U|X}(u|x)P_{V|UY}(v|u,y)$ that attains the maximum in Proposition 3. Let

$$P_U(u) \equiv \sum_{x\in\mathcal{X}} P_{U|X}(u|x)P_X(x) \ , \ P_{V|U}(v|u) \equiv \sum_{y\in\mathcal{Y}} P_{V|UY}(v|u,y)P_Y(y). \tag{C.3}$$

For this choice of RVs, set the rates (R_U, R_V) to be

$$I(U;X) + \epsilon(\delta) \coloneqq R_U \ , \ I(V;Y|U) + \epsilon(\delta') \coloneqq R_V \tag{C.4}$$

with $\epsilon(\delta) \to 0$ as $\delta \to 0$. By the definition of the set $\mathscr{S}(R)$, it is clear that $R_U + R_V \leq R + \epsilon(\delta) + \epsilon(\delta')$. Randomly and independently draw 2^{nR_U} sequences $\mathbf{u} = (u_1, \ldots, u_n)$,

each according to $\prod_{i=1}^{n} P_U(u_i)$. Index these sequences by $m_U \in [1 : M_U \coloneqq 2^{nR_U}]$ to form the random codebook $\mathcal{C}_{\mathbf{u}} \coloneqq \big\{\mathbf{u}(m_U) : m_U \in [1 : M_U]\big\}$. As a second step, for each word $\mathbf{u} \in \mathcal{C}_{\mathbf{u}}$, build a codebook $\mathcal{C}_{\mathbf{v}}(m_U)$ by randomly and independently drawing 2^{nR_V} sequences $\mathbf{v}$, each according to $\prod_{i=1}^{n} P_{V|U}(v_i|u_i(m_U))$. Index these sequences by $m_V \in [1 : M_V \coloneqq 2^{nR_V}]$ to form the collection of codebooks $\mathcal{C}_{\mathbf{v}}(m_U) \coloneqq \big\{\mathbf{v}(m_U, m_V) : m_V \in [1 : M_V]\big\}$ for $m_U \in [1 : M_U]$.

Encoding and decision mappings. Given a sequence $\mathbf{x}$, node A searches in the codebook $\mathcal{C}_{\mathbf{u}}$ for an index m_U such that $(\mathbf{u}(m_U), \mathbf{x}) \in \mathcal{T}^n_{[UX]_\delta}$ (note that this notation denotes the δ-typical set with relation to the probability measure implied by H_0). If no such index is found, node A declares H_1. If more than one sequence is found, node A chooses one at random. Node A then communicates the chosen index m_U to node B, using a portion R_U bits of the available exchange rate. Upon receiving the index m_U, node B checks if $(\mathbf{u}(m_U), \mathbf{y}) \in \mathcal{T}^n_{[UY]_{\delta'}}$. If not, node B declares H_1. If the received sequence $\mathbf{u}$ and $\mathbf{y}$ (the observed sequence at node B) are jointly typical, node B looks in the specific codebook $\mathcal{C}_{\mathbf{v}}(m_U)$, for an index m_V such that $\big(\mathbf{u}(m_U), \mathbf{v}(m_U, m_V), \mathbf{y}\big) \in \mathcal{T}^n_{[UVY]_{\delta'}}$. If such an index is not found, node B declares H_1. If node B finds more than one such index, it chooses one of them at random. Node B then transmits the chosen index m_V to node A. Upon reception of the index m_V, node A checks if $\big(\mathbf{u}(m_U), \mathbf{v}(m_U, m_V), \mathbf{x}\big) \in \mathcal{T}^n_{[UVX]_{\delta''}}$. If so, it declares H_0 and otherwise, it declares H_1. The relation between δ, δ' and δ'' can be deducted from Lemma 7. It is, however, important to emphasize that $\delta'(\delta) \to 0$ as $\delta \to 0$, and $\delta''(\delta') \to 0$ as $\delta' \to 0$ with $n \to \infty$.

Analysis of α_n (Type I). The analysis of α_n is identical to the one proposed in [20], for the case of testing against independence. We give here a short summary of the analysis available in [20]. Assuming that the measure that controls X and Y is P_{XY}, and denoting the chosen indices at nodes A and B by m_U and m_V respectively, the error probability of the Type I can be expressed as follows

$$\alpha_n \equiv \Pr(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) \leq \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_1^c \cap \mathcal{E}_2) + \Pr(\mathcal{E}_1^c \cap \mathcal{E}_2^c \cap \mathcal{E}_3) \ , \tag{C.5}$$

where $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ represent the following error events:

$$\mathcal{E}_1 \equiv \big\{(\mathbf{U}(m_U), \mathbf{X}) \notin \mathcal{T}^n_{[UX]\delta} \, \forall \, m_U \in [1 : M_U]\big\} \ , \tag{C.6a}$$

$$\mathcal{E}_2 \equiv \big\{(\mathbf{V}(m_U, m_V), \mathbf{U}(m_U), \mathbf{Y}) \notin \mathcal{T}^n_{[VUY]_{\delta'}} \, \forall \, m_V \in [1 : M_V] \text{ and the specific } m_U \text{ selected at node } A\big\} \ , \tag{C.6b}$$

$$\mathcal{E}_3 \equiv \big\{(\mathbf{V}(m_U, m_V), \mathbf{U}(m_U), \mathbf{X}) \notin \mathcal{T}^n_{[VUX]_{\delta''}}, \text{ for the specific } m_U \text{ and } m_V \text{ previously chosen}\big\} \ . \tag{C.6c}$$

Analyzing each of the probabilities in (C.5) separately, $\Pr(\mathcal{E}_1) \to 0$ as $n \to \infty$ by the *covering lemma* [105], provided that $R_U \geq I(U;X) + \epsilon(\delta)$, with $\epsilon(\delta) \to 0$ as $\delta \to 0$. $\Pr(\mathcal{E}_1^c \cap \mathcal{E}_2) \to 0$ when $n \to \infty$ by the *conditional typicality lemma* [105], in addition to the covering lemma, provided that $R_V \geq I(V;Y|U) + \epsilon(\delta')$. Finally, the third term in (C.5) can be shown to tend to zero through the use of the Markov lemma (see Lemma 8), as well as Lemma 6 and Lemma 7 in Chapter 2. Thus, as all three components tend to

zero with large n, we may conclude that $\alpha_n \leq \epsilon$ for any constraint $0 < \epsilon < 1$ and n large enough.

Analysis of β_n (Type II). The error probability of Type II is defined by

$$\beta_n(R, \epsilon \,|K = 1) \equiv \Pr\left(\text{decide } H_0 | XY \sim P_{\bar{X}\bar{Y}}\right) . \tag{C.7}$$

Thus, we assume that $P_{\bar{X}\bar{Y}}$ controls the measure of the observed RVs throughout this analysis. We use similar methods to what was done in [9], although we choose to work with random codebooks. The influence of this choice is on the analysis of α_n only, as seen above, and not on β_n.

For a given pair of sequences $(\mathbf{x}, \mathbf{y})$ with type variables $X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$, we count all possible events that lead to an error. We notice first, that given a pair of vectors $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ the probability that these vectors will be the result of n i.i.d. draws, according to the measure implied by H_1, is given by Lemma 6 to be:

$$\Pr\{\bar{X}^n\bar{Y}^n = (\mathbf{x}, \mathbf{y})\} = \exp\left[-n\left(H(X^{(n)}Y^{(n)}) + \mathcal{D}(X^{(n)}Y^{(n)}||\bar{X}\bar{Y})\right)\right] , \tag{C.8}$$

where $X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ are the type variables of the realizations $(\mathbf{x}, \mathbf{y})$ (see Chapter 2). For each pair of codewords $\mathbf{u}_i \in \mathcal{C}_\mathbf{u}$ and $\mathbf{v}_{ij} \in \mathcal{C}_\mathbf{v}(i)$, we define the set:

$$\mathcal{S}_{ij}(\mathbf{x}) \coloneqq \{\mathbf{u}_i\} \times \{\mathbf{v}_{ij}\} \times \mathcal{G}_{ij} \times \{\mathbf{x}\} , \tag{C.9}$$

where $\mathcal{G}_{ij} \subseteq \mathcal{Y}^n$ is the set of all vectors $\mathbf{y}$ that, given the received message $\mathbf{u}_i$, will result in the message $\mathbf{v}_{ij}$ being transmitted back to node A. Denoting by $K_{ij}(\mathbf{x})$ the number of elements $(\mathbf{u}_i, \mathbf{v}_{ij}, \mathbf{x}, \mathbf{y}) \in \mathcal{S}_{ij}(\mathbf{x})$ whose type variables coincide with $U^{(n)}V^{(n)}X^{(n)}Y^{(n)}$, we have by Lemma 5 that:

$$K_{ij}(\mathbf{x}) \leq \exp\left[nH(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)})\right] . \tag{C.10}$$

Let $K(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ denote the number of all elements:

$$(\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y}) \in \mathscr{S}_n \coloneqq \bigcup_{i=1}^{M_U}\bigcup_{j=1}^{M_V} \bigcup_{\mathbf{x} \in \mathcal{T}^n_{[X|UV]\delta''}(\mathbf{u}_i\mathbf{v}_{ij})} \mathcal{S}_{ij}(\mathbf{x})$$

that have type variable $U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathcal{U} \times \mathcal{V} \times \mathcal{X} \times \mathcal{Y})$, then

$$\begin{aligned} K(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) &\leq \sum_{i=1}^{M_U}\sum_{j=1}^{M_V} \exp\left[nH(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)})\right]\left|\mathcal{T}^n_{[X|UV]\delta''}(\mathbf{u}_i\mathbf{v}_{ij})\right| \\ &\leq \exp\Big[n\Big(H(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)}) \\ &\qquad + I(U;X) + I(V;Y|U) + H(X|UV) + \mu_n\Big)\Big] , \end{aligned} \tag{C.11}$$

where M_U and M_V are the sizes of the codebooks $\mathcal{C}_\mathbf{u}$ and $\mathcal{C}_\mathbf{v}(\cdot)$. The first and second additional terms in the final expression come from the size of the codebooks and the third

is a bound over the size of the delta-typical set (see Lemma 9). The resulting sequence μ_n is a function of $\delta, \delta', \delta''$ that complies with $\mu_n \to 0$ as $n \to \infty$. The error probability of Type II satisfies:

$$\beta_n(R, \epsilon \,|K=1) \leq \sum_{U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathscr{S}_n} \exp\left[-n\left(k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) - \mu_n\right)\right] \ , \tag{C.12}$$

where the function $k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ is defined by

$$\begin{aligned} k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) \coloneqq\ & H(X^{(n)}Y^{(n)}) + \mathcal{D}(X^{(n)}Y^{(n)}||\bar{X}\bar{Y}) \\ & - H(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)}) - H(X|UV) \\ & - I(U;X) - I(V;Y|U) \ . \end{aligned} \tag{C.13}$$

Note that we deliberately made an abuse of notation in (C.12) to indicate that the sum is taken over all possible type-variables $U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathcal{U}\times\mathcal{V}\times\mathcal{X}\times\mathcal{Y})$ formed by empirical probability measures from elements $(\mathbf{u},\mathbf{v},\mathbf{x},\mathbf{y}) \in \mathscr{S}_n$.

