entropy

MDPI

Article

Joint Communication and Channel Discrimination

Han Wu and Hamdi Joudeh *

Information and Communication Theory Lab, Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands; h.wu1@tue.nl
* Correspondence: h.joudeh@tue.nl

Abstract: We consider a basic joint communication and sensing setup comprising a transmitter, a receiver and a sensor. The transmitter sends a codeword to the receiver through a discrete memoryless channel, and the receiver is interested in decoding the transmitted codeword. At the same time, the sensor picks up a noisy version of the transmitted codeword through one of two possible discrete memoryless channels. The sensor knows the codeword and wishes to discriminate between the two possible channels, i.e., to identify the channel that has generated the output given the input. We study the trade-off between communication and sensing in the asymptotic regime, captured in terms of the channel coding rate against the two types of discrimination error exponents. We characterize the optimal trade-off between the rate and the exponents for general discrete memoryless channels with an input cost constraint.

Keywords: capacity; discrimination exponents; joint communication and sensing

Citation: Wu, H.; Joudeh, H. Joint Communication and Channel Discrimination. *Entropy* **2024**, *26*, 1089. https://doi.org/10.3390/e26121089

Academic Editors: Mehrasa Ahmadipour, Shlomo Shamai (Shitz) and Michele Wigger

Received: 5 November 2024
Revised: 29 November 2024
Accepted: 6 December 2024
Published: 13 December 2024

1. Introduction

We consider a setting comprising a transmitter, a receiver and a sensor. The transmitter has a random message W which it encodes into a sequence $X^n \triangleq X_1, X_2, \ldots, X_n$ of length n, drawn from an alphabet $\mathcal{X}^n$. This sequence serves as an input to a pair of channels $P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n$ and $P^{\theta}_{Z^n|X^n} : \mathcal{X}^n \to \mathcal{Z}^n$, where $\mathcal{Y}^n$ and $\mathcal{Z}^n$ are the corresponding output alphabets. The receiver observes $Y^n \triangleq Y_1, Y_2, \ldots, Y_n$ through $P_{Y^n|X^n}$ and wishes to retrieve the message W from Y^n. The sensor, on the other hand, observes $Z^n \triangleq Z_1, Z_2, \ldots, Z_n$ through $P^{\theta}_{Z^n|X^n}$, which depends on a fixed yet unknown parameter $\theta \in \Theta$. The sensor has access to W as side information (the transmitter and sensor are, e.g., co-located) and wishes to estimate the channel parameter θ from (Z^n, W). An illustration is shown in Figure 1.

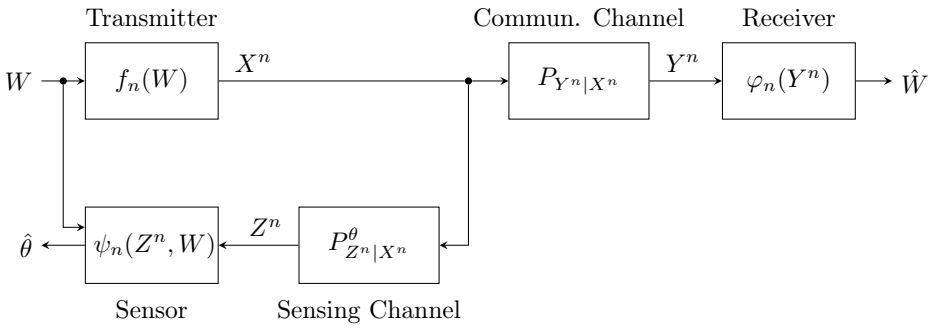

Figure 1. Illustration of the considered setting. A precise definition of all blocks is given in Section 2.

The above setting is a very basic model for joint communication and sensing (JCAS), or integrated sensing and communication (ISAC). The JCAS/ISAC paradigm has emerged with the aim of designing wireless systems in which transceivers utilize the same hardware and spectrum resources efficiently to both communicate and sense, and has received

increased research attention in recent years; see, e.g., [1,2]. Potential practical use cases include next-generation cellular networks, where base stations will be able to communicate with active devices and simultaneously detect and track passive moving targets from backscattered signals [3]; and automotive applications, where vehicles will be able to sense their surroundings to identify and track road obstacles, and communicate with other vehicles on the road and infrastructure [4,5].

In this paper, our aim is to shed some light on the fundamental performance limits and trade-offs in JCAS systems. As a step in this direction, we focus on discrete memoryless settings: the input and output alphabets are finite, and the noisy channels are stationary and memoryless. Moreover, we also limit our attention to the case where the parameter θ is drawn from $\Theta = \{0, 1\}$, which represents the most basic sensing task of target detection or classification. That is, with knowledge of W (and, hence, X^n) and upon observing Z^n, the sensor wishes to discriminate between the two channels $P^0_{Z^n|X^n}$ and $P^1_{Z^n|X^n}$.

Note that in the setting we consider, the sensing channel parameter (or state) to be estimated does not influence the communication channel. This is an abstraction of practical scenarios where the phenomenon or target to be sensed is distinct from the device or user involved in communication; see, e.g., [1,2].

1.1. Related Work

A basic information-theoretic formulation for JCAS was proposed by Kobayashi et al. [6], involving a terminal communicating with a second terminal over a state-dependent memoryless channel, and simultaneously estimating the channel state sequence from generalized feedback (in relation to the model in Figure 1, the transmit-estimate terminal in [6] includes both the transmitter and sensor). The performance trade-off between communication and state estimation is characterized in terms of a capacity-distortion function, a quantity borrowed from earlier works on state amplification [7–9]. The results and insights from [6] were extended by several authors in multiple directions, including multi-terminal settings [10–12], secrecy-constrained settings [13,14] and multi-antenna Gaussian settings [15], to mention a few. The capacity-distortion trade-off in the original setting of [6] has also been studied under a special logarithmic loss distortion measure, yielding a simple characterization in terms of mutual information quantities [16].

All the above-mentioned works follow the same modeling logic in [6]. That is, the state to be estimated varies in an i.i.d. fashion from one channel use to the other. This model fails to capture scenarios where the sensing task involves estimating a state (or parameter) that changes at a much slower timescale compared to channel uses. To study this latter case, one may consider an abstraction where the state to be estimated remains fixed throughout the whole transmission period, a model that sits at the extreme opposite of the i.i.d. state model. This approach was first taken in by Joudeh and Willems in [17], where a special case of the setting in Figure 1 with discrete binary channels was considered, as well as a case with continuous Gaussian channels. In both cases studied in [17], the sensing task considered is that of target detection, i.e., discriminating between a target response and pure noise. The extension to discriminating between an arbitrary pair of channels in discrete memoryless settings, as described in Figure 1, was considered in our preliminary works [18,19]. Concurrently, Chang et al. [20] studied an almost identical model to the one in Figure 1 and further investigated the case of discriminating between more than two channels and the role of adaptive schemes, a work that was later extended in [21]. Other related works include the extension of the target detection setting in [17] to vector Gaussian channels (i.e., multiple antennas) [22], as well as strong converse results for discrete memoryless settings in [23,24].

Another seemingly related line of work considers the problem of joint detection and decoding at the receiver [25,26]. In the most basic instance of this problem, the receiver wishes to detect the presence of a codeword, i.e., discriminate between codeword and noise, and decode it in case it is present. In contrast with the setting in Figure 1, the joint detection

and decoding problem co-locates the sensor with the receiver and not the transmitter, and is hence quite distinct from the problem we study in the current paper.

Most relevant to the present paper are the results of Chang et al. in [20,21] and our preliminary results in [18,19]. The nuanced differences between these works are further elaborated in light of our contributions in the next subsection. Before we proceed, we highlight a few more relevant works, some from the classical literature. The basic sensing task that we consider, with a binary parameter θ, is a simple binary hypothesis testing problem. This is a canonical problem in both statistics and information theory, and notable works that characterize the asymptotic performance limits include those by Chernoff [27], Hoeffding [28], Shannon-Gallager-Berlekamp [29], Csiszár-Longo [30] and Blahut [31]. The specific version of the problem that we consider here, where the sensor knows (and may control) the input X^n and wishes to distinguish between two channels $P^0_{Z^n|X^n}$ and $P^1_{Z^n|X^n}$ from an observation Z^n, is also known as channel discrimination [32], hypothesis testing with feedback [33] or controlled sensing [34]. This problem has been considered in a number of works under various assumptions, including fixed-length non-adaptive transmission in Shannon-Gallager-Berlekamp [29] and Blahut [31]; fixed-length adaptive transmission in Hayashi [32], Polyanskiy-Verdú [33] and Nitinawarat et al. [34]; and variable-length adaptive transmission in Polyanskiy-Verdú [33] and Nitinawarat et al. [34].

In this paper, we consider fixed-length transmission focusing mainly on the non-adaptive case (i.e., no feedback). More importantly, in addition to facilitating channel discrimination at the sensor, the input sequence X^n in our setting must also carry a message to the receiver, which distinguishes our problem from the ones previously considered in the channel discrimination and controlled sensing literature.

1.2. Contribution and Comparison

We consider the setting illustrated in Figure 1 with discrete memoryless channels, a binary parameter θ and an average input cost constraint; and we study the trade-off between reliable message communication and efficient channel discrimination in the asymptotic regime (i.e., $n \to \infty$). This trade-off is captured in terms of the channel coding rate against the two channel discrimination error exponents (i.e., the rate–exponent region). Note that the error exponents capture the exponential decay rates of the two types of channel discrimination errors, known as type I and type II errors in the hypothesis testing literature.