From the construction of $\mathscr{S}_n$, it is clear that if $(\mathbf{u},\mathbf{v},\mathbf{x},\mathbf{y}) \in \mathscr{S}_n$, then at least $(\mathbf{u},\mathbf{v},\mathbf{x}) \in \mathcal{T}^n_{[UVX]_{\delta''}}$ and $(\mathbf{u},\mathbf{v},\mathbf{y}) \in \mathcal{T}^n_{[UVY]_{\delta'}}$. Thus, the summation in (C.12) is only over all types satisfying:

$$\begin{aligned} &|P_{U^{(n)}V^{(n)}X^{(n)}}(u,v,x) - P_{UVX}(u,v,x)| \leq \delta'' \ , \\ &|P_{U^{(n)}V^{(n)}Y^{(n)}}(u,v,y) - P_{UVY}(u,v,y)| \leq \delta' \ , \end{aligned} \tag{C.14}$$

for all $(u,v,x) \in \operatorname{supp}(P_{UVX})$ and $(u,v,y) \in \operatorname{supp}(P_{UVY})$. In addition, it follows by Lemma 4 from the total number of types of length n that:

$$\begin{aligned} \beta_n(R, \epsilon \,|K=1) \leq\ & (n+1)^{|\mathcal{U}||\mathcal{V}||\mathcal{X}||\mathcal{Y}|} \\ & \times \max_{U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathscr{S}_n} \exp\left[-n\left(k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) - \mu_n\right)\right] \ . \end{aligned} \tag{C.15}$$

By (C.14) and the continuity of the entropy function as well as the KL divergence, we can conclude that

$$\begin{aligned} k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) = & H(\tilde{X}\tilde{Y}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - H(\tilde{Y}|\tilde{U}\tilde{V}\tilde{X}) \\ & - H(\tilde{X}|\tilde{U}\tilde{V}) - I(\tilde{U};\tilde{X}) - I(\tilde{V};\tilde{Y}|\tilde{U}) + \mu'_n \ , \end{aligned} \tag{C.16}$$

with $\tilde{U}\tilde{V}\tilde{X}\tilde{Y} \in \mathscr{L}(U,V)$ and $\mu'_n \to 0$ when $n \to \infty$. We can further simplify the expression of $k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ by observing that:

$$k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) = H(\tilde{X}\tilde{Y}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - H(\tilde{Y}|\tilde{U}\tilde{V}\tilde{X}) \tag{C.17a}$$

$$- H(\tilde{X}|\tilde{U}\tilde{V}) - I(\tilde{U};\tilde{X}) - I(\tilde{V};\tilde{Y}|\tilde{U}) + \mu'_n \tag{C.17b}$$

$$= H(\tilde{X}\tilde{Y}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - H(\tilde{X}\tilde{Y}|\tilde{U}\tilde{V}) - I(\tilde{U};\tilde{X}) - I(\tilde{V};\tilde{Y}|\tilde{U}) + \mu'_n \tag{C.17c}$$

$$= I(\tilde{X}\tilde{Y};\tilde{U}\tilde{V}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - I(\tilde{U};\tilde{X}) - I(\tilde{V};\tilde{Y}|\tilde{U}) + \mu'_n \tag{C.17d}$$

$$= I(\tilde{X}\tilde{Y};\tilde{U}) + I(\tilde{X}\tilde{Y};\tilde{V}|\tilde{U}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - I(\tilde{U};\tilde{X}) - I(\tilde{V};\tilde{Y}|\tilde{U}) + \mu'_n \tag{C.17e}$$

$$= \mathcal{D}(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) + I(\tilde{X}\tilde{Y};\tilde{V}|\tilde{U}) - I(\tilde{Y};\tilde{V}|\tilde{U}) + \mu'_n \tag{C.17f}$$

$$= \mathcal{D}(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) + I(\tilde{X};\tilde{V}|\tilde{U}\tilde{Y}) + \mu'_n \ , \tag{C.17g}$$

where equality (C.17f) stems from the identity (see [9]):

$$I(\tilde{X}\tilde{Y};\tilde{U}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) - I(\tilde{U};\tilde{X}) = I(\tilde{U};\tilde{Y}|\tilde{X}) + \mathcal{D}(\tilde{X}\tilde{Y}||\bar{X}\bar{Y}) \tag{C.18}$$

$$= \mathcal{D}(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) \ . \tag{C.19}$$

Note that the following Markov chain: $X - (U,Y) - V$ holds under both hypotheses (i.e., the same chain can be written with a bar over all variables), *but not* for the auxiliary RVs, marked with a tilde.

Finally, we conclude our development of $k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ as follows:

$$k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) = \mathcal{D}(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) + I(\tilde{X};\tilde{V}|\tilde{U}\tilde{Y}) + \mu'_n \tag{C.20a}$$

$$= \sum_{\forall(u,v,x,y)} P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)\times$$

$$\times \log\left(\frac{P_{\tilde{U}\tilde{X}\tilde{Y}}(u,x,y)}{P_{\bar{U}\bar{X}\bar{Y}}(u,x,y)}\frac{P_{\tilde{X}\tilde{V}|\tilde{U}\tilde{Y}}(x,v|u,y)}{P_{\tilde{X}|\tilde{U}\tilde{Y}}(x|u,y)P_{\tilde{V}|\tilde{U}\tilde{Y}}(v|u,y)}\right) + \mu'_n \tag{C.20b}$$

$$= \sum P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)\times$$

$$\times \log\left(\frac{P_{\tilde{U}\tilde{X}\tilde{Y}}(u,x,y)}{P_{\bar{U}\bar{X}\bar{Y}}(u,x,y)}\frac{\cancel{P_{\tilde{X}|\tilde{U}\tilde{Y}}(x|u,y)}P_{\tilde{V}|\tilde{U}\tilde{X}\tilde{Y}}(v|u,x,y)}{\cancel{P_{\tilde{X}|\tilde{U}\tilde{Y}}(x|u,y)}P_{\tilde{V}|\tilde{U}\tilde{Y}}(v|u,y)}\right) + \mu'_n \tag{C.20c}$$

$$= \sum_{\forall(u,v,x,y)} P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)\log\left(\frac{P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)}{P_{\bar{U}\bar{X}\bar{Y}}(u,x,y)P_{\tilde{V}|\bar{U}\bar{Y}}(v|u,y)}\right) + \mu'_n \tag{C.20d}$$

$$= \sum_{\forall(u,v,x,y)} P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)\log\left(\frac{P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}(u,v,x,y)}{P_{\bar{U}\bar{V}\bar{X}\bar{Y}}(u,v,x,y)}\right) + \mu'_n \tag{C.20e}$$

$$= \mathcal{D}(\tilde{U}\tilde{V}\tilde{X}\tilde{Y}||\bar{U}\bar{V}\bar{X}\bar{Y}) + \mu'_n \ , \tag{C.20f}$$

where the sums are over the $\text{supp}(P_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}})$; and (C.20d) is due to the definition of the set $\mathscr{L}(U,V)$ that implies $P_{\tilde{V}|\tilde{U}\tilde{Y}}(v|u,y) = P_{V|UY}(v|u,y)$. In addition, as coding (at each side) is performed before a decision is made, it is clear it is done in the same way under both hypotheses. Thus, while $P_{UVY}(u,v,y) \neq P_{\bar{U}\bar{V}\bar{Y}}(u,v,y)$, it is true that $P_{\bar{V}|\bar{U}\bar{Y}}(v|u,y) = P_{V|UY}(v|u,y) = P_{\tilde{V}|\tilde{U}\tilde{Y}}(v|u,y)$. As μ_n, μ'_n are arbitrarily small, as a function of the choices of δ and δ' provided that n is large enough, this concludes the proof of Proposition 3.

C.2 Proof of Proposition 4

The proof of Proposition 4 is very similar to the one presented above for Proposition 3. Codebook construction, as well as encoding and decision mappings remain similar. At each round, a codebook is built based on any possible combination of the previous messages.

Given previous messages, each node chooses a message in the relevant codebook and communicates its index to the other statistician. The process continues until a message cannot be found, which is jointly typical with all previous messages as well as the observed sequence, in which case H_1 is declared. Otherwise, until the end of round K in which case H_0 is declared, provided that all the messages are jointly typical with the observed sequence. We next provide a sketch of the proof to this simple extension.

The analysis of α_n applies similarly to the previous case, as long as a finite number of rounds is considered. Regarding the analysis of β_n, the following important changes are needed:

- The set $\mathcal{S}_{\mathbf{ij}}(\mathbf{x})$ is now defined by using all exchanged messages:

$$\mathcal{S}_{\mathbf{ij}}(\mathbf{x}) \coloneqq \{\mathbf{u}_{[1],i_1}\} \times \{\mathbf{v}_{[1],i_1 j_1}\} \times \cdots \times \{\mathbf{u}_{[K],i_K}\} \times \{\mathbf{v}_{[K],i_K j_K}\} \times \mathcal{G}_{\mathbf{ij}} \times \{\mathbf{x}\} \ , \quad \text{(C.21)}$$

  where $(\mathbf{i},\mathbf{j}) \coloneqq (i_1,j_1),\ldots,(i_K,j_K)$ and $\mathbf{u}_{[k],i_k}$ is the i_k-th message in the codebook $\mathcal{C}_{\mathbf{u}_{[k]}}$, similarly for the other random variables.