In our main result (Theorem 1 in Section 3), we establish the optimal trade-off between the channel coding rate and the two channel discrimination exponents (i.e., the optimal rate–exponent region). We also provide insights into the trade-off and demonstrate it through a couple of simple examples. The proof of our main result is obtained by adapting classical results on binary hypothesis testing, combined with a channel coding argument with constrained input sequences (see Sections 4 and 5). In the proof, we encountered an interesting technical challenge, specifically in showing the converse to the channel coding rate. The coupling with channel discrimination imposes a constraint on the types of admissible input sequences, i.e., their empirical distributions. However, unlike additive cost constraints [35,36], or similarly, constraints imposed by sensing an i.i.d. state sequence subject to an additive distortion measure [6,9,11], the constraint we encounter here is non-convex in the input sequence type. This prohibits us from directly applying the common approach of upper-bounding the multi-letter mutual information through concavity and Jensen's inequality, which strongly relies on the convexity of the set of input distributions that satisfy the cost (or, similarly, the distortion) constraint. The approach we take here reduces the problem to upper-bounding the rate of a constant-composition code, where all input sequences are of the same type, and uses a slightly more refined analysis of the multi-letter mutual information (see Section 5.2). It is worth noting that our converse proof, which relies on reducing the problem to that of constant-composition codes, first appeared in an earlier version of the current paper [19], posted on arXiv on 15 August 2022. Since

then, the exact same approach was also adopted in [21] (Section V.B), which first appeared on arXiv on 14 October 2022.

In Section 6, we consider two important special cases of our general result in Theorem 1. In the first case, we consider a minimax error criterion for channel discrimination, where the goal is to minimize the worst of the two types of error, and we characterize the optimal rate–exponent trade-off region in this case (see Theorem 2). In the second special case, we adopt a Neyman–Pearson channel discrimination error criterion, where the goal is to minimize one type of error while keeping the other type below a predefined threshold, and we derive the optimal trade-off in this case as well (see Theorem 3). This latter case is relevant in many practical applications, e.g., in obstacle detection to avoid road collisions in automotive scenarios, a missed detection is much worse than a false alarm.

As mentioned earlier, a special case of the problem addressed in this paper, with binary channels and an on–off channel parameter (i.e., target detection), was considered in [17] under a minimax channel discrimination error criterion. The results we present here generalize [17] (Theorem 1) to arbitrary discrete memoryless channels with input cost, and to the entire trade-off between the rate and the two types of channel discrimination exponents. Other very closely related results were also reported in [21] and its preliminary version [20]. In these works, the authors consider a setting where θ is not necessarily binary, but belongs to a finite set and affects both the communication and sensing channels; hence, the communication problem is of a compound nature. Moreover, they investigate both non-adaptive and adaptive schemes. Nevertheless, refs. [20,21] focus only on the minimax channel discrimination criterion (they also do not consider an input cost constraint, which is, however, a minor difference and can be easily incorporated). Our results are somewhat comparable to [20] (Theorem 3) and [21] (Theorem 1), yet are more general in some sense, as we characterize trade-off between the two exponents, and are more restricted in another sense, as we focus on the binary parameter case. In this context, it should be noted that for binary θ and under fixed-length transmission, adaptivity does not improve the channel discrimination error exponents [32,33]; hence, there is no loss in generality in our restriction to non-adaptive schemes.

A more subtle difference compared to the preliminary work of Chang et al. [20] is in the definition of the discrimination error. As we shall see in Section 2, we define the two types of discrimination error probabilities by taking the maximum (i.e., worst-case) over all codewords in the codebook. We believe this to be a natural definition from an operational perspective, since it provides performance guarantees for sensing regardless of which codeword is used for communication, as it is not known beforehand which codeword (or message) will be selected. This worst-case formulation, however, also requires more involved analysis. For instance, i.i.d. code ensembles are insufficient for proving achievability in this case, due to the fact that bad codewords from a channel discrimination perspective, while improbable, are still possible under such ensembles. In ref. [20], the authors alleviate this challenge by considering the average discrimination error over all codewords, which provides no guarantees on the sensing performance for the worst-case codeword, but renders i.i.d. code ensembles sufficient. In our preliminary work [19], we dealt with worst-case sensing errors by resorting to constant-composition codes (see also [17]), which were also later adopted by Chang et al. in [21]. In the current paper, we prove achievability using almost-constant-composition codes through a constrained version of the channel coding theorem, which has the advantage of being directly applicable to channels with continuous alphabets.

1.3. Notation

Upper-case letters, e.g., X, Y, Z, often denote random variables and the corresponding lower-case letters, e.g., x, y, z, denote their realizations. Calligraphic letters, e.g., $\mathcal{M}$, denote sets. $|\mathcal{M}|$ denotes the cardinality of set $\mathcal{M}$. The indicator function $\mathbb{1}[\mathcal{A}]$ is equal to 1 if the event $\mathcal{A}$ is true, and is 0 otherwise. Let X and Y be, respectively, an input and output to a channel $P_{Y|X}$, which is a stochastic mapping from the input alphabet $\mathcal{X}$ to the output

alphabet $\mathcal{Y}$. The mutual information $I(X;Y)$ is also denoted by $I(P_X, P_{Y|X})$. The Bernoulli distribution with parameter p is denoted by $\text{Bern}(p)$ and the binary symmetric channel with parameter q is denoted by $\text{BSC}(q)$. For $p, q \in [0,1]$, we define $p * q \triangleq (1-q)p + q(1-p)$.

Next, we present some notation and preliminaries on types from [35], which will be essential in the technical development of our results. Given a sequence $x^n \in \mathcal{X}^n$, we define

$$N(a|x^n) \triangleq \sum_{i=1}^{n} \mathbb{1}[x_i = a], a \in \mathcal{X}. \tag{1}$$

The type of x^n, denoted by $\hat{P}_{x^n}$, is a distribution on $\mathcal{X}$ defined as

$$\hat{P}_{x^n}(a) = \frac{N(a|x^n)}{n}, \ a \in \mathcal{X}. \tag{2}$$

Let $\mathcal{P}(\mathcal{X})$ be the set of all distributions (i.e., probability mass functions) on $\mathcal{X}$ and let $\mathcal{P}_n(\mathcal{X})$ be the set of all types of sequences in $\mathcal{X}^n$. Note that $\mathcal{P}_n(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$. A very important property is that the number of types in $\mathcal{P}_n(\mathcal{X})$ is at most a polynomial in n, which follows from the upper bound [35] (Lemma 2.2)

$$|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}. \tag{3}$$

2. Problem Setting

We consider the setting introduced in Section 1 and illustrated in Figure 1 with finite alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ and a binary parameter $\theta \in \{0,1\}$. The channels are stationary and memoryless, that is,

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \quad \text{and} \quad P^{\theta}_{Z^n|X^n}(z^n|x^n) = \prod_{i=1}^{n} P^{\theta}_{Z|X}(z_i|x_i) \tag{4}$$

where n is a positive integer that denotes the block length (i.e., number of channel uses). An admissible input sequence $x^n \in \mathcal{X}^n$ must satisfy an average cost constraint of

$$\frac{1}{n}\sum_{i=1}^{n} b(x_i) \leq B \tag{5}$$

where $b : \mathcal{X} \to \mathbb{R}_+$ is some non-negative cost function and $B \geq 0$ is the average cost constraint. To simplify the notation, we use

$$P^n_0(z^n|x^n) = \prod_{i=1}^{n} P_0(z_i|x_i) \ \text{ and } \ P^n_1(z^n|x^n) = \prod_{i=1}^{n} P_1(z_i|x_i)$$

to denote $P^0_{Z^n|X^n}(z^n|x^n)$ and $P^1_{Z^n|X^n}(z^n|x^n)$, respectively, where P_0 and P_1 denote $P^0_{Z|X}$ and $P^1_{Z|X}$, respectively. Throughout the paper, we assume that $P_0(z|x)P_1(z|x) \neq 0$, for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, i.e., the two channels have a shared support under every input. This mild regularity condition helps avoid unnecessary technical complications, and is satisfied from many channels of interest.

2.1. Codes and Error Probabilities

For fixed block length n, let $\mathcal{M}_n \triangleq \{1, 2, \ldots, M_n\}$ be a message set of M_n message indices. An (n, M_n)-code for the above setting consists of the following mappings:

- An encoding function $f_n : \mathcal{M}_n \to \mathcal{X}^n$ that maps each message index $w \in \mathcal{M}_n$ into a codeword $x^n(w) = f_n(w)$ from $\mathcal{X}^n$, which satisfies the cost constraint in (5). The corresponding set of all M_n codewords, given by $\mathcal{C}_n \triangleq \{x^n(1), x^n(2), \ldots, x^n(M_n)\}$, is known as a codebook.

- A message decoding function $\varphi_n : \mathcal{Y}^n \to \mathcal{M}_n$ that maps each output sequence $y^n \in \mathcal{Y}^n$ into a decoded message $\hat{w} = \varphi_n(y^n)$ from the message set $\mathcal{M}_n$.
- A channel discrimination function $\psi_n : \mathcal{Z}^n \times \mathcal{M}_n \to \{0,1\}$ that maps each output sequence and message pair $(z^n, w) \in \mathcal{Z}^n \times \mathcal{M}_n$ into a decision (i.e., a hypothesis) $\hat{\theta} = \psi_n(z^n, w)$ from $\{0,1\}$.

The message W, which is drawn uniformly at random from $\mathcal{M}_n$, is encoded into $X^n = f_n(W)$ and then sent over the channels. Upon observing Y^n, the receiver produces a decoded message $\hat{W} = \varphi_n(Y^n)$. On the other end, upon observing Z^n and with knowledge of W, the sensor produces a binary decision $\hat{\theta} = \psi_n(Z^n, W)$. Next, we examine the message decoding and channel discrimination error probabilities.

Decoding error: For a given code, the probability of decoding error given that message $W = w$ has been sent is $\mathbb{P}[\varphi_n(Y^n) \neq w \mid X^n = x^n(w)]$. The maximum probability of decoding error is defined as

$$p_{\mathrm{e},n} \triangleq \max_{w \in \mathcal{M}_n} \mathbb{P}[\varphi_n(Y^n) \neq w \mid X^n = x^n(w)] \tag{6}$$

which is a common performance measure that reflects the assumption that messages are equally important. It is sometimes more convenient to work with the average probability of decoding error, defined as

$$p_{\mathrm{e},n}^{\mathrm{av}} \triangleq \frac{1}{M_n} \sum_{w \in \mathcal{M}_n} \mathbb{P}[\varphi_n(Y^n) \neq w \mid X^n = x^n(w)]. \tag{7}$$

As is often the case with asymptotic rate results in DMCs, the results we present here remain unchanged regardless of whether we choose $p_{\mathrm{e},n}$ or $p_{\mathrm{e},n}^{\mathrm{av}}$ as the measure of decoding error probability. In some cases, where it helps to emphasize the codebook being used, we write $p_{\mathrm{e},n}(\mathcal{C}_n)$ and $p_{\mathrm{e},n}^{\mathrm{av}}(\mathcal{C}_n)$.

Discrimination error: Without loss of generality, any discrimination function can be written as

$$\psi_n(z^n, w) = \mathbb{1}[z^n \notin \mathcal{A}_n(w)] \tag{8}$$

for some decision region $\mathcal{A}_n(w) \subset \mathcal{Z}^n$ which comprises output sequences that map to hypothesis $\hat{\theta} = 0$ (without loss of generality, we prohibit the trivial degenerate cases of $\mathcal{A}_n(w) = \emptyset$ and $\mathcal{A}_n(w) = \mathcal{Z}^n$). There are two types of discrimination error events associated with the two hypotheses, defined for $W = w$ as

$$\varepsilon_{0,n}(w) \triangleq \mathbb{P}[Z^n \notin \mathcal{A}_n(w) \mid \theta = 0, X^n = x^n(w)] \tag{9}$$

$$\varepsilon_{1,n}(w) \triangleq \mathbb{P}[Z^n \in \mathcal{A}_n(w) \mid \theta = 1, X^n = x^n(w)] \tag{10}$$

known as the type I error and type II error, respectively. Since it is not known beforehand which message will be sent, it is reasonable to define the discrimination error probabilities for a code as

$$\varepsilon_{0,n} = \max_{w \in \mathcal{M}_n} \varepsilon_{0,n}(w) \tag{11}$$

$$\varepsilon_{1,n} = \max_{w \in \mathcal{M}_n} \varepsilon_{1,n}(w). \tag{12}$$

This worst-case definition guarantees that by considering the pair $(\varepsilon_{0,n}, \varepsilon_{1,n})$, a certain discrimination error performance is guaranteed regardless of the selected message.

2.2. Rate–Exponent Region

We are interested in the asymptotic performance limits measured in terms of the channel coding rate and the channel discrimination error exponents. This is formalized as follows.

Definition 1. *A rate–exponent tuple (R, E_0, E_1) is said to be achievable if for every $\epsilon > 0$, and provided that $n \geq n_\epsilon$ for some (possibly large) n_ϵ, there exists an (n, M_n)-code with*

$$p_{e,n} \leq \epsilon \tag{13}$$

$$\frac{1}{n} \log M_n \geq R - \epsilon \tag{14}$$

$$-\frac{1}{n} \log \varepsilon_{0,n} \geq E_0 - \epsilon \tag{15}$$

$$-\frac{1}{n} \log \varepsilon_{1,n} \geq E_1 - \epsilon. \tag{16}$$

The rate–exponent region $\mathcal{R}$ is the closure of the set of all achievable tuples (R, E_0, E_1).

The main result of this paper is a characterization of the rate–exponent region $\mathcal{R}$ for the general discrete memoryless channels in (4) under the input cost constraint in (5).

3. Main Result

We now present the main result of the paper, which is a complete characterization of the rate–exponent region in Definition 1. To this end, we define the channel $P_s : \mathcal{X} \to \mathcal{Z}$ for $s \in [0, 1]$ as

$$P_s(z|x) = \frac{P_0(z|x)^{1-s} P_1(z|x)^s}{\sum_{z' \in \mathcal{Z}} P_0(z'|x)^{1-s} P_1(z'|x)^s}. \tag{17}$$

For any fixed input x, $P_s(\cdot|x)$ is a parameterized output distribution on $\mathcal{Z}$ known in the literature as the "tilted" distribution, which moves from $P_0(\cdot|x)$ to $P_1(\cdot|x)$ as the parameter s increases. We also have the following standard definition of the conditional relative entropy (or conditional information divergence) between any pair of channels $P_{Z|X}$ and $Q_{Z|X}$ from $\mathcal{X}$ to $\mathcal{Z}$ given an input distribution P_X on $\mathcal{X}$:

$$\begin{aligned} D\Big(P_{Z|X} \| Q_{Z|X} | P_X\Big) &= \sum_{x \in \mathcal{X}} P_X(x) D\Big(P_{Z|X}(\cdot|x) \| Q_{Z|X}(\cdot|x)\Big) \\ &= \sum_{x \in \mathcal{X}} P_X(x) \sum_{z \in \mathcal{Z}} P_{Z|X}(z|x) \log \frac{P_{Z|X}(z|x)}{Q_{Z|X}(z|x)}. \end{aligned}$$

We assume that $P_{Z|X}(z|x) = 0$ whenever $Q_{Z|X}(z|x) = 0$, and, hence, $D(P_{Z|X} \| Q_{Z|X} | P_X) < \infty$. This is satisfied for all channels of interest to us here. We are now ready to present the main result.

Theorem 1. *$\mathcal{R}$ is given by the set of all non-negative pairs (R, E_0, E_1) such that*

$$R \leq I(P_X, P_{Y|X}) \tag{18}$$

$$E_0 \leq D(P_s \| P_0 | P_X) \tag{19}$$

$$E_1 \leq D(P_s \| P_1 | P_X) \tag{20}$$

for some input distribution $P_X \in \mathcal{P}(\mathcal{X})$ that satisfies $\mathbb{E}_{X \sim P_X}[b(X)] \leq B$, and some $s \in [0, 1]$.

The proof of the above theorem is presented in Section 5. As a first step towards the proof, we first study the trade-off between the two discrimination exponents E_0 and E_1 in Section 4. Moreover, two important special cases of the result in Theorem 1, where we specialize to minimax and Neyman–Pearson discrimination error metrics, are presented and discussed in Section 6.

Insights and Examples

We now provide some insights into the result in Theorem 1. To this end, it helps to interpret the input distribution P_X in Theorem 1 as being equal, or arbitrarily close, to

the types (i.e., empirical distribution) of codewords in the employed codebooks. Under such codebooks, classical results on binary hypothesis testing are used to establish that the discrimination errors are given by

$$\varepsilon_{0,n} \approx e^{-nD(P_s\|P_0|P_X)} \text{ and } \varepsilon_{1,n} \approx e^{-nD(P_s\|P_1|P_X)}. \tag{21}$$

for some parameter $s \in [0,1]$, which is used to tune the trade-off between the two types of error probabilities, and is directly related to the choice of channel discrimination function (or decision region). On the other hand, a channel coding argument with constrained input sequences is used to show that reliable communication is achieved provided that the number of codewords in the codebook is $M_n \approx e^{nI(P_X,P_{Y|X})}$.

As seen through the above explanation, and as one may expect, the type P_X of codeword dictates the performance of both communication and channel discrimination. In general, a trade-off between the two arises since P_X, which maximizes the rate, may not simultaneously provide the best discrimination exponents. This trade-off is further explored through the following basic examples.

Example 1. *Consider a setting with a binary input, and binary outputs given by*

$$Y = X \oplus N_Y \text{ and } Z = \theta X \oplus N_Z \tag{22}$$

where N_Y and N_Z are Bernoulli with parameters p and q, respectively, and $\theta \in \{0,1\}$. Here, $P_{Y|X}$ is a BSC(p) *and P_0 is a* BSC(q)*, while P_1 satisfies $P_1(1|\cdot) = q$ and $P_1(0|\cdot) = 1-q$. Therefore, we have*

$$P_s(0|0) = 1 - P_s(1|0) = 1 - q \tag{23}$$

$$P_s(0|1) = 1 - P_s(1|1) = \frac{(1-q)^{1-s}q^s}{(1-q)^{1-s}q^s + q^{1-s}(1-q)^s} = \hat{q} \tag{24}$$

where $\hat{q} \in [\min\{q, 1-q\}, \max\{q, 1-q\}]$, depending on s. Let $P_X \sim$ Bern(ρ)*. It follows that*

$$I(P_X, P_{Y|X}) = H(\rho * p) - H(p) \tag{25}$$

$$D(P_s\|P_0|P_X) = \rho d(\hat{q}\|1-q) \tag{26}$$

$$D(P_s\|P_1|P_X) = \rho d(\hat{q}\|q) \tag{27}$$

where $H(p)$ is the entropy of the distribution Bern(p)*, while $d(q\|1-q)$ is the information divergence between* Bern(q) *and* Bern$(1-q)$*. From the above, it follows that $\mathcal{R}$ for this setting is described by*

$$R \leq H(\rho * p) - H(p) \tag{28}$$

$$E_0 \leq \rho d(\hat{q}\|1-q) \tag{29}$$

$$E_1 \leq \rho d(\hat{q}\|q) \tag{30}$$

for some $\hat{q} \in [\min\{q, 1-q\}, \max\{q, 1-q\}]$, and $\rho \leq B$ (we assume $b(1) = 1$ and $B \leq 1$). The maximum rate is achieved when $\rho = \min\{0.5, B\}$, while the best set of exponent pairs is achieved when $\rho = B$. Hence, there is a trade-off between the rate on one hand and the pair of exponents on the other whenever $B > 0.5$. Note that $\hat{q}$ controls the exponent trade-off for any fixed ρ, and does not influence the rate.

Example 2. *Consider a binary input binary output setting as in the previous example, but here P_0 is a* BSC(p_0) *and P_1 is a* BSC(p_1)*. In this case, for any $s \in [0,1]$ we have*

$$P_s(0|0) = P_s(1|1) = \frac{(1-p_0)^{1-s}(1-p_1)^s}{(1-p_0)^{1-s}(1-p_1)^s + p_0^{1-s}p_1^s} = 1 - q \tag{31}$$

where $q \in [\min\{p_0, p_1\}, \max\{p_0, p_1\}]$. *It follows that*

$$D(P_s \| P_0 | P_X) = d(q \| p_0) \tag{32}$$
$$D(P_s \| P_1 | P_X) = d(q \| p_1) \tag{33}$$

which do not depend on the input cost constraint. Therefore, $\mathcal{R}$ *in this case is given by*

$$R \leq H(\min\{0.5, B\} * p) - H(p) \tag{34}$$
$$E_0 \leq d(q \| p_0) \tag{35}$$
$$E_1 \leq d(q \| p_1) \tag{36}$$

for some $q \in [\min\{p_0, p_1\}, \max\{p_0, p_1\}]$. *Hence, in this case, the rate R and the exponent pair* (E_0, E_1) *are not coupled, and there is no trade-off between the two tasks.*

4. Channel Discrimination

In this section, we focus on the problem of channel discrimination with the aid of a fixed input sequence $x^n \in \mathcal{X}^n$, also known as controlled sensing [34]. We review classical results from the literature, mainly due to Shannon, Gallager and Berlekamp [29], and adapt them to the setting considered here.