- Similarly, $\mathscr{S}_n$ is now defined by the union over the codewords of *all* auxiliary RVs.

- The bound over $K_{\mathbf{ij}}$ (analogues to expression (C.10) before) writes:

$$K_{\mathbf{ij}}(\mathbf{x}) \leq \exp\left[nH\big(Y^{(n)}|U^{(n)}_{[1:K]}V^{(n)}_{[1:K]}X^{(n)}\big)\right] \ . \quad \text{(C.22)}$$

- Finally, $K\big(U^{(n)}_{[1:K]}V^{(n)}_{[1:K]}X^{(n)}Y^{(n)}\big)$, i.e., see (C.11), is now calculated through the summation over the codebooks of all messages, considering the cardinality of the conditional set: $\big|\mathcal{T}^n_{[X|\mathbf{u}_{[1:K]},\mathbf{i}\mathbf{v}_{[1:K]},\mathbf{ij}]_\delta}\big|$.

- As more steps are performed, each of which requires encoding, we also need to define new δ's for each of these steps. We refrain from this for the sake of readability, as all of these δ's go to 0 together, as was seen in the case of a single round.

Considering these differences, after k rounds of interactions, $k\big(U^{(n)}_{[1:k]}V^{(n)}_{[1:k]}\big)$ can be shown to be equal to (e.g. see (C.13)):

$$\begin{aligned} k\big(U^{(n)}_{[1:k]}V^{(n)}_{[1:k]}\big) &= \mathcal{D}\big(P_{\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}\tilde{Y}}||P_{\tilde{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}\bar{Y}}\big) \\ &+ I(\tilde{Y};\tilde{U}_{[k]}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}) + I(\tilde{X};\tilde{V}_{[k]}|\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{Y}) + \mu'_n \ . \end{aligned} \quad \text{(C.23)}$$

By continuing in the same manner as in (C.20a), we show:

$$\begin{aligned} k\big(U^{(n)}_{[1:k]}V^{(n)}_{[1:k]}\big) - \mu'_n &= \sum_{\forall} P_{\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}\tilde{Y}} \log \frac{P_{\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}\tilde{Y}}}{P_{\tilde{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}\bar{Y}}} \quad \text{(C.24a)} \\ + \sum_{\forall} P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{X}\tilde{Y}} \log &\frac{P_{\tilde{U}_{[k]}\tilde{Y}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}}}{P_{\tilde{U}_{[k]}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}} P_{\tilde{Y}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}}} \end{aligned}$$

$$+\sum_{\forall} P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}} \log \frac{P_{\tilde{V}_{[k]}\tilde{X}|\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{Y}}}{P_{\tilde{V}_{[k]}|\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{Y}} P_{\tilde{X}|\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{Y}}}$$

$$= \sum_{\forall} P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}} \log \left[\frac{P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}}}{P_{\bar{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}\bar{Y}} P_{\tilde{U}_{[k]}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}} P_{\tilde{V}_{[k]}|\tilde{U}_{[1:k]}\tilde{V}_{[1:k-1]}\tilde{Y}}} \right] \tag{C.24b}$$

$$= \sum_{\forall} P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}} \log \left[\frac{P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}}}{P_{\bar{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}\bar{Y}} P_{\bar{U}_{[k]}|\bar{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}} P_{\bar{V}_{[k]}|\bar{U}_{[1:k]}\bar{V}_{[1:k-1]}\bar{Y}}} \right] \tag{C.24c}$$

$$= \sum_{\forall} P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}} \log \left[\frac{P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}}}{P_{\bar{U}_{[1:k]}\bar{V}_{[1:k]}\bar{X}\bar{Y}}} \right] \tag{C.24d}$$

$$= \mathcal{D}\big(P_{\tilde{U}_{[1:k]}\tilde{V}_{[1:k]}\tilde{X}\tilde{Y}} || P_{\bar{U}_{[1:k]}\bar{V}_{[1:k]}\bar{X}\bar{Y}}\big) \ , \tag{C.24e}$$

where all sums are over all the alphabets of the relevant RVs. Here, (C.24c), much like in the case of single-round exchange above, is due to the definition of the set $\mathscr{L}(U_{[1:k]}, V_{[1:k]})$ and to the fact that encoding occurs without knowledge of the PM controlling the RVs, and thus behaves the same under each of the hypotheses. Thus,

$$P_{\tilde{U}_{[k]}|\tilde{U}_{[1:k-1]}\tilde{V}_{[1:k-1]}\tilde{X}} = P_{U_{[k]}|U_{[1:k-1]}V_{[1:k-1]}X} = P_{\bar{U}_{[k]}|\bar{U}_{[1:k-1]}\bar{V}_{[1:k-1]}\bar{X}},$$

and similarly for the messages $V_{[k]}$ at node B. Pursuing this until round K, the proposition is proved.

C.3 Proof of Converse for Theorem 3

In this appensix, we complete the proof of Theorem 3 by proving a weak unfeasibility (converse) property. We start by proposing a *multi-letter* expression that constitutes an upper bound over performance in this case, as summarized in the following lemma:

Lemma 23 (Multi-letter representation for testing against independence with $K = 1$ [20])**.** *The error exponent to the error probability of Type II for testing against independence with one round satisfies:*

$$\lim_{\epsilon\to 0} \liminf_{n\to\infty} -\frac{1}{n} \log \beta_n(R, \epsilon | K = 1) \leq \frac{1}{n}\big[I(I_A; \mathbf{Y}^n) + I(I_B; \mathbf{X}^n | I_A)\big] \ , \tag{C.25}$$

$$R \geq \frac{1}{n}\big[I(I_A; \mathbf{X}^n) + I(I_B; \mathbf{Y}^n | I_A)\big] \ , \tag{C.26}$$

where $I_A \coloneqq f_1(\mathbf{X}^n)$ and $I_B \coloneqq g_1\big(f_1(\mathbf{X}^n), \mathbf{Y}^n\big)$ for any mappings (f_1, g_1), as given in Defintion 13.

Proof. For block-length n, given a code characterized by the encoding mappings $f_{[1]}, g_{[1]}$ at nodes A and B respectively, and a decoding mapping ϕ at node A, let the acceptance region be denoted by

$$\mathcal{A}_n \coloneqq \big\{(\mathbf{x}, j) \in \mathcal{X}^n \times \{1, \ldots, |g_{[1]}|\} : g_{[1]}\big(\mathbf{y}, f_{[1]}(\mathbf{x})\big) = j,\ \mathbf{y} \in \mathcal{Y}^n,\ \phi\big(\mathbf{x}, j\big) = 0\big\} \ . \tag{C.27}$$

Let P and Q denote the probability measures on $\mathcal{X}^n \times \{1, \ldots, |g_{[1]}|\}$ induced by H_0 and H_1, respectively. From the *log-sum inequality* [82], we have:

$$\begin{aligned} &\mathcal{D}\left(P_{\mathbf{X}^n I_A I_B} \| Q_{\mathbf{X}^n I_A I_B}\right) = \mathcal{D}\left(P_{\mathbf{X}^n I_B} \| Q_{\mathbf{X}^n I_B}\right) \\ &\geq (1-\alpha_n) \log \frac{1-\alpha_n}{\beta_n(R, \epsilon \,|K=1)} + \alpha_n \log \frac{\alpha_n}{1-\beta_n(R, \epsilon \,|K=1)} \ , \end{aligned} \tag{C.28}$$

where $I_A \coloneqq f_{[1]}(\mathbf{X}^n)$, $I_B \coloneqq g_{[1]}(I_A, \mathbf{Y}^n)$, $\alpha_n(R|K=1) \coloneqq P(\mathcal{A}_n^c) \leq \epsilon$ and $\beta_n(R, \epsilon \,|K=1) \coloneqq Q(\mathcal{A}_n)$. Through some algebra this yields:

$$\mathcal{D}\left(P_{\mathbf{X}^n I_A I_B} \| Q_{\mathbf{X}^n I_A I_B}\right) \geq (1-\alpha_n) \log \frac{1}{\beta_n(R, \epsilon \,|K=1)} - H_2(\alpha_n) \ , \tag{C.29}$$

where $H_2(\rho) \coloneqq -\rho \log \rho - (1-\rho)\log(1-\rho)$ is the *binary entropy* function. By assumption $\epsilon \to 0$ as $n \to \infty$, one concludes that for n large enough

$$-\frac{1}{n} \log \beta_n(R, \epsilon \,|K=1) \leq \frac{1}{n} \mathcal{D}\left(P_{\mathbf{X}^n I_A I_B} \| Q_{\mathbf{X}^n I_A I_B}\right) - \delta_n \ , \tag{C.30}$$

with $\delta_n \to 0$ as $n \to \infty$. Using the chain rule, we continue to get:

$$\begin{aligned} \mathcal{D}\big(P_{\mathbf{X}^n I_A I_B} \| Q_{\mathbf{X}^n I_A I_B}\big) &= I(I_B; \mathbf{X}^n | I_A) + \mathcal{D}\left(P_{I_B|I_A} \| Q_{I_B|I_A} | P_{I_A}\right) && \text{(C.31a)} \\ &\leq I(I_B; \mathbf{X}^n | I_A) + \mathcal{D}\left(P_{\mathbf{Y}^n I_A I_B} \| Q_{\mathbf{Y}^n I_A I_B}\right) && \text{(C.31b)} \\ &= I(I_B; \mathbf{X}^n | I_A) + \mathcal{D}\left(P_{\mathbf{Y}^n I_A} \| P_Y^n P_{I_A}\right) && \text{(C.31c)} \\ &= I(I_B; \mathbf{X}^n | I_A) + I(I_A; \mathbf{Y}^n) \ . && \text{(C.31d)} \end{aligned}$$

Here, (C.31a) and (C.31b) stem from the chain rule for the KL-divergence, and (C.31c) is due to the fact that we consider the case of testing against independence. With this, the *weak unfeasibility* proof of a multi-letter expression is completed. □

From Lemma 23, it follows that:

$$\begin{aligned} &\lim_{\epsilon \to 0} \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n(R, \epsilon | K=1) \\ &\qquad \leq \limsup_{n \to \infty} \frac{1}{n} [I(I_A; \mathbf{Y}^n) + I(I_B; \mathbf{X}^n | I_A)] \coloneqq \limsup_{n \to \infty} \Delta_n \ , \end{aligned} \tag{C.32}$$

where I_A is the message sent from node A while I_B is its reply from node B. In order to derive a single-letter expression, we expand (C.32) as follows:

$$\begin{aligned} \Delta_n &= \frac{1}{n} \sum_{i=1}^{n} \left[I(I_A; Y_i | \mathbf{Y}_{i+1}^n) + I(I_B; X_i | I_A \mathbf{X}^{i-1}) \right] && \text{(C.33a)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[I(I_A \mathbf{Y}_{i+1}^n; Y_i) + I(I_B \mathbf{Y}_{i+1}^n; X_i | I_A \mathbf{X}^{i-1}) - I(\mathbf{Y}_{i+1}^n; X_i | I_A I_B \mathbf{X}^{i-1}) \right] && \text{(C.33b)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[I(I_A \mathbf{X}^{i-1} \mathbf{Y}_{i+1}^n; Y_i) - I(\mathbf{X}^{i-1}; Y_i | I_A \mathbf{Y}_{i+1}^n) + I(\mathbf{Y}_{i+1}^n; X_i | I_A \mathbf{X}^{i-1}) \right. && \text{(C.33c)} \end{aligned}$$

$$+I(I_B;X_i|I_A\mathbf{X}^{i-1}\mathbf{Y}_{i+1}^n)-I(\mathbf{Y}_{i+1}^n;X_i|I_AI_B\mathbf{X}^{i-1})\big]$$
$$=\frac{1}{n}\sum_{i=1}^{n}\Big[I(\hat{U}_i;Y_i)+I(V_i;X_i|\hat{U}_i)-I(\mathbf{Y}_{i+1}^n;X_i|I_AI_B\mathbf{X}^{i-1})\Big]\ , \tag{C.33d}$$

where $\mathbf{X}^i$ denotes the first i samples and $\mathbf{X}_i^n=(X_i,\ldots,X_n)$; (C.33a) stems from the chain rule and (C.33b) from the assumed i.i.d. nature of the sources. In (C.33d), the following identity is used [82]:

$$\sum_{i=1}^{n}I(\mathbf{A}^{i-1};B_i|C,\mathbf{B}_{i+1}^n)=\sum_{i=1}^{n}I(\mathbf{B}_{i+1}^n;A_i|C,\mathbf{A}^{i+1})\ , \tag{C.34}$$

where C can be arbitrarily dependent to the vectors $\mathbf{A}$ and $\mathbf{B}$, as long as it does not change with i, and the following auxiliary RVs are defined on measurable spaces $(\mathcal{U}_i\times\mathcal{V}_i,\mathcal{B}_{\mathcal{U}_i\times\mathcal{V}_i})$ by setting:

$$\hat{U}_i\coloneqq(I_A,\mathbf{X}^{i-1},\mathbf{Y}_{i+1}^n)\quad\text{and}\quad V_i\coloneqq I_B\ ,\ \forall\, i=[1:n]\ . \tag{C.35}$$

It is important to emphasize that the required Markov chains in (4.17) are verified for each $i=[1:n]$. this is proved at the end of this appendix. Let Q be a RV uniformly distributed over $[1:n]$, then:

$$\begin{aligned}\Delta_n&\le I(\hat{U}_Q;Y_Q|Q)+I(V_Q;X_Q|\hat{U}_Q,Q)-\frac{1}{n}\sum_{i=1}^{n}I(\mathbf{Y}_{i+1}^n;X_i|I_AI_B\mathbf{X}^{i-1})\\&\triangleq I(U;Y)+I(V;X|U)-T\ ,\end{aligned} \tag{C.36}$$

where $U\coloneqq(\hat{U}_Q,Q)$.

We now bound the required rate, from the size of the mappings, we have

$$nR\ge I(I_A;\mathbf{X}^n)+I(I_B;\mathbf{Y}^nI_A)\ge I(I_A;\mathbf{X}^n)+I(I_B;\mathbf{Y}^n|I_A)\ . \tag{C.37}$$

For convenience, we analyze each of these terms separately:

$$I(I_A;\mathbf{X}^n)=\sum_{i=1}^{n}I(I_A\mathbf{X}^{i-1};X_i) \tag{C.38a}$$
$$=\sum_{i=1}^{n}\big[I(I_A\mathbf{X}^{i-1}\mathbf{Y}_{i+1^n};X_i)-I(\mathbf{Y}_{i+1}^n;X_i|I_A\mathbf{X}^{i-1})\big]\ , \tag{C.38b}$$

where (C.38a) is due to the i.i.d nature of samples. The second term writes as:

$$\begin{aligned}I(I_B;\mathbf{Y}^n|I_A)&=\sum_{i=1}^{n}\big[I(I_B\mathbf{X}^{i-1};Y_i|I_A\mathbf{Y}_{i+1}^n)-I(\mathbf{X}^{i-1};Y_i|I_AI_B\mathbf{Y}_{i+1}^n)\big]\\&=\sum_{i=1}^{n}\big[I(\mathbf{X}^{i-1};Y_i|I_A\mathbf{Y}_{i+1}^n)+I(I_B;Y_i|I_A\mathbf{X}^{i-1}\mathbf{Y}_{i+1}^n)-I(\mathbf{X}^{i-1};Y_i|I_AI_B\mathbf{Y}_{i+1}^n)\big]\\&=\sum_{i=1}^{n}\big[I(I_B;Y_i|I_A\mathbf{X}^{i-1}\mathbf{Y}_{i+1}^n)+I(X_i;\mathbf{Y}_{i+1}^n|I_A\mathbf{X}^{i-1})-I(\mathbf{X}^{i-1};Y_i|I_AI_B\mathbf{Y}_{i+1}^n)\big]\ ,\end{aligned} \tag{C.39}$$

where the final step is due to identity (C.34). These inequalities lead to

$$nR \geq \sum_{i=1}^{n} \left[I(I_A \mathbf{X}^{i-1} \mathbf{Y}_{i+1}^{n}; X_i) + I(I_B; Y_i | I_A \mathbf{X}^{i-1} \mathbf{Y}_{i+1}^{n}) - I(\mathbf{X}^{i-1}; Y_i | I_A I_B \mathbf{Y}_{i+1}^{n}) \right] . \quad \text{(C.40)}$$

Using the same definitions for the auxiliary RVs as above, this result can be expressed as follows:

$$R \geq I(\hat{U}_Q; X_Q | Q) + I(V_Q; Y_Q | \hat{U}_Q, Q) - T , \quad \text{(C.41)}$$

and thus, the following region is an outer bound:

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) - T , \\ R \geq I(U;X) + I(V;Y|U) - T , \end{cases} \quad \text{(C.42)}$$

where (U, V) are auxiliary RVs that respect the required Markov chains in (4.17). It is left to show that (C.42) is equivalent or stricter than:

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) , \\ R \geq I(U;X) + I(V;Y|U) . \end{cases} \quad \text{(C.43)}$$

That is, all pairs (R, Δ_n) that are forbidden in the region in (C.42) are also forbidden in (C.43). In order to do so we use *Fourier-Motzkin* elimination [106] over $T \geq 0$.

By removing T, we get:

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) , \\ R \geq I(U;X) + I(V;Y|U) - I(U;Y) - I(V;X|U) + \Delta_n , \end{cases} \quad \text{(C.44)}$$

and using the Markovian relations between the different RVs we obtain:

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) , \\ R \geq I(U;X|Y) + I(V;Y|UX) + \Delta_n . \end{cases} \quad \text{(C.45)}$$

In order to show the equivalence between the two regions, we need to check the extremal points. The point where $\Delta_n = 0$ is trivial, as $R = 0$ is optimal under both regions. When checking $\Delta_n = I(U;Y) + I(V;X|U)$ we have:

$$\begin{aligned} R &\geq I(U;X|Y) + I(V;Y|UX) + I(U;Y) + I(V;X|U) \\ &= I(U;X) + I(V;Y|U) , \end{aligned} \quad \text{(C.46)}$$

which completes the proof of the weak unfeasibility.

C.3.1 Proving the Required Markov Chains

In order to complete the proof of converse, we need to show that the required Markov chains are indeed respected. Two Markov chains are necessary:

$$\begin{cases} \hat{U}_i - X_i - Y_i , \ \forall\, i = [1:n] \\ V_i - (\hat{U}_i, Y_i) - X_i , \ \forall\, i = [1:n]. \end{cases} \quad \text{(C.47)}$$

Using the chosen RVs from (C.35), these Markov chains are represented by

$$\begin{cases} (I_A, \mathbf{X}^{i-1}, \mathbf{Y}_{i+1}^n) - X_i - Y_i \ , \ \forall\, i = [1:n] \\ I_B - (I_A, \mathbf{X}^{i-1}, \mathbf{Y}_i^n) - X_i \ , \ \forall\, i = [1:n]. \end{cases} \tag{C.48}$$

In order to check this, we use the next lemma.