With knowledge of x^n, channel discrimination boils down to simple binary hypothesis testing between $P_0^n(\cdot|x^n)$ and $P_1^n(\cdot|x^n)$. It is known that the likelihood ratio test (LRT) is the optimal (deterministic) test in this setting. The decision region can be written in terms of the log-likelihood ration (LLR) as

$$\mathcal{A}_n = \left\{ z^n \in \mathcal{Z}^n : \log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)} \leq \gamma_n \right\} \tag{37}$$

for some threshold $\gamma_n \in \mathbb{R}$, which can be tuned to achieve different trade-offs between the two types of error probability (we omit the dependency of the decision region and error probabilities on the message index w in this part, as we focus on only one transmitted sequence. This is equivalent to having a codebook of size $M_n = 1$). Now suppose that the channel is used once, i.e., $n = 1$. In analyzing the error probabilities, the following function, defined on $s \in [0, 1]$ for every input symbol $x \in \mathcal{X}$, is instrumental:

$$\mu_x(s) = \log \sum_{z \in \mathcal{Z}} P_0(z|x)^{1-s} P_1(z|x)^s. \tag{38}$$

$\mu_x(s)$ is the cumulant-generating function of the LLR $\log \frac{P_1(Z|x)}{P_0(Z|x)}$, evaluated under $Z \sim P_0(\cdot|x)$. The first and second derivatives of $\mu_x(s)$, denoted by $\mu_x'(s)$ and $\mu_x''(s)$, respectively, are given by

$$\mu_x'(s) = \sum_{z \in \mathcal{Z}} P_s(z|x) \log \frac{P_1(z|x)}{P_0(z|x)} \tag{39}$$
$$\mu_x''(s) = \sum_{z \in \mathcal{Z}} P_s(z|x) \left(\log \frac{P_1(z|x)}{P_0(z|x)} \right)^2 - \left(\mu_x'(s) \right)^2 \tag{40}$$

which coincide, respectively, with the mean and variance of $\log \frac{P_1(Z|x)}{P_0(Z|x)}$, evaluated under $Z \sim P_s(\cdot|x)$. It follows that $\mu_x''(s) \geq 0$ always holds. Moreover, it can be verified that $\mu_x''(s) > 0$ unless $\frac{P_1(z|x)}{P_0(z|x)}$ is a constant over all z. Since we have assumed that the two channels have the same support, this would occur only if they are identical. Otherwise,

if $P_1(\cdot|x)$ and $P_0(\cdot|x)$ are distinct, then $\mu_x(s)$ is a strictly convex function on $s \in [0,1]$. The following identities can be verified from the expressions in (38) and (39)

$$D(P_s(\cdot|x)\|P_0(\cdot|x)) = s\mu'_x(s) - \mu_x(s) \tag{41}$$
$$D(P_s(\cdot|x)\|P_1(\cdot|x)) = (s-1)\mu'_x(s) - \mu_x(s). \tag{42}$$

When the channel is used multiple times, in which a sequence x^n is sent over n channel uses, multi-letter extensions of the above quantities become relevant. We denote these by $\mu_{x^n}(s)$, $\mu'_{x^n}(s)$ and $\mu''_{x^n}(s)$. Since the channels $P_0^n(\cdot|x^n)$ and $P_1^n(\cdot|x^n)$ are memoryless, it can be verified that

$$\mu_{x^n}(s) = \sum_{i=1}^{n} \mu_{x_i}(s) = n \sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu_x(s) \tag{43}$$
$$\mu'_{x^n}(s) = \sum_{i=1}^{n} \mu'_{x_i}(s) = n \sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu'_x(s) \tag{44}$$
$$\mu''_{x^n}(s) = \sum_{i=1}^{n} \mu''_{x_i}(s) = n \sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu''_x(s). \tag{45}$$

The first two identities follow directly from the corresponding definitions, while noting that the tilted channel $P_0^n(\cdot|x^n)$ is memoryless for every s. The last identity is obtained by noting that $\mu''_{x^n}(s)$ is the variance of a sum of independent random variables and, therefore, is equal to the sum of the individual variances. Combining the first two identities with (41) and (42), it follows that

$$nD(P_s\|P_0|\hat{P}_{x^n}) = s\mu'_{x^n}(s) - \mu_{x^n}(s) \tag{46}$$
$$nD(P_s\|P_1|\hat{P}_{x^n}) = (s-1)\mu'_{x^n}(s) - \mu_{x^n}(s). \tag{47}$$

These conditional divergence terms are useful for writing bounds on the two types of error probabilities.

Lemma 1. *Fix n and $x^n \in \mathcal{X}^n$. There exists a decision region $\mathcal{A}_n \subset \mathcal{Z}^n$ and $s \in (0,1)$ such that*

$$\varepsilon_{0,n} \le \exp\{-nD(P_s\|P_0|\hat{P}_{x^n})\} \tag{48}$$
$$\varepsilon_{1,n} \le \exp\{-nD(P_s\|P_1|\hat{P}_{x^n})\} \tag{49}$$

Moreover, for every decision region $\mathcal{A}_n \subset \mathcal{Z}^n$ and $s \in (0,1)$, at least one of the inequalities

$$\varepsilon_{0,n} > \frac{1}{4}\exp\left\{-nD(P_s\|P_0|\hat{P}_{x^n}) - s\sqrt{2\mu''_{x^n}(s)}\right\} \tag{50}$$
$$\varepsilon_{1,n} > \frac{1}{4}\exp\left\{-nD(P_s\|P_1|\hat{P}_{x^n}) - (1-s)\sqrt{2\mu''_{x^n}(s)}\right\} \tag{51}$$

must hold.

The above lemma essentially follows from [29] (Theorem 5) by Shannon, Gallager and Berlekamp. We find the proof quite insightful, and hence we present a version that is adapted to our setting next.

Proof. Fix a parameter $s \in (0,1)$. We choose the threshold in (37) such that

$$\mathcal{A}_n = \left\{z^n \in \mathcal{Z}^n : \log\frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)} \le \mu'_{x^n}(s)\right\}. \tag{52}$$

It follows that the first error probability can be bounded above as

$$\begin{aligned}
\varepsilon_{0,n} &= \sum_{z^n \in \mathcal{A}_n^c} P_0^n(z^n|x^n) && (53)\\
&= \sum_{z^n \in \mathcal{A}_n^c} \exp\left\{\mu_{x^n}(s) - s \log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)}\right\} P_s^n(z^n|x^n) && (54)\\
&\leq \sum_{z^n \in \mathcal{A}_n^c} \exp\left\{\mu_{x^n}(s) - s\mu'_{x^n}(s)\right\} P_s^n(z^n|x^n) && (55)\\
&\leq \exp\left\{\mu_{x^n}(s) - s\mu'_{x^n}(s)\right\}. && (56)
\end{aligned}$$

The main step in obtaining the above bound is the change of measure trick in the second line. Similarly, we bound the second error probability as

$$\begin{aligned}
\varepsilon_{1,n} &= \sum_{z^n \in \mathcal{A}_n} P_1^n(z^n|x^n) && (57)\\
&= \sum_{z^n \in \mathcal{A}_n} \exp\left\{\mu_{x^n}(s) + (1-s) \log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)}\right\} P_s^n(z^n|x^n) && (58)\\
&\leq \sum_{z^n \in \mathcal{A}_n} \exp\left\{\mu(s) + (1-s)\mu'_{x^n}(s)\right\} P_s^n(z^n|x^n) && (59)\\
&\leq \exp\left\{\mu_{x^n}(s) + (1-s)\mu'_{x^n}(s)\right\}. && (60)
\end{aligned}$$

The upper bounds in Lemma 1 follow by substituting (46) and (47) into the above bounds.

We now turn to proving the lower bounds. To this end, fix a parameter $s \in (0,1)$ and define the following subset of output sequences:

$$\mathcal{D}_n(s) = \left\{z^n \in \mathcal{Z}^n : \left|\log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)} - \mu'_{x^n}(s)\right| \leq \sqrt{2\mu''_{x^n}(s)}\right\}. \quad (61)$$

For an arbitrary decision region $\mathcal{A}_n \subset \mathcal{Z}^n$, the error probability of the first type is bounded below as

$$\begin{aligned}
\varepsilon_{0,n} &\geq \sum_{z^n \in \mathcal{A}_n^c \cap \mathcal{D}_n(s)} P_0^n(z^n|x^n) && (62)\\
&= \sum_{z^n \in \mathcal{A}_n^c \cap \mathcal{D}_n(s)} \exp\left\{\mu_{x^n}(s) - s \log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)}\right\} P_s^n(z^n|x^n) && (63)\\
&\geq \exp\left\{\mu_{x^n}(s) - s\mu'_{x^n}(s) - s\sqrt{2\mu''_{x^n}(s)}\right\} \sum_{z^n \in \mathcal{A}_n^c \cap \mathcal{D}_n(s)} P_s^n(z^n|x^n) && (64)
\end{aligned}$$

where, in the second line, we use the change of measure argument used in the upper bound, while the last inequality follows from the definition of $\mathcal{D}_n(s)$. Similarly, we bound the second error probability as

$$\begin{aligned}
\varepsilon_{1,n} &= \sum_{z^n \in \mathcal{A}_n \cap \mathcal{D}_n(s)} P_1^n(z^n|x^n) && (65)\\
&= \sum_{z^n \in \mathcal{A}_n \cap \mathcal{D}_n(s)} \exp\left\{\mu_{x^n}(s) + (1-s) \log \frac{P_1^n(z^n|x^n)}{P_0^n(z^n|x^n)}\right\} P_s^n(z^n|x^n) && (66)\\
&\geq \exp\left\{\mu_{x^n}(s) + (1-s)\mu'_{x^n}(s) - (1-s)\sqrt{2\mu''_{x^n}(s)}\right\} \sum_{z^n \in \mathcal{A}_n \cap \mathcal{D}_n(s)} P_s^n(z^n|x^n). && (67)
\end{aligned}$$