Lemma 24. *Let A_1, A_2, B_1, B_2 be RVs with joint probability measure $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$ and assume that $\{f^i\}_{i=1}^k, \{g^i\}_{i=1}^k$ are any collection of P-measurable mappings with domain structure given by:*

$$f^1(A_1, A_2); f^2(A_1, A_2, g^1); \ldots; f^k(A_1, A_2, g^1, \ldots, g^{k-1}) \ , \tag{C.49}$$

$$g^1(B_1, B_2, f^1); g^2(B_1, B_2, f^1, f^2); \ldots; g^k(B_1, B_2, f^1, \ldots, f^k) \ . \tag{C.50}$$

Then,

$$I(A_2; B_1 | f^1, f^2, \ldots, f^k, g^1, g^2, \ldots, g^k, A_1, B_2) = 0 \ . \tag{C.51}$$

Proof. Refer to reference [50, Lemma 1]. □

In order to prove the first Markov chain, we simply let:

$$\begin{cases} A_1 \coloneqq X_i, \quad B_1 \coloneqq Y_i \ , \\ A_2 \coloneqq (\mathbf{X}^{i-1}, \mathbf{X}_{i+1}^n, \mathbf{Y}_{i+1}^n) \ , \quad B_2 \coloneqq \mathbf{Y}^{i-1} \ . \end{cases} \tag{C.52}$$

It can be easily verified that $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$, which stems directly from the i.i.d. nature of the samples. Thus, according to Lemma 24:

$$\begin{aligned} 0 &= I(\mathbf{X}^{i-1}\mathbf{X}_{i+1}^n\mathbf{Y}_{i+1}^n; Y_i | X_i\mathbf{Y}^{i-1}) \\ &= I(\mathbf{X}^{i-1}\mathbf{X}_{i+1}^n\mathbf{Y}^{i-1}\mathbf{Y}_{i+1}^n; Y_i | X_i) - I(\mathbf{Y}^{i-1}; Y_i | X_i) \\ &= I(\mathbf{X}^{i-1}\mathbf{X}_{i+1}^n\mathbf{Y}^{i-1}\mathbf{Y}_{i+1}^n; Y_i | X_i) \ , \end{aligned} \tag{C.53}$$

which shows the Markov chain:

$$(\mathbf{X}^{i-1}, \mathbf{X}_{i+1}^n, \mathbf{Y}^{i-1}, \mathbf{Y}_{i+1}^n) - X_i - Y_i \ , \ \forall\, i = [1:n]. \tag{C.54}$$

As $I_A \coloneqq f_{[1]}(\mathbf{X}^n)$, the following Markov chain is also true:

$$(I_A, \mathbf{X}^{i-1}, \mathbf{Y}_{i+1}^n) - X_i - Y_i \ , \ \forall\, i = [1:n] \tag{C.55}$$

which proves the first Markov chain in (C.48).

As for the second one, we let:

$$\begin{cases} A_1 \coloneqq \mathbf{X}^{i-1} \ , \quad B_1 \coloneqq \mathbf{Y}^{i-1} \ , \\ A_2 \coloneqq (X_i, \mathbf{X}_{i+1}^n) \ , \quad B_2 \coloneqq (Y_i, \mathbf{Y}_{i+1}^n) \ . \end{cases} \tag{C.56}$$

Under this choice, $I_A \coloneqq f_{[1]}(A_1, A_2)$ and thus,

$$I(X_i \mathbf{X}_{i+1}^n; \mathbf{Y}^{i-1} | I_A \mathbf{X}^{i-1} Y_i \mathbf{Y}_{i+1}^n) = 0 \ , \ \forall\, i = [1:n]. \tag{C.57}$$

The later identity proves the following Markov chain:

$$(X_i, \mathbf{X}_{i+1}^n) - (I_A, \mathbf{X}^{i-1}, Y_i, \mathbf{Y}_{i+1}^n) - \mathbf{Y}^{i-1} \ , \ \forall\, i = [1:n]. \tag{C.58}$$

As $I_B \coloneqq g_{[1]}(I_A, Y^n)$, it also holds that:

$$X_i - (I_A, \mathbf{X}^{i-1}, Y_i^n) - I_B \ , \ \forall\, i = [1:n] \tag{C.59}$$

which yields the desired Markov chain.

C.4 Explanation of Remark 16

Having proven the optimality (at least in the weak sense) of the region proposed in Theorem 3 for the problem of cooperative testing against independence over one round of communication, we discuss the case of multiple communication rounds in this appendix. As mentioned in Remark 16, the achievability of (4.24) can be shown through the result for the error exponent with general hypotheses of Section 4.4, in much the same way as was the case for a single round of communication.

In [51] the special case of testing against independence is explored, over multiple rounds of communication. While our achievability result (acquired through the general error exponent), matches the one in [51], the authors of that work unfortunately missed a significant detail in attempting to prove a converse, similar to the one proved in Appendix C.3. In fact, an equivalent to the *multi-letter expression* of Lemma 23 cannot be shown to constitute a multi-letter converse for the case of multiple rounds of communications, making the passage from a multi-letter to a single-letter expression a moot point. We show this in this appendix.

Consider as an example a two-round scenario. As was done as part of the proof of Lemma 23, let P and Q denote the probability measures on $\mathcal{X}^n \times \{1, \ldots, |g_1|\} \times \{1, \ldots, |g_2|\}$ induced by H_0 and H_1, respectively. It is straight-forward to show that

$$-\frac{1}{n} \log \beta_n(R, \epsilon | K = 2) \leq \frac{1}{n} \mathcal{D}(P_{\mathbf{X}^n, I_A^{(1)} I_B^{(1)} I_A^{(2)} I_B^{(2)}} || Q_{\mathbf{X}^n, I_A^{(1)} I_B^{(1)} I_A^{(2)} I_B^{(2)}}) - \delta_n \ , \tag{C.60}$$

where $I_A^{(j)}$ and $I_B^{(j)}$ are the messages sent from nodes A and B, respectively, at round j. In the case of a single round of communication it was shown that

$$\mathcal{D}(P_{\mathbf{X}^n, I_A^{(1)} I_B^{(1)}} || Q_{\mathbf{X}^n, I_A^{(1)} I_B^{(1)}}) \leq I(I_A^{(1)}; \mathbf{Y}^n) + I(I_B^{(1)}; \mathbf{X}^n | I_A^{(1)}) \ , \tag{C.61}$$

thus proving the multi-letter version of the converse. When two rounds of communication are allowed, the expression in (C.60) can be expressed through the chain rule for KL-

divergence as follows:

$$\begin{aligned}&\mathcal{D}(P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}||Q_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}) = \\ &\quad = \mathcal{D}(P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}) + \mathcal{D}(P_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}|P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}})\end{aligned} \tag{C.62}$$

Analyzing each of the two arguments of (C.62) separately, the first one can be expressed as follows:

$$\mathcal{D}(P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}) \tag{C.63a}$$
$$= \mathcal{D}(P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}||Q_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}) + \mathcal{D}(P_{I_A^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}|P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}) \tag{C.63b}$$
$$= \mathcal{D}(P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}||Q_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}}) \tag{C.63c}$$
$$\leq I(I_A^{(1)};\mathbf{Y}^n) + I(I_B^{(1)};\mathbf{X}^n|I_A^{(1)}) \ . \tag{C.63d}$$

Here, (C.63b) is due to the chain rule (C.63c) stems from the fact that under both hypotheses $I_A^{(2)} = f_{[2]}(\mathbf{X}^n, I_B^{(1)})$ with the same function $f_{[2]}$, and (C.63d) is the result demonstrated for the first round in the proof of Lemma 23. The second argument in (C.62) can be expressed as follows:

$$\mathcal{D}(P_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}|P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}) \tag{C.64a}$$
$$= \sum P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}} \log \frac{P_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}}{Q_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}} \tag{C.64b}$$
$$= \sum P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}} \log \left[\frac{P_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}}{Q_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}} \times \frac{P_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}}{P_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}}\right] \tag{C.64c}$$
$$= I(I_B^{(2)};\mathbf{X}^n|I_A^{(1)}I_B^{(1)}I_A^{(2)}) + \sum P_{\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}} \log \frac{P_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}}{Q_{I_B^{(2)}|\mathbf{X}^n,I_A^{(1)}I_B^{(1)}I_A^{(2)}}} \tag{C.64d}$$
$$= I(I_B^{(2)};\mathbf{X}^n|I_A^{(1)}I_B^{(1)}I_A^{(2)}) + \mathcal{D}(P_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}|P_{I_A^{(1)}I_B^{(1)}I_A^{(2)}}) \tag{C.64e}$$

Here, all the sums are over the super-alphabet of the $\mathbf{x}$-sequence, $\mathcal{X}^n$, as well as the messages in each of the codebooks for $I_A^{(1)},\ldots,I_B^{(2)}$. The arguments of the probabilities within the sums were excluded for convenience of notation. (C.64e) is due to the fact that we are considering *testing against independence*. Thus, under probability distribution Q, the message $I_B^{(2)}$ is independent of $\mathbf{x}$, given all previous messages. Taking a closer look at the KL-divergence in (C.64e), it can be shown to comply with

$$\mathcal{D}(P_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{I_B^{(2)}|I_A^{(1)}I_B^{(1)}I_A^{(2)}}|P_{I_A^{(1)}I_B^{(1)}I_A^{(2)}}) \tag{C.65a}$$
$$\leq \mathcal{D}(P_{I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}||Q_{I_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}) \tag{C.65b}$$
$$\leq \mathcal{D}(P_{\mathbf{Y}^nI_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}||Q_{\mathbf{Y}^nI_A^{(1)}I_B^{(1)}I_A^{(2)}I_B^{(2)}}) \tag{C.65c}$$
$$= \mathcal{D}(P_{\mathbf{Y}^nI_A^{(1)}I_B^{(1)}I_A^{(2)}}||Q_{\mathbf{Y}^nI_A^{(1)}I_B^{(1)}I_A^{(2)}}) \tag{C.65d}$$

$$= \mathcal{D}(P_{\mathbf{Y}^n I_A^{(2)}|I_A^{(1)}I_B^{(1)}}||Q_{\mathbf{Y}^n I_A^{(2)}|I_A^{(1)}I_B^{(1)}}|P_{I_A^{(1)}I_B^{(1)}}) + \mathcal{D}(P_{I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(1)}I_B^{(1)}}) \tag{C.65e}$$

$$= \mathcal{D}(P_{\mathbf{Y}^n I_A^{(2)}|I_A^{(1)}I_B^{(1)}}||P_{\mathbf{Y}^n|I_A^{(1)}I_B^{(1)}}P_{I_A^{(2)}|I_A^{(1)}I_B^{(1)}}|P_{I_A^{(1)}I_B^{(1)}}) + \mathcal{D}(P_{I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(1)}I_B^{(1)}}) \tag{C.65f}$$

$$= I(I_A^{(2)};\mathbf{Y}^n|I_A^{(1)}I_B^{(1)}) + \mathcal{D}(P_{I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(1)}I_B^{(1)}}) \ . \tag{C.65g}$$

Here, (C.65b), (C.65c) and (C.65e) are due to the chain rule, (C.65d) to the fact that $I_B^{(2)}$ is a function of $\mathbf{Y}^n$ and all previous messages (unchanged under both hypotheses), and (C.65f) to the fact we are testing against independence.

Reassembling all of the expressions above, we conclude that a multi-letter upper bound to the error exponent of Type II is:

$$\begin{aligned} -\frac{1}{n}\log\beta_n(R,\epsilon|K=2) \leq & I(I_A^{(1)};\mathbf{Y}^n) + I(I_B^{(1)};\mathbf{X}^n|I_A^{(1)}) \\ & + I(I_A^{(2)};\mathbf{Y}^n|I_A^{(1)}I_B^{(1)}) + I(I_B^{(2)};\mathbf{X}^n|I_A^{(1)}I_B^{(1)}I_A^{(2)}) \\ & + \mathcal{D}(P_{I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(1)}I_B^{(1)}}) \ . \end{aligned} \tag{C.66}$$

It can be seen that this bound does include the extension to Lemma 23 we would have liked to have for cooperative communication over multiple rounds, but it adds to it another expression. This expression could be thought of as a quantification of the difference in the correlation of the two messages, between the two hypotheses. While this *does not prove* that the desired expression does not constitute a converse to the error exponent of Type II, it does show that the multi-letter step of the proof is non-trivial, and cannot be ignored. Moreover, it seems at least intuitively, that the added expression $\mathcal{D}(P_{I_A^{(1)}I_B^{(1)}}||Q_{I_A^{(1)}I_B^{(1)}})$ should be significant, as the opposite suggests that the message of node B depends only on the received message from node A, and not at all on its observed sequence $\mathbf{Y}^n$.

C.5 Proof of Theorem 4

From the expression of the error exponent in (4.26), it is clear that it is enough to show the result for $K = 1$, since it is feasible with one round and the extension of the unfeasibility proof is straightforward. We start by proving the feasibility of the error exponent in (4.26), and then we prove the unfeasibility result using methods similar to the ones in [40] for the case of a unidirectional exchanges.

C.5.1 Proof of Achievability

As the error exponent in (4.26) is feasible with single-side exchange, we use Proposition 3 setting $V = \phi$. Thus, a feasible error exponent for zero-rate, as defined in Theorem 4:

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(R=0,\epsilon\,|K) \geq \max_{\mathscr{S}(R=0)} \min_{\mathscr{L}(U,X,Y)} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}}||P_{\tilde{U}\tilde{X}\bar{Y}}) \ , \tag{C.67}$$

where $\mathscr{S}$ and $\mathscr{L}$ are the sets defined in Proposition 3. Using the chain rule for KL divergence, this exponent can be bounded as follows:

$$\begin{aligned}
&\max_{\mathscr{S}(R=0)} \min_{\mathscr{L}(U,X,Y)} \mathcal{D}(P_{\tilde{U}\tilde{X}\tilde{Y}}||P_{\bar{U}\bar{X}\bar{Y}}) \\
&\quad= \max_{\mathscr{S}(R=0)} \min_{\mathscr{L}(U,X,Y)} \Big[\mathcal{D}(P_{\tilde{X}\tilde{Y}}||P_{\bar{X}\bar{Y}}) + \mathcal{D}(P_{\tilde{U}|\tilde{X}\tilde{Y}}||P_{\bar{U}|\bar{X}\bar{Y}}|P_{\tilde{X}\tilde{Y}})\Big] \\
&\quad= \max_{\mathscr{S}(R=0)} \min_{\mathscr{L}_0(X,Y)} \Big[\mathcal{D}(P_{\tilde{X}\tilde{Y}}||P_{\bar{X}\bar{Y}}) + \min_{P_{\tilde{U}|\tilde{X}\tilde{Y}}} \mathcal{D}(P_{\tilde{U}|\tilde{X}\tilde{Y}}||P_{\bar{U}|\bar{X}\bar{Y}}|P_{\tilde{X}\tilde{Y}})\Big] \\
&\quad\geq \min_{\mathscr{L}_0(X,Y)} \mathcal{D}(P_{\tilde{X}\tilde{Y}}||P_{\bar{X}\bar{Y}}) \ .
\end{aligned} \tag{C.68}$$

Here, the minimum over $P_{\tilde{U}|\tilde{X}\tilde{Y}}$ is such that $\tilde{U}\tilde{X}\tilde{Y} \in \mathscr{L}(U,X,Y)$, $\mathcal{L}_0(X,Y)$ is as defined in Theorem 4 and the final inequality is due to the non-negativity of the KL divergence.

C.5.2 Proof of Strong Converse

We now prove the optimality of Theorem 4, by showing that the error exponent of $\beta_n(R = 0, \epsilon)$ does not depend on $\epsilon \in (0,1)$, and that (4.26) cannot be beaten. We follow a similar approach to [40], which addressed this proof for the case of unidirectional exchanges.

Let $f_{[1]} : \mathcal{X}^n \to \{1, \ldots, |f_{[1]}|\}$ and $g_{[1]} : \mathcal{Y}^n \times \{1, \ldots, |f_{[1]}|\} \to \{1, \ldots, |g_{[1]}|\}$ be the encoding functions at node A and B, respectively, and let $\phi\big(X^n, g_{[1]}(Y^n,\ f_{[1]}(X^n))\big) \in \{0,1\}$ be the decoding function at node A. Define sets:

$$\begin{aligned}
&\mathcal{C}_{ij} \coloneqq \big\{\mathbf{x} \in \mathcal{X}^n : f_{[1]}(\mathbf{x}) = i \ \text{ and } \phi(\mathbf{x}, j) = 0\big\} \ , \ \mathcal{C}_i \coloneqq \bigcup_{i=1}^{|f_{[1]}|} \mathcal{C}_{ij} \ , \\
&\mathcal{F}_{ij} \coloneqq \big\{\mathbf{y} \in \mathcal{Y}^n : g_{[1]}(\mathbf{y}, i) = j\big\} \ , \ (i,j) \in \{1, \ldots, |f_{[1]}|\} \times \{1, \ldots, |g_{[1]}|\} \ .
\end{aligned} \tag{C.69}$$

Note that $\mathcal{C}_{ij}$ (respectively, $\mathcal{F}_{ij}$) cannot be said to be pairwise disjoint in $\mathcal{X}^n$ (respectively, $\mathcal{Y}^n$) while the sets $\mathcal{C}_i$ are pairwise disjoint. Similarly, for each index i_0, the sets $\mathcal{F}_{i_0 j}$ are disjoint. The acceptance set of H_0 can be expressed by

$$\mathcal{A}_n \coloneqq \bigcup_{i=1}^{|f_{[1]}|} \bigcup_{j=1}^{|g_{[1]}|} \mathcal{C}_{ij} \times \mathcal{F}_{ij} \ . \tag{C.70}$$

That is, if $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}_n$, $\phi\big(\mathbf{x}, g_{[1]}(\mathbf{y}, f_{[1]}(\mathbf{x}))\big) = 0$ and otherwise, the result is H_1. By the definition, $P_{XY}^n(\mathcal{A}_n^c) \leq \epsilon$, or equivalently

$$P_{XY}^n(\mathcal{A}_n) = P_{XY}^n\left(\bigcup_{i=1}^{|f_{[1]}|} \bigcup_{j=1}^{|g_{[1]}|} \mathcal{C}_{ij} \times \mathcal{F}_{ij}\right) > 1 - \epsilon \ . \tag{C.71}$$

Since the sets $\mathcal{B}_i \coloneqq \bigcup_{j=1}^{|g_{[1]}|} \mathcal{C}_{ij} \times \mathcal{F}_{ij}$ are disjoint, by relying on (C.71) and on the size $|f_{[1]}|$, there exists an index i_0 such that

$$P_{XY}^n \left(\bigcup_{j=1}^{|g_{[1]}|} \mathcal{C}_{i_0 j} \times \mathcal{F}_{i_0 j} \right) \geq \frac{1-\epsilon}{|f_{[1]}|} \ . \tag{C.72}$$

As the sets $F_{i_0 j}$ are disjoint, there exists an index j_0 such that

$$P_{XY}^n (\mathcal{C}_{i_0 j_0} \times \mathcal{F}_{i_0 j_0}) \geq \frac{1-\epsilon}{|f_{[1]}||g_{[1]}|} \ . \tag{C.73}$$

Letting $\mathcal{C} \equiv \mathcal{C}_{i_0 j_0}$ and $\mathcal{F} \equiv \mathcal{F}_{i_0 j_0}$, we rewrite this as:

$$P_{XY}^n (\mathcal{C} \times \mathcal{F}) \geq \frac{1-\epsilon}{|f_{[1]}||g_{[1]}|} \equiv \exp(-n\delta_n) \ , \tag{C.74}$$

with $\delta_n \equiv \frac{1}{n} \log \big(|f_{[1]}||g_{[1]}| \big) - \frac{1}{n} \log(1-\epsilon)$. As the log-function is monotonic and both $|f_{[1]}|$ and $|g_{[1]}|$ are non-negative, expression (4.25) implies that $\log |f_{[1]}| = o(n)$ and $\log |g_{[1]}| = o(n)$ and thus $\delta_n = o(1)$.