We now proceed by observing that the probability of $\mathcal{D}_n(s)$ under $P_s^n(\cdot|x^n)$ is lower bounded as follows:

$$P_s^n(\mathcal{D}_n(s)|x^n) = 1 - \mathbb{P}_{Z^n \sim P_s^n(\cdot|x^n)}\left[\left|\log \frac{P_1^n(Z^n|x^n)}{P_0^n(Z^n|x^n)} - \mu'_{x^n}(s)\right|^2 > 2\mu''_{x^n}(s)\right] > \frac{1}{2} \tag{68}$$

where we use Markov's (or Chebyshev's) inequality, combined with the fact that $\mu'_{x^n}(s)$ and $\mu''_{x^n}(s)$ are the mean and variance of $\log \frac{P_1^n(Z^n|x^n)}{P_0^n(Z^n|x^n)}$ under $Z^n \sim P_s^n(\cdot|x^n)$. Therefore, it follows that

$$P_s^n(\mathcal{D}_n(s) \cap \mathcal{A}_n^c|x^n) + P_s^n(\mathcal{D}_n(s) \cap \mathcal{A}_n|x^n) > \frac{1}{2}. \tag{69}$$

which, in turn, implies that either $P_s^n(\mathcal{D}_n(s) \cap \mathcal{A}_n^c|x^n) > \frac{1}{4}$ or $P_s^n(\mathcal{D}_n(s) \cap \mathcal{A}_n|x^n) > \frac{1}{4}$ must hold. Combining this with the above lower bounds on the error probabilities completes the proof. □

We conclude this section by rewriting the result in Lemma 1 in a form that will be more useful for us when proving Theorem 1. To this end, and for fixed n and $x^n \in \mathcal{X}^n$, define the region

$$\mathcal{E}_n(x^n) = \left\{(E_{0,n}, E_{1,n}) : E_{0,n} = -\frac{1}{n}\log \varepsilon_{0,n} \text{ and } E_{1,n} = -\frac{1}{n}\log \varepsilon_{1,n}, \text{ for some } \mathcal{A}_n \subset \mathcal{Z}^n\right\}. \tag{70}$$

This can be thought of as a non-asymptotic error exponent trade-off region for a fixed input sequence x^n. An inner bound and outer bound for this region are obtained from Lemma 1 as follows.

Corollary 1. *$\mathcal{E}_n(x^n)$ includes the region given by all non-negative tuples $(E_{0,n}, E_{1,n})$ satisfying*

$$E_{0,n} \leq D(P_s\|P_0|\hat{P}_{x^n}) \tag{71}$$
$$E_{1,n} \leq D(P_s\|P_1|\hat{P}_{x^n}) \tag{72}$$

for some $s \in (0,1)$, and is included in the region given by all non-negative tuples satisfying

$$E_{0,n} \leq D(P_s\|P_0|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}} \tag{73}$$
$$E_{1,n} \leq D(P_s\|P_1|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}} \tag{74}$$

for some $s \in (0,1)$, where c is a finite non-negative constant.

Proof. The inner bound follows directly by rewriting the error probability upper bounds in (48) and (49). The outer bound, on the other hand, can be shown to hold from the error probability lower bounds in (50) and (51), as we demonstrate next. We start by restating a slightly loosened version of these bounds.

Fix a channel discrimination decision region, associated with the tuple $(E_{0,n}, E_{0,n})$. Then for every $s \in (0,1)$, at least one of the following inequalities must hold:

$$E_{0,n} < D(P_s\|P_0|\hat{P}_{x^n}) + \sqrt{\frac{2}{n}\sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu''_x(s)} + \frac{1}{n}\log 4 \tag{75}$$

$$E_{1,n} < D(P_s\|P_1|\hat{P}_{x^n}) + \sqrt{\frac{2}{n}\sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu''_x(s)} + \frac{1}{n}\log 4. \tag{76}$$

Recalling that $\mu''_x(s)$ is the variance of the log-likelihood under the tilted distribution,

we have

$$\sum_{x\in\mathcal{X}} \hat{P}_{x^n}(x)\mu_x''(s) \leq \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \hat{P}_{x^n}(x)P_s(z|x)\left(\log\frac{P_1(z|x)}{P_0(z|x)}\right)^2 \tag{77}$$

$$\leq \max_{(x,z)\in\mathcal{X}\times\mathcal{Z}}\left(\log\frac{P_1(z|x)}{P_0(z|x)}\right)^2 \tag{78}$$

where the right-most upper bound is finite due to the assumption that $P_0(z|x)P_1(z|x) \neq 0$ for every $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Therefore, the bounds in (75) and (76) imply

$$E_{0,n} < D(P_s\|P_0|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}} \tag{79}$$

$$E_{1,n} < D(P_s\|P_1|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}} \tag{80}$$

for some finite non-negative constant c. Now, since at least one of these bounds must hold for every s, it follows that $0 \leq E_{0,n} \leq D(P_1\|P_0|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}}$ and $0 \leq E_{1,n} \leq D(P_0\|P_1|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}}$. In the proof of Lemma 3 in Section 6, we shall see that $D(P_s\|P_0|\hat{P}_{x^n})$ is continuous and strictly increasing in $s \in [0,1]$, from 0 to $D(P_1\|P_0|\hat{P}_{x^n})$. Therefore, there exists $s' \in [0,1]$ such that $E_{0,n} = D(P_{s'}\|P_0|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}}$. For such a choice of s', we must have $E_{1,n} \leq D(P_{s'}\|P_1|\hat{P}_{x^n}) + \frac{c}{\sqrt{n}}$. This concludes the proof. □

5. Proof of Theorem 1

Equipped with the bounds on the error probabilities of channel discrimination, we now proceed to prove the result in Theorem 1. We start with the achievability and then present the converse.

5.1. Achievability

We start with a channel coding achievability bound under a general input sequence set constraint.

Lemma 2. *Fix an input distribution $P_X \in \mathcal{P}(\mathcal{X})$ and block length n. Let $\mathcal{B}_n$ be a subset of $\mathcal{X}^n$. Then, there exists a codebook $\mathcal{C}_n \subseteq \mathcal{B}_n$ of size M_n such that for every $\tau > 0$, the average error probability satisfies*

$$p_{e,n}^{av} \times \mathbb{P}[X^n \in \mathcal{B}_n] \leq \mathbb{P}[\imath(X^n;Y^n) \leq n\tau + \log M_n] + \exp\{-n\tau\} \tag{81}$$

where $(X^n, Y^n) \sim P_X^n \times P_{Y|X}^n$, and $\imath(x^n;y^n) = \log\frac{P_{Y|X}^n(y^n|x^n)}{P_Y^n(y^n)}$ is the information density.

The unconstrained version of the above result, i.e., with $\mathcal{B}_n = \mathcal{X}^n$, was first derived by Shannon in [37] using a random coding argument. In Appendix A, we present a proof for the constrained version. A similar yet stronger result, known as Feinstein's lemma, holds but for the maximal error probability, and is proved using Feinstein's maximal coding technique (see [38] for a modern treatment of these results).

We are now in a position to show that every tuple that lies in the region described in Theorem 1 is achievable in the sense of Definition 1. Let (R, E_0, E_1) be one such tuple. This means that

$$I(P_X, P_{Y|X}) \geq R,\ D(P_s\|P_0|P_X) \geq E_0,\ D(P_s\|P_1|P_X) \geq E_1, \text{ and } E_{X\sim P_X}[b(X)] \leq B \tag{82}$$

hold for some input distribution $P_X \in \mathcal{P}(\mathcal{X})$ and constant $s \in [0,1]$. In what follows, we fix a block length n and a pair (P_X, s) sastifying (82). We define the following subsets of input sequences

$$\mathcal{B}_{0,n}(P_X, s) = \left\{x^n \in \mathcal{X}^n : D(P_s \| P_0 | \hat{P}_{x^n}) \geq D(P_s \| P_0 | P_X) - \epsilon\right\} \tag{83}$$

$$\mathcal{B}_{1,n}(P_X, s) = \left\{x^n \in \mathcal{X}^n : D(P_s \| P_0 | \hat{P}_{x^n}) \geq D(P_s \| P_0 | P_X) - \epsilon\right\} \tag{84}$$

$$\mathcal{B}_{b,n}(P_X) = \left\{x^n \in \mathcal{X}^n : \mathbb{E}_{X\sim\hat{P}_{x^n}}[b(X)] \leq \mathbb{E}_{X\sim P_X}[b(X)] + \epsilon\right\}. \tag{85}$$

for some small constant $\epsilon > 0$. The inner bound in Corollary 1, combined with (82)–(84), implies that for every $x^n \in \mathcal{B}_{0,n}(P_X, s) \cap \mathcal{B}_{1,n}(P_X, s)$, there exists a decision region that satisfies

$$-\frac{1}{n}\log \varepsilon_{0,n}(x^n) \geq D(P_s \| P_0 | \hat{P}_{x^n}) \geq D(P_s \| P_0 | P_X) - \epsilon \geq E_0 - \epsilon \tag{86}$$

$$-\frac{1}{n}\log \varepsilon_{1,n}(x^n) \geq D(P_s \| P_1 | \hat{P}_{x^n}) \geq D(P_s \| P_1 | P_X) - \epsilon \geq E_1 - \epsilon. \tag{87}$$