Having shown that there exist sets $\mathcal{C}$ and $\mathcal{F}$, such that $\mathcal{C} \times \mathcal{F} \in \mathcal{A}_n$, and the probability $P_{XY}(\mathcal{C} \times \mathcal{F})$ does not approach 0 exponentially with n, the rest of the proof follows along the lines in [40]. We finish it here, for the sake of completeness. We now evoke the "Blowing-Up" Lemma:

Lemma 25 (Blowing-up Lemma)**.** *Let $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$ be independent random variables in $(\mathcal{Y}^n, \mathcal{B}_{\mathcal{Y}^n})$ distributed according to $W^n(\mathbf{Y}^n | \mathbf{X}^n = \mathbf{x})$ for some fixed vector $\mathbf{x} \in \mathcal{X}^n$ and a stochastic mapping $W : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$ and let $\delta_n \to 0$ be a given sequence. There exist sequences $k_n \equiv o(n)$ and $\gamma_n \equiv o(1)$, such that for every subset $\mathcal{A}_n \subset \mathcal{Y}^n$:*

$$W^n(\mathcal{A}_n | \mathbf{X}^n = \mathbf{x}) \geq \exp(-n\delta_n) \text{ implies } W^n\big(\Gamma^{k_n} \mathcal{A}_n | \mathbf{X}^n = \mathbf{x}\big) \geq 1 - \gamma_n \tag{C.75}$$

where $\Gamma^{k_n} \mathcal{A}_n$ denotes the Γ^{k_n}-neighborhood of the set $\mathcal{A}_n$ defined by

$$\Gamma^{k_n} \mathcal{A}_n \coloneqq \left\{ \hat{\mathbf{y}} \in \mathcal{Y}^n : \min_{\mathbf{y} \in \mathcal{A}_n} \rho_n(\hat{\mathbf{y}}, \mathbf{y}) \leq k_n \right\} \ , \tag{C.76}$$

where $\rho_n(\hat{\mathbf{y}}, \mathbf{y}) \coloneqq \sum_{i=1}^n \mathbb{1}\{\hat{y}_i \neq y_i\}$ and $\mathbb{1}\{\hat{y} \neq y\} = 1$ if $\hat{y} \neq y$ or $= 0$ otherwise.

Proof. Refer to references [78, 107] □

As $P_{XY}^n(\mathcal{C} \times \mathcal{F}) \geq \exp(-n\delta_n)$, clearly $P_X^n(\mathcal{C}) \geq \exp(-n\delta_n)$ and $P_Y^n(\mathcal{F}) \geq \exp(-n\delta_n)$. Using the non-conditional version of Lemma 25, there exist sequences $k_n = o(n)$ and $\gamma_n = o(1)$ s.t.:

$$P_X^n\big(\Gamma^{k_n} \mathcal{C}\big) \geq 1 - \gamma_n \ , \quad P_Y^n\big(\Gamma^{k_n} \mathcal{F}\big) \geq 1 - \gamma_n \ , \tag{C.77}$$

where k_n, γ_n only depend on $|\mathcal{X}|, |\mathcal{Y}|$ and δ_n, but not on P_{XY}. Equation (C.77) holds true if we change P_X to $P_{\tilde{X}}$ and P_Y to $P_{\tilde{Y}}$, for some $\tilde{X}\tilde{Y} \in \mathscr{L}_0$. As we wish to analyze the error probability for fixed n, during most of this proof we take the liberty to dismiss the subscript n from k_n, for the sake of readability.

Using the fact $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$ and (C.77), we obtain:

$$P^n_{\tilde{X}\tilde{Y}}\big(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}\big) \geq P^n_{\tilde{X}}\big(\Gamma^k\mathcal{C}\big) + P^n_{\tilde{Y}}\big(\Gamma^k\mathcal{F}\big) - 1 \geq 1 - 2\gamma_n \ . \tag{C.78}$$

Consider the set of η-typical sequences defined by $P_{\tilde{X}\tilde{Y}}$. By Lemma 9,

$$P^n_{\tilde{X}\tilde{Y}}(\mathcal{T}_{[\tilde{X}\tilde{Y}]\eta}) \geq 1 - \mathcal{O}\left(\frac{1}{n\eta^2}\right) = 1 - \mathcal{O}\left(n^{-\frac{1}{3}}\right) \ , \tag{C.79}$$

where the last equality is a result of the choice $\eta \equiv \eta_n \coloneqq n^{-\frac{1}{3}}$. Combining (C.78) and (C.79), it is clear that for sufficiently large n,

$$P^n_{\tilde{X}\tilde{Y}}\big((\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}) \cap \mathcal{T}_{[\tilde{X}\tilde{Y}]_\eta}\big) \geq \frac{1}{2} \ . \tag{C.80}$$

By the definition of the η-typical set (see Definition 7 and in particular Remark 2), we have:

$$\mathcal{T}_{[\tilde{X}\tilde{Y}]_\eta} = \bigcup_{\substack{P_{\hat{X}\hat{Y}} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y}) \\ |P_{\hat{X}\hat{Y}} - P_{\tilde{X}\tilde{Y}}| \leq \eta \ , \ P_{\hat{X}\hat{Y}} \ll P_{\tilde{X}\tilde{Y}}}} \mathcal{T}_{[\hat{X}\hat{Y}]} \ , \tag{C.81}$$

where $|P_{\hat{X}\hat{Y}} - P_{\tilde{X}\tilde{Y}}| \leq \eta$ refers to the maximum over all the arguments in $\mathcal{X} \times \mathcal{Y}$. As all elements of $\mathcal{T}_{[\hat{X}\hat{Y}]}$ are equiprobable under an i.i.d measure, (C.80) can be rewritten as

$$\sum_{\substack{P_{\hat{X}\hat{Y}} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y}) \\ |P_{\hat{X}\hat{Y}} - P_{\tilde{X}\tilde{Y}}| \leq \eta \ , \ P_{\hat{X}\hat{Y}} \ll P_{\tilde{X}\tilde{Y}}}} P^n_{\tilde{X}\tilde{Y}}\big(\mathcal{T}_{[\hat{X}\hat{Y}]}\big) \frac{|(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}) \cap \mathcal{T}_{[\hat{X}\hat{Y}]_\eta}|}{|\mathcal{T}_{[\hat{X}\hat{Y}]_\eta}|} \geq \frac{1}{2} \ . \tag{C.82}$$

As $P^n_{\tilde{X}\tilde{Y}}(\mathcal{T}_{[\hat{X}\hat{Y}]}) \leq 1$, by using the bound over the size of the set $\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ in Lemma 4, there must be *at least one type* $\mathcal{T}_{[\hat{X}\hat{Y}]}$, for which

$$\frac{|(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}) \cap \mathcal{T}_{[\hat{X}\hat{Y}]_\eta}|}{|\mathcal{T}_{[\hat{X}\hat{Y}]_\eta}|} \geq \frac{1}{2}(n+1)^{-|\mathcal{X}||\mathcal{Y}|} = \frac{1}{2}\exp(-n\epsilon_n) \ , \tag{C.83}$$

with $\epsilon_n = \mathcal{O}(n^{-1}\log(n+1)) \to 0$ as $n \to \infty$. The equiprobability property is also true for the probability measure implied by H_1, that is $P_{\bar{X}\bar{Y}}$. Thus,

$$\begin{aligned} P^n_{\bar{X}\bar{Y}}\big(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}\big) &\geq P^n_{\bar{X}\bar{Y}}\big((\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}) \cap \mathcal{T}_{\hat{X}\hat{Y}}\big) \\ &= P^n_{\bar{X}\bar{Y}}(\mathcal{T}_{\hat{X}\hat{Y}}) \frac{|(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}) \cap \mathcal{T}_{\hat{X}\hat{Y}}|}{|\mathcal{T}_{\hat{X}\hat{Y}}|} \\ &\geq \frac{1}{2}\exp(-n\epsilon_n) P^n_{\bar{X}\bar{Y}}(\mathcal{T}_{\hat{X}\hat{Y}}) \ , \end{aligned} \tag{C.84}$$

where the final inequality stems from (C.83).

Consider now an arbitrary element $(\mathbf{u}, \mathbf{v}) \in \Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}$. By definition, there exist an element $(\mathbf{x}, \mathbf{y}) \in \mathcal{C} \times \mathcal{F}$, such that $(u_i, v_i) \neq (x_i, y_i)$ at most in $2k$ locations. Thus,

$$P^n_{\bar{X}\bar{Y}}(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{n} P_{\bar{X}\bar{Y}}(u_i, v_i) \leq \rho^{-2k} \prod_{i=1}^{n} P_{\bar{X}\bar{Y}}(x_i, y_i) = \rho^{-2k} P^n_{\bar{X}\bar{Y}}(\mathbf{x}, \mathbf{y}) \ , \tag{C.85}$$

with $\rho = \min\limits_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{\bar{X}\bar{Y}}(x, y)$, and we assume that $\rho > 0$ (which complies with the preliminaries of Theorem 4). As $(\mathbf{u}, \mathbf{v})$ range over $\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}$, each element $(\mathbf{x}, \mathbf{y}) \in \mathcal{C} \times \mathcal{F}$ will be chosen as the closest neighbor at most $|\Gamma^k(\mathbf{x})| \times |\Gamma^k(\mathbf{y})|$ times. Thus,

$$P^n_{\bar{X}\bar{Y}}\big(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}\big) \leq \rho^{-2k} |\Gamma^k(\mathbf{x})| \times |\Gamma^k(\mathbf{y})| P^n_{\bar{X}\bar{Y}}(\mathcal{C} \times \mathcal{F}) \ . \tag{C.86}$$

From [82, Lemma 5.1] we have:

$$|\Gamma^k_n(\mathbf{x})| \leq \exp\left[n\left(h_2\left(\frac{k_n}{n}\right) + \frac{k_n}{n}\log|\mathcal{X}|\right)\right] \equiv \exp(n\zeta'_n) \ , \tag{C.87}$$

with $h_2(\cdot)$ being the *binary entropy* function and $\zeta'_n \to 0$ as $n \to \infty$. This implies that

$$P^n_{\bar{X}\bar{Y}}\big(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}\big) \leq \exp(n\zeta_n) P^n_{\bar{X}\bar{Y}}(\mathcal{C} \times \mathcal{F}) \ , \tag{C.88}$$

with $\zeta_n \coloneqq 2h_2\left(\frac{k_n}{n}\right) + \frac{k_n}{n}\log(|\mathcal{X}|\mathcal{Y}|) - \frac{2k_n}{n}\log\rho \to 0$ as $n \to \infty$. Combining this with (C.84), we finally get

$$\begin{aligned} P^n_{\bar{X}\bar{Y}}(\mathcal{C} \times \mathcal{F}) &\geq \exp(-n\zeta_n) P^n_{\bar{X}\bar{Y}}\big(\Gamma^k\mathcal{C} \times \Gamma^k\mathcal{F}\big) \\ &\geq \frac{1}{2}\exp\left[-n(\zeta_n + \epsilon_n)\right] P^n_{\bar{X}\bar{Y}}(\mathcal{T}_{\hat{X}\hat{Y}}) \\ &\geq \frac{(n+1)^{|\mathcal{X}|\mathcal{Y}|}}{2}\exp\left[-n\big(\mathcal{D}(P_{\hat{X}\hat{Y}}\|P_{\bar{X}\bar{Y}}) + \zeta_n + \epsilon_n\big)\right] \\ &\geq \exp\left[-n\big(\mathcal{D}(P_{\hat{X}\hat{Y}}\|P_{\bar{X}\bar{Y}}) + \mu_n\big)\right] \ , \end{aligned} \tag{C.89}$$

and $\mu_n \equiv \mu_n(\rho, \epsilon, M_n, N_n, |\mathcal{X}|, |\mathcal{Y}|) \to 0$ as $n \to \infty$.