Note that for convenience, we highlight the dependency of $\varepsilon_{0,n}(x^n)$ and $\varepsilon_{1,n}(x^n)$ on x^n instead of the message w. By imposing the codebook constraint $\mathcal{C}_n \subseteq \mathcal{B}_n$, where

$$\mathcal{B}_n = \mathcal{B}_{0,n}(P_X, s) \cap \mathcal{B}_{1,n}(P_X, s) \cap \mathcal{B}_{b,n}(P_X), \tag{88}$$

the desired channel discrimination error exponents (E_0, E_1) are achieved, while satisfying the input cost constraint (strictly speaking, an additional cost of ϵ is incurred. However, ϵ can be made to vanish as n grows large). It remains to show that the rate R is also achievable. To show this, we use Lemma 2 while setting $\frac{1}{n}\log M_n = I(P_X, P_{Y|X}) - (\tau + \delta)$ for some $\delta > 0$, from which we obtain

$$p_{\mathrm{e},n}^{\mathrm{av}} \leq \frac{\mathbb{P}\left[\frac{1}{n}\imath(X^n; Y^n) \leq I(P_X, P_{Y|X}) - \delta\right] + \exp\{-n\tau\}}{\mathbb{P}[X^n \in \mathcal{B}_n]}. \tag{89}$$

Note that since $X^n \sim P_X^n$, then, by the weak law of large numbers (WLLN), we have

$$\mathbb{P}\left[\frac{1}{n}\imath(X^n; Y^n) \leq I(P_X, P_{Y|X}) - \delta\right] \to 0, \text{ as } n \to \infty. \tag{90}$$

Similarly, by the WLLN, we also have

$$\mathbb{P}[X^n \in \mathcal{B}_{0,n}(P_X, s)] \to 1,\ \mathbb{P}[X^n \in \mathcal{B}_{1,n}(P_X, s)] \to 1,\ \mathbb{P}[X^n \in \mathcal{B}_{b,n}(P_X, s)] \to 1, \text{ as } n \to \infty. \tag{91}$$

Combining this with the fact that $\mathbb{P}[X^n \in \mathcal{B}_n]$ is lower-bounded by

$$\mathbb{P}[X^n \in \mathcal{B}_n] \geq 1 - \mathbb{P}[X^n \notin \mathcal{B}_{0,n}(P_X, s)] - \mathbb{P}[X^n \notin \mathcal{B}_{0,n}(P_X, s)] - \mathbb{P}[X^n \notin \mathcal{B}_{0,n}(P_X, s)], \tag{92}$$

it directly follows that $\mathbb{P}[X^n \in \mathcal{B}_n] \to 1$ as $n \to \infty$, and therefore $p_{\mathrm{e},n}^{\mathrm{av}} \to 0$ as $n \to \infty$. We conclude that for any $\epsilon > 0$, and by making n large enough, there exists a codebook of M_n codewords in $\mathcal{B}_n$ such that

$$p_{\mathrm{e},n}^{\mathrm{av}} \leq \frac{1}{4}\epsilon \quad \text{and} \quad \frac{1}{n}\log M_n = I(P_X, P_{Y|X}) - (\tau + \delta) \geq R - (\epsilon - \frac{1}{n}\log 2). \tag{93}$$

Note that the last inequality holds from (82) and by choosing τ, δ and n appropriately.

Finally, using an expurgation argument, where we remove the worst half of the codewords (see, e.g, [39] (Ch. 7.7)), it follows that there exists a codebook of size $M_n' = M_n/2$ in $\mathcal{B}_n$ for which $\frac{1}{n}\log M_n' \geq R - \epsilon$ and $p_{\mathrm{e},n} \leq 4p_{\mathrm{e},n}^{\mathrm{av}} \leq \epsilon$. This concludes the proof of achievability.

5.2. Converse

We now turn to the converse. First, for every E_0, E_1 and B, we define the following sets of distributions on $\mathcal{X}$ (i.e., subsets of $\mathcal{P}(\mathcal{X})$), which will prove useful further on:

$$\mathcal{P}_E(\mathcal{X}, E_0, E_1) = \bigcup_{s\in[0,1]} \{P_X \in \mathcal{P}(\mathcal{X}) : D(P_s\|P_0|P_X) \geq E_0,\ D(P_s\|P_1|P_X) \geq E_1\} \tag{94}$$

$$\mathcal{P}_b(\mathcal{X}, B) = \big\{P_X \in \mathcal{P}(\mathcal{X}) : \mathbb{E}_{X\sim P_X}[b(X)] \leq B\big\}. \tag{95}$$

Now suppose that the rate–exponent tuple (R, E_0, E_1) is achievable. Then, for every $\delta > 0$, and by making n large enough, there exists an (n, M_n)-code in which every codeword $x^n \in \mathcal{C}_n$ satisfies

$$E_0 - \delta \leq -\frac{1}{n}\log \varepsilon_{0,n}(x^n) \leq D(P_s\|P_0\|\hat{P}_{x^n}) + \delta \tag{96}$$

$$E_1 - \delta \leq -\frac{1}{n}\log \varepsilon_{1,n}(x^n) \leq D(P_s\|P_1\|\hat{P}_{x^n}) + \delta \tag{97}$$

for some parameter $s \in (0, 1)$, which may depend on the codeword. The left-hand-side inequalities follow from Definition 1, while the right-hand-side inequalities follow from the outer bound in Corollary 1. The constraints in (96) and (97) imply that every codeword x^n in $\mathcal{C}_n$ must have a type $\hat{P}_{x^n}$ that satisfies

$$\hat{P}_{x^n} \in \mathcal{P}_E(\mathcal{X}, E_0 - \epsilon, E_1 - \epsilon) \cap \mathcal{P}_b(\mathcal{X}, B) \tag{98}$$

where $\epsilon = 2\delta$. Under this codeword constraint, our next goal is find an upper bound on the codebook size M_n, given the additional constraint $p_{e,n}(\mathcal{C}_n) \leq \epsilon$. A challenge here is that the set of input distributions in (98) is non-convex in general due to the union in (94). This prohibits us from applying existing approaches for proving channel coding converses under additive cost [35,36] or additive sensing distortion [6,9,11] constraints, as they crucially rely on the convexity of the underlying input distribution constraint set.

To circumvent this challenge, we first observe that $\mathcal{C}_n$ can be partitioned into constant-composition sub-codebooks, each comprising codewords of the same type. Let $P_X \in \mathcal{P}_n(\mathcal{X})$ be the type associate with the largest of such sub-codebooks, denoted by $\mathcal{C}_n(P_X)$, and let $M_n(P_X)$ be the corresponding number of codewords in this sub-codebook. It is straightforward to see

$$M_n \leq |\mathcal{P}_n(\mathcal{X})| M_n(P_X) \leq (n+1)^{|\mathcal{X}|} M_n(P_X) \tag{99}$$

where the right-most inequality is due to (3). Therefore, bounding the rate $\frac{1}{n}\log M_n$ of $\mathcal{C}_n$ is asymptotically equivalent to bounding the rate $\frac{1}{n}\log M_n(P_X)$ of $\mathcal{C}_n(P_X)$, since $\frac{|\mathcal{X}|}{n}\log(n+1)$ tends to zero as n approaches infinity. Going forward, we focus on bounding $\frac{1}{n}\log M_n(P_X)$, which is more suited to our purpose as it can be performed without relying on the convexity of the underlying input distribution constraint set.

Note that $p_{e,n}(\mathcal{C}_n) \leq \epsilon$ implies $p^{\text{av}}_{e,n}(\mathcal{C}_n) \leq \epsilon$, which in turn implies $p^{\text{av}}_{e,n}(\mathcal{C}_n(P_X)) \leq \epsilon$, which holds because $\mathcal{C}_n(P_X) \subseteq \mathcal{C}_n$. Therefore, from Fano's inequality (see, e.g., [36] Section 3.1.4), we obtain

$$\frac{1}{n}\log M_n(P_X) \leq \frac{1}{n} I(X^n; Y^n) + \frac{1}{n} + \frac{\epsilon}{n}\log M_n(P_X) \tag{100}$$

where X^n in (100) is uniformly distributed on the constant-composition sub-codebook $\mathcal{C}^n(P_X)$, since the underlying message is uniform.We proceed by bounding the per-channel-use mutual information as

$$\frac{1}{n} I(X^n; Y^n) \leq \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y_i) \tag{101}$$

$$= \frac{1}{n} \sum_{i=1}^{n} I(P_{X_i}, P_{Y|X}) \tag{102}$$

$$\leq I(Q_X, P_{Y|X}) \tag{103}$$

where P_{X_i} is obtained by marginalizing P_{X^n}, while Q_X in (103) is given by the following mixture:

$$Q_X(x) = \frac{1}{n} \sum_{i=1}^{n} P_{X_i}(x), \ \forall x \in \mathcal{X}. \tag{104}$$

The inequality in (103) holds due to the concavity of the mutual information in the input distribution for a fixed channel and by Jensen's inequality. The next step is to show that the mixture input distribution Q_X coincides with the sub-codebook's type P_X. This is seen from the following:

$$Q_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M_n(P_X)} \sum_{w=1}^{M_n(P_X)} \mathbb{1}[x_i(w) = x] \tag{105}$$

$$= \frac{1}{M_n(P_X)} \sum_{w=1}^{M_n(P_X)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[x_i(w) = x] \tag{106}$$

$$= \frac{1}{M_n(P_X)} \sum_{w=1}^{M_n(P_X)} \hat{P}_{x^n(w)}(x) \tag{107}$$

$$= P_X(x). \tag{108}$$

Putting together (108), (103), (100) and (99), we obtain

$$\frac{1}{n} \log M_n \leq \frac{1}{n} \log M_n(P_X) + \frac{|\mathcal{X}|}{n} \log(n+1) \tag{109}$$

$$\leq I(P_X, P_{Y|X}) + \frac{1}{n}\big(1 + \epsilon \log M_n(P_X) + \log(n+1)\big). \tag{110}$$

Since the rate R is achievable, then $R - \delta \leq \frac{1}{n} \log M_n \leq I(P_X, P_{Y|X}) + \delta$ holds for sufficiently large n. Recall that P_X must be in the set (98), from which we conclude that (R, E_0, E_1) must satisfy

$$R \leq I(P_X, P_{Y|X}) + \epsilon \tag{111}$$

$$E_0 \leq D(P_s \| P_0 | P_X) + \epsilon \tag{112}$$

$$E_1 \leq D(P_s \| P_1 | P_X) + \epsilon \tag{113}$$

for some $s \in [0, 1]$ and $P_X \in \mathcal{P}_b(\mathcal{X}, B)$. These inequalities hold for every $\epsilon > 0$, as ϵ can be made as small as desired in the above proof by making n sufficiently large. Therefore, we conclude that (R, E_0, E_1) must satisfy the above inequalities while setting $\epsilon = 0$. This concludes the converse proof. □

6. Minimax and Neyman–Pearson

In this section, we consider minimax and Neyman–Pearson metrics for channel discrimination and characterize the resulting rate–exponent trade-offs. As we will see, the emerging trade-offs are special cases of the general trade-off in Theorem 1.

6.1. Minimax Discrimination Criterion

In some applications, the two types of discrimination error probabilities are treated equally, and one wishes to control (i.e., minimize) the maximum of the two, defined for a given code as

$$\varepsilon_n = \max\{\varepsilon_{0,n}, \varepsilon_{1,n}\}. \tag{114}$$

This gives rise to the following asymptotic trade-off.

Definition 2. *A tuple (R, E) is said to be achievable under the minimax channel discrimination criterion if for every $\epsilon > 0$, and provided that n is large enough, there exists an (n, M_n)-code with*

$$p_{\mathrm{e},n} \leq \epsilon, \quad \frac{1}{n}\log M_n \geq R - \epsilon \quad \textit{and} \quad \frac{1}{n}\log\frac{1}{\varepsilon_n} \geq E - \epsilon.$$

The rate–exponent region $\mathcal{R}_{\mathrm{mini}}$ is the closure of the set of all achievable pairs (R, E).

It can be verified that $(R, E) \in \mathcal{R}_{\mathrm{mini}}$ if and only if $(R, E, E) \in \mathcal{R}$. From this observation and Theorem 1, it is readily seen that $\mathcal{R}_{\mathrm{mini}}$ is given by the set of all non-negative tuples (R, E) such that

$$R \leq I(P_X, P_{Y|X}) \tag{115}$$

$$E \leq \max_{s\in[0,1]} \min\{D(P_s\|P_0|P_X), D(P_s\|P_1|P_X)\} \tag{116}$$

for some input distribution $P_X \in \mathcal{P}(\mathcal{X})$ that satisfies $\mathbb{E}_{X\sim P_X}[b(X)] \leq B$. Before we proceed, we define

$$\begin{aligned} C(P_0, P_1|P_X) &\triangleq -\min_{s\in[0,1]} \sum_{x\in\mathcal{X}} P_X(x) \log \sum_{z\in\mathcal{Z}} P_0(z|x)^{1-s} P_1(z|x)^s \\ &= -\min_{s\in[0,1]} \sum_{x\in\mathcal{X}} P_X(x)\mu_x(s) \end{aligned} \tag{117}$$

The above quantity is a generalized form of the Chernoff information (see, e.g., [39] (Ch. 11.9)), which is suited for channel discrimination. In the following result, we establish the equivalence between $C(P_0, P_1|P_X)$ and the right-hand side of (116), which parallels a known equivalence in classical hypothesis testing.

Lemma 3. *The following identity holds:*

$$\max_{s\in[0,1]} \min\{D(P_s\|P_0|P_X), D(P_s\|P_1|P_X)\} = C(P_0, P_1|P_X) \tag{118}$$

Proof. Taking the expectations of (41) and (42) with respect to $X \sim P_X$, we obtain

$$D(P_s\|P_0|P_X) = \sum_{x\in\mathcal{X}} P_X(x)\big(s\mu'_x(s) - \mu_x(s)\big) = s\mu'_{P_X}(s) - \mu_{P_X}(s) \tag{119}$$

$$D(P_s\|P_1|P_X) = \sum_{x\in\mathcal{X}} P_X(x)\big((s-1)\mu'_x(s) - \mu_x(s)\big) = (s-1)\mu'_{P_X}(s) - \mu_{P_X}(s) \tag{120}$$

from which it follows directly that

$$\frac{\mathrm{d}}{\mathrm{d}s} D(P_s\|P_0|P_X) = \sum_{x\in\mathcal{X}} P_X(x) s\mu''_x(s) = s\mu''_{P_X}(s) \tag{121}$$

$$\frac{\mathrm{d}}{\mathrm{d}s} D(P_s\|P_1|P_X) = \sum_{x\in\mathcal{X}} P_X(x)(s-1)\mu''_x(s) = (s-1)\mu''_{P_X}(s). \tag{122}$$

Recall from Section 4 that $\mu''_x(s) > 0$, unless the two channels $P_0(\cdot|x)$ and $P_1(\cdot|x)$ are identical under x. If such inputs exist and P_X is supported on them only, then we will have

$D(P_s\|P_0|P_X) = D(P_s\|P_0|P_X) = 0$ for every $s \in [0,1]$, and hence $C(P_0, P_1|P_X) = 0$, and there is nothing left to prove. Therefore, we assume that $P_X(x) > 0$ and $\mu''_x(s) > 0$ for at least one x. From (121) and (122), we see that $D(P_s\|P_0|P_X)$ is strictly increasing in s and $D(P_s\|P_1|P_X)$ is strictly decreasing in s; hence,

$$\min\{D(P_s\|P_0|P_X), D(P_s\|P_1|P_X)\} = \begin{cases} D(P_s\|P_0|P_X), & \text{if } D(P_s\|P_0|P_X) \le D(P_s\|P_1|P_X) \\ D(P_s\|P_1|P_X), & \text{if } D(P_s\|P_0|P_X) > D(W_s\|P_1|P_X). \end{cases} \tag{123}$$

It follows that the maximum is achieved at $s = s^\star$ that satisfies $D(P_{s^\star}\|P_0|P_X) = D(P_{s^\star}\|P_1|P_X)$, for which we also have $\mu'_{P_X}(s^\star) = 0$. Due to $\mu''_{P_X}(s) > 0$ and $\mu_{P_X}(0) = \mu_{P_X}(1) = 0$, it follows that

$$\max_{s\in[0,1]} \min\{D(P_s\|P_0|P_X), D(P_s\|P_1|P_X)\} = -\min_{s\in[0,1]} \mu_{P_X}(s) \tag{124}$$

which completes the proof of the lemma. □

We are now ready to present the rate–exponent trade-off under the minimax error criterion.

Theorem 2. *$\mathcal{R}_{\text{mini}}$ is given by the set of all non-negative pairs (R, E) such that*

$$R \le I(P_X, P_{Y|X}) \tag{125}$$

$$E \le C(P_0, P_1|P_X) \tag{126}$$

for some input distribution $P_X \in \mathcal{P}(\mathcal{X})$ that satisfies $\mathbb{E}_{X\sim P_X}[b(X)] \le B$.

The proof follows directly from the above discussion.

Example 3. *For the setting in Example 1, we have*

$$C(P_0, P_1|P_X) = -\min_{s\in[0,1]} \rho \log\Big((1-q)^{1-s}q^s + q^{1-s}(1-q)^s\Big). \tag{127}$$

Recall that $\mathbb{E}_{X\sim P_X}[\mu_X(s)] = \rho\log\big((1-q)^{1-s}q^s + q^s(1-q)^{1-s}\big)$ is strictly convex in s. Since here we have $\mathbb{E}_{X\sim P_X}[\mu_X(s)] = \mathbb{E}_{X\sim P_X}[\mu_X(1-s)]$, it follows that $s = 0.5$ is the minimizer in (127), and we obtain

$$C(P_0, P_1|P_X) = -\rho\log\left(2\sqrt{(1-q)q}\right) = -\rho\log e^{-D(0.5\|q)} = \rho D(0.5\|q). \tag{128}$$

Therefore, $\mathcal{R}_{\text{mini}}$ in this case is described by

$$R \le H(\rho * p) - H(p) \tag{129}$$

$$E \le \rho D(0.5\|q) \tag{130}$$

for some $\rho \le B$. This recovers the result in [17] (Theorem 1). As seen in Example 1, there is generally a trade-off between R and E whenever $B > 0.5$. On the other hand, for the setting in Example 2, we have

$$C(P_0, P_1|P_X) = -\min_{s\in[0,1]} \log\Big((1-p_0)^{1-s}(1-p_1)^{1-s} + p_0^s p_1^s\Big). \tag{131}$$

Since there is no trade-off between R and E in this case, the corresponding $\mathcal{R}_{\text{mini}}$ is a rectangle.

6.2. Neyman–Pearson Discrimination Criterion

Here, we consider the case where the two types of discrimination errors are treated unequally. We adopt the Neyman–Pearson criterion, where the focus is on minimizing one type of error while keeping the other under control. Here, we choose to minimize the type II error probability while requiring that the type I error probability does not exceed a desired threshold $\alpha \in (0,1)$.

For a codebook $\mathcal{C}_n$ and given that the codeword x^n has been sent, the decision region $\mathcal{A}_n(x^n)$ is chosen according to the above criterion, and the resulting type II discrimination error is given by

$$\beta_{\alpha,n}(x^n) \triangleq \min_{\mathcal{A}_n(x^n):\varepsilon_{0,n}(x^n)\leq\alpha} \varepsilon_{1,n}(x^n). \tag{132}$$

As argued in Section 2.1, since it is not known beforehand which codeword in $\mathcal{C}_n$ will be sent, we take the maximum over all codewords in $\mathcal{C}_n$ and obtain an error probability of

$$\beta_{\alpha,n} \triangleq \max_{x^n\in\mathcal{C}_n} \beta_{\alpha,n}(x^n). \tag{133}$$

Under the Neyman–Pearson criterion, the asymptotic trade-off is formalized as follows.

Definition 3. *Under the Neyman–Pearson discrimination criterion, (R, E_α) is said to be achievable if for every $\epsilon > 0$, and provided that n is large enough, there exists an (n, M_n)-code with*

$$p_{e,n} \leq \epsilon, \quad \frac{1}{n}\log M_n \geq R - \epsilon \quad \textit{and} \quad \frac{1}{n}\log\frac{1}{\beta_{\alpha,n}} \geq E_\alpha - \epsilon.$$

The rate–exponent region $\mathcal{R}_\alpha$ is the closure of the set of all achievable pairs (R, E_α).

We now present the trade-off under the Neyman–Pearson channel discrimination criterion.