The previous conclusion is true for *some type* $P_{\hat{X}\hat{Y}}$ over the range of all types that are η-typical for the measure $P_{\tilde{X}\tilde{Y}}$. As the divergence functional $\mathcal{D}(\cdot\|\cdot)$ is convex and bounded, it is also uniformly continuous. It follows that we can find a sequence $\mu'_n \equiv \mu'_n(\rho, |\mathcal{X}|, |\mathcal{Y}|)$ such that $|P_{\hat{X}\hat{Y}} - P_{\tilde{X}\tilde{Y}}| \leq \eta = o(n^{-\frac{1}{3}})$ implies that $|\mathcal{D}(P_{\hat{X}\hat{Y}}\|P_{\bar{X}\bar{Y}}) - \mathcal{D}(P_{\tilde{X}\tilde{Y}}\|P_{\bar{X}\bar{Y}})| \leq \mu'_n$. Hence

$$P^n_{\bar{X}\bar{Y}}(\mathcal{C} \times \mathcal{F}) \geq \exp\left[-n\big(\mathcal{D}(P_{\tilde{X}\tilde{Y}}\|P_{\bar{X}\bar{Y}}) + \mu_n + \mu'_n\big)\right] \ , \tag{C.90}$$

and consequently

$$\begin{aligned} -\liminf_{n\to\infty}\frac{1}{n}\log P^n_{\bar{X}\bar{Y}}(\mathcal{A}_n) &= -\lim_{n\to\infty}\frac{1}{n}\log\beta_n(R = 0, \epsilon\,|K = 1) \\ &\leq \mathcal{D}(P_{\tilde{X}\tilde{Y}}\|P_{\bar{X}\bar{Y}}) \ , \end{aligned} \tag{C.91}$$

and the RVs $\tilde{X}\tilde{Y}$ are chosen from the set $\mathscr{L}_0$, which concludes the proof.

Appendix D

Résumé

Ces dernières années, l'intérêt scientifique porté aux différents aspects des systèmes autonomes est en plein croissance. Des voitures autonomes jusqu'à l'Internet des objets, il est clair que la capacité des systèmes à prendre des décisions de manière autonome devient cruciale. De plus, ces systèmes opèreront avec des ressources limitées. Dans cette thèse, ces systèmes sont étudiés sous l'aspect de la théorie de l'information, dans l'espoir qu'une compréhension fondamentale de leurs limites et de leurs utilisations pourra aider leur conception par les futurs ingénieurs.

Dans ce travail, divers problèmes de décision binaire distribuée et collaborative sont considérés. Deux participants doivent "déclarer" la mesure de probabilité de deux variables aléatoires, distribuées conjointement par un processus sans mémoire et désignées par $\mathbf{X}^n = (X_1, \ldots, X_n)$ et $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$. Cette décision est prise entre deux mesures de probabilité possibles sur un alphabet fini, désignées P_{XY} et $P_{\bar{X}\bar{Y}}$. Les prélèvements marginaux des variables aléatoires, $\mathbf{X}^n$ et $\mathbf{Y}^n$ sont supposés disponibles aux différents sites.

Il est permis aux participants d'échanger des quantités limitées d'information sur un canal parfait avec une contrainte de débit maximal. Durant cette thèse, la nature de cette communication varie. D'abord, seule une communication unidirectionnelle est permise. Le récepteur de cette communication doit, en utilisant également sa propre information, identifier d'abord la légitimité de son expéditeur, en déclarant la distribution conjointe des processus. Il peut ensuite devoir, selon cette authentification, générer une reconstitution adéquate des observations de l'émetteur, qui satisfait une contrainte de distorsion moyenne. La performance de cette configuration est étudiée via la région réalisable de débit-erreur-distorsion, qui décrit le compromis entre: le débit de communication, la probabilité d'erreur en détection et la distorsion attendue de la reconstitution de la source.

Nous séparons le cas général d'un cas spécial où il est supposé que le test est fait *contre l'indépendance*. Dans ce cas, nous supposons que l'hypothèse H_1 implique l'indépendance statistique entre les variables X et Y, qui conservent leur distributions marginales respectives: $P_{\bar{X}\bar{Y}} = P_X P_Y$. Il s'avère que dans ce cas la région débit-erreur-distorsion réalisable est aussi *optimale*. Dans le cas général, tandis qu'une théorème d'optimalité

reste illusoire, nous présentons deux nouvelles approches au problème. Ici, une stratégie de groupage aléatoire de mots de code ("binning") s'avère très bénéfique pour la performance du système, même si elle paraît risquée au premier regard. Nous démontons ce fait à travers un exemple, et continuons à prouver que la relaxation de la demande de reconstitution au récepteur est bénéfique en général, ce qui diffère du cas de test contre l'indépendance.

Un scénario différent est étudié ensuite, dans lequel les participants peuvent utiliser un lien bidirectionnel pour arriver à leur conclusion. Un tel scénario permet la considération de multiples tours d'interactions, un choix qui diverge des études précédentes. Un tour unique de communication est d'abord considéré, avant que le résultat soit généralisé pour inclure un nombre indéfini (mais non pas infini) de tours. Un résultat de faisabilité est démontré pour le cas général de chaque hypothèse. Le cas spécial de test contre l'indépendance, où il est supposé que l'hypothèse alternative implique des sources indépendantes, est revisité comme une instance du cas général. Il est démontré que le résultat général conduit au résultat connu pour ce cas spécial. Un résultat de non-faisabilité est démontré pour le cas spécial, prouvant par conséquent l'optimalité de ce résultat, au moins dans le cas d'un seul tour de communication. In est expliqué pourquoi ce résultat de non-faisabilité (et donc d'optimalité) n'est pas généralisable pour un scénario de multiples tours de communication. Un autre cas spécial est considéré, où la communication est faite avec débit nul, pour lequel il est démontré que l'interaction n'améliore pas les performances.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Détection Binaire Distribuée sous Contraintes de Communication

Mots clefs : Test d'hypothèse, exposants d'erreur, codage de source avec pertes, débit-distorsion

Résumé : Ces dernières années, l'intérêt scientifique porté aux différents aspects des systèmes autonomes est en plein croissance. Des voitures autonomes jusqu'à l'Internet des objets, il est clair que la capacité des systèmes à prendre des décisions de manière autonome devient cruciale. De plus, ces systèmes opèreront avec des ressources limitées. Dans cette thèse, ces systèmes sont étudiés sous l'aspect de la théorie de l'information, dans l'espoir qu'une compréhension fondamentale de leurs limites et de leurs utilisations pourra aider leur conception par les futurs ingénieurs.
Dans ce travail, divers problèmes de décision binaire distribuée et collaborative sont considérés. Deux participants doivent "déclarer" la mesure de probabilité de deux variables aléatoires, distribuées conjointement par un processus sans mémoire et désignées par $\mathbf{X}^n = (X_1, \ldots, X_n)$ et $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$. Cette décision est prise entre deux mesures de probabilité possibles sur un alphabet fini, désignées P_{XY} et $P_{\bar{X}\bar{Y}}$. Les prélèvements marginaux des variables aléatoires, $\mathbf{X}^n$ et $\mathbf{Y}^n$ sont supposés disponibles aux différents sites.
Il est permis aux participants d'échanger des quantités limitées d'information sur un canal parfait avec une contrainte de débit maximal. Durant cette thèse, la nature de cette communication varie. La communication unidirectionnelle est considérée d'abord, suivie par la considération de communication bidirectionnelle, qui permet des échanges interactifs entre les participants.

Title : Distributed Binary Detection with Communication Constraints

Keywords : Hypothesis testing, Error exponents, Lossy source coding, Re-distortion

Abstract : In recent years, interest has been growing in research of different autonomous systems. From the self-driving car to the Internet of Things (IoT), it is clear that the ability of automated systems to make autonomous decisions in a timely manner is crucial in the 21st century. These systems will often operate under strict constraints over their resources. In this thesis, an information-theoretic approach is taken to this problem, in hope that a fundamental understanding of the limitations and perspectives of such systems can help future engineers in designing them.
Throughout this thesis, collaborative distributed binary decision problems are considered. Two statisticians are required to declare the correct probability measure of two jointly distributed memoryless processes, denoted by $\mathbf{X}^n = (X_1, \ldots, X_n)$ and $\mathbf{Y}^n = (Y_1, \ldots, Y_n)$, out of two possible probability measures on finite alphabets, namely P_{XY} and $P_{\bar{X}\bar{Y}}$. The marginal samples given by $\mathbf{X}^n$ and $\mathbf{Y}^n$ are assumed to be available at different locations. The statisticians are allowed to exchange limited amounts of data over a perfect channel with a maximum-rate constraint. Throughout the thesis, the nature of communication varies. First, only unidirectional communication is allowed. Using its own observations, the receiver of this communication is required to first identify the legitimacy of its sender by declaring the joint distribution of the process, and then depending on such authentication it generates an adequate reconstruction of the observations, satisfying an average per-letter distortion. Bidirectional communication is subsequently considered, in a scenario that allows interactive communication between the participants.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France