Theorem 3. *$\mathcal{R}_\alpha$ for any $\alpha \in (0,1)$ is given by the set of all non-negative pairs (R, E) such that*

$$R \leq I(P_X, P_{Y|X}) \tag{134}$$

$$E \leq D(P_0\|P_1|P_X) \tag{135}$$

for some input distribution $P_X \in \mathcal{P}(\mathcal{X})$ that satisfies $\mathbb{E}_{X\sim P_X}[b(X)] \leq B$.

The above theorem follows from Theorem 1. In particular, we know that the pair (R, E_1) must satisfy $R \leq I(P_X, P_{Y|X})$ and $E_1 \leq D(P_0\|P_1|P_X)$ for some $P_X \in \mathcal{P}(\mathcal{X})$ that satisfies $\mathbb{E}_{X\sim P_X}[b(X)] \leq B$. Achievability follows by choosing an arbitrarily small $s > 0$ in Theorem 1, such that $s \to 0$ as $n \to \infty$.

Example 4. *For the setting in Example 1, the rate–exponent region $\mathcal{R}_\alpha$ is given by*

$$R \leq H(\rho * p) - H(p) \tag{136}$$

$$E \leq \rho d(q\|1-q) \tag{137}$$

for some $\rho \leq B$. For the setting in Example 2, we have $R \leq H(\min\{0.5, B\} * p) - H(p)$ and $E \leq d(p_0\|p_1)$.

7. Concluding Remarks

We considered a problem of joint message communication and channel discrimination in discrete memoryless systems with an additive input cost, and we have established the optimal trade-off between the rate of reliable communication and the two types of channel discrimination error exponents. The simple instances given in Examples 1 and 2 provide

insights into the fundamental trade-off arising in JCAS systems, where the same resources are used simultaneously for both tasks.

The setting we proposed in Figure 1 can be extended to enable adaptivity by modifying the encoding function such that at every channel use i, the input X_i is made to be a function of both the message W and previous sensor observations Z^{i-1}. However, it turns out that the results we have reported here will remain unchanged. This is because adaptivity does not improve the channel discrimination error exponents in block transmissions [32,33], and will not improve the channel coding rate since the data channel does not depend on the parameter θ. This picture is different if we move beyond binary channel discrimination or if the data channel is made to depend on the parameter, as explored in [20,21].

Our setting can be further extended in several other directions. For instance, it is of interest to extend the results to channels with general (e.g., continuous) alphabets. Although we do not expect the results to change dramatically, alternative proof techniques may be required. Note that, unlike our previous preliminary works [18,19] and the works of Chang et al. [20,21], the achievability part in the current paper is directly applicable to channels with general alphabets; so is the converse part for channel discrimination. Nevertheless, our channel coding rate converse proof relies on constant-composition sub-codes and a type counting argument which, at least in its current form, only holds for discrete finite alphabets. It is of interest to try to extend the converse argument to general alphabets.

Another extension is to consider, in addition to the coding rate and discrimination exponents, the channel coding exponent (i.e., reliability function) and to study the trade-off between all three. Some progress along these lines has been reported in [21]. It may also be of interest for practical purposes to derive refined bounds and second-order asymptotics, which tend to be tighter in finite blocklength regimes. Progress along these lines for the i.i.d. state model has been recently reported in [40]. Finally, one may also consider extending the setup to incorporate variable-length sequential transmissions, which can result in significant gains for channel discrimination as recently reported in [41].

Author Contributions: Conceptualization, H.J.; methodology, H.W. and H.J.; formal analysis, H.W. and H.J.; writing—original draft preparation, H.W.; writing—review and editing, H.W. and H.J.; funding acquisition, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded in part by the European Union (ERC, IT-JCAS, 101116550). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: Han Wu would like to thank Giuseppe Durisi for introducing hypothesis testing to him when he was studying at Chalmers University of Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Lemma 2

The proof relies on random coding and threshold decoding (also known as Feinstein's decoding rule). Define the following random coding distribution on $\mathcal{X}^n$:

$$Q_{X^n}(x^n) = \frac{P_X^n(x^n)}{P_X^n(\mathcal{B}^n)} \mathbb{1}[x^n \in \mathcal{B}_n], \text{ for all } x^n \in \mathcal{X}^n. \tag{A1}$$

Let $\boldsymbol{C} = \{X^n(1), X^n(2), \ldots, X^n(M_n)\}$ be a random codebook ensemble, in which codewords are independently drawn at random with distribution Q_{X^n}. Since Q_{X^n} is supported on $\mathcal{B}_n$, every codebook in the random codebook ensemble satisfies the constraint that codewords all belong to $\mathcal{B}_n$.

For a fixed codebook $\boldsymbol{C} = \mathcal{C}_n$ and threshold $\gamma_n > 0$, and given a channel observation $Y^n = y^n$, the threshold decoder selects the unique message index $\hat{w} \in \mathcal{M}_n$ such that

$$\imath(x^n(\hat{w}); y^n) = \log \frac{P_{Y|X}^n\big(x^n(\hat{w})|y^n\big)}{P_Y^n(y^n)} > \gamma_n \tag{A2}$$

where $P_Y^n(y^n) = \mathbb{E}_{X^n \sim P_X^n}\Big[P_{Y|X}^n(y^n|X^n)\Big]$. Note that the information density $\imath(x^n(\hat{w}); y^n)$ is evaluated using P_X^n as the input distribution, and not Q_{X^n}. This is done so that the final bound is in terms of $P_{Y|X}^n \times P_X^n$, which will become more apparent later in the proof.

The threshold decoder makes an error if the information density of the transmitted codeword does not exceed γ_n, or if the information density of any other codeword does. For a given codebook $\mathcal{C}_n$, we have

$$p_{\mathrm{e},n}^{\mathrm{av}}(\mathcal{C}_n) \le \frac{1}{M_n} \sum_{w \in \mathcal{M}_n} \Big(\mathbb{P}[\imath(x^n(w); Y^n) \le \gamma_n | W = w] + \mathbb{P}[\imath(x^n(\bar{w}); Y^n) > \gamma_n, \text{ for some } \bar{w} \ne w | W = w] \Big).$$

Taking the expectation with respect to the codebook ensemble, we obtain

$$\mathbb{E}\big[p_{\mathrm{e},n}^{\mathrm{av}}(\boldsymbol{C})\big] \le \mathbb{P}[\imath(X^n(1); Y^n) \le \gamma_n | W = 1] + \sum_{\bar{w} \ne 1} \mathbb{P}[\imath(X^n(\bar{w}); Y^n) > \gamma_n | W = 1] \tag{A3}$$

This is obtained by noting that the codewords in $\boldsymbol{C}$ are independent and identically distributed; hence, it suffices to restrict to $W = 1$ and apply the union bound to obtain the second term on the right-hand side.

We now treat the two terms on the right-hand side of (A3) separately. The first is bounded as

$$\mathbb{P}[\imath(X^n(1); Y^n) \le \gamma_n | W = 1] = \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} Q_{X^n}(x^n) P_{Y|X}^n(y^n|x^n) \mathbb{1}[\imath(x^n; y^n) \le \gamma_n] \tag{A4}$$

$$\le \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} \frac{P_X^n(x^n)}{P_X^n(\mathcal{B}^n)} P_{Y|X}^n(y^n|x^n) \mathbb{1}[\imath(x^n; y^n) \le \gamma_n] \tag{A5}$$

$$= \frac{\mathbb{P}[\imath(X^n; Y^n) \le \gamma_n]}{\mathbb{P}[X^n \in \mathcal{B}_n]} \tag{A6}$$

where, in the last line, we have $(X^n, Y^n) \sim P_X^n \times P_{Y|X}^n$. As for the second term, for any $\bar{w} \ne 1$, we have

$$\mathbb{P}[\imath(X^n(\bar{w}); Y^n) > \gamma_n | W = 1] = \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} Q_{X^n}(x^n) Q_{Y^n}(y^n) \mathbb{1}\left[\frac{P_{Y|X}^n(y^n|x^n)}{P_Y^n(y^n)} > \exp\{\gamma_n\}\right] \tag{A7}$$

$$\le \exp\{-\gamma_n\} \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} Q_{X^n}(x^n) Q_{Y^n}(y^n) \times \frac{P_{Y|X}^n(y^n|x^n)}{P_Y^n(y^n)} \tag{A8}$$

where $Q_{Y^n}(y^n) = \mathbb{E}_{X^n \sim Q_{X^n}}\Big[P_{Y|X}^n(y^n|X^n)\Big]$. We bound the product $Q_{X^n}(x^n) Q_{Y^n}(y^n)$ as follows

$$Q_{X^n}(x^n)Q_{Y^n}(y^n) = Q_{X^n}(x^n)\mathbb{E}_{X^n\sim Q_{X^n}}\left[P^n_{Y|X}(y^n|X^n)\right] \tag{A9}$$

$$= \frac{\mathbb{1}[x^n \in \mathcal{B}_n]}{P^n_X(\mathcal{B}^n)} P^n_X(x^n)\mathbb{E}_{X^n\sim P^n_X}\left[P^n_{Y|X}(y^n|X^n)\frac{\mathbb{1}[X^n \in \mathcal{B}_n]}{P^n_X(\mathcal{B}_n)}\right] \tag{A10}$$

$$\leq \frac{\mathbb{1}[x^n \in \mathcal{B}_n]}{P^n_X(\mathcal{B}^n)} P^n_X(x^n)\mathbb{E}_{X^n\sim P^n_X}\left[P^n_{Y|X}(y^n|X^n)\right] \tag{A11}$$

$$= \frac{\mathbb{1}[x^n \in \mathcal{B}_n]}{\left(P^n_X(\mathcal{B}^n)\right)^2} P^n_X(x^n)P^n_Y(y^n). \tag{A12}$$

Substituting this into (A8), we obtain the following.

$$\mathbb{P}[\imath(X^n(\bar{w});Y^n) > \gamma_n|W=1] \leq \frac{\exp\{-\gamma_n\}}{\left(P^n_X(\mathcal{B}^n)\right)^2} \sum_{x^n\in\mathcal{X}^n} P^n_X(x^n)\mathbb{1}[x^n \in \mathcal{B}_n] \sum_{y^n\in\mathcal{Y}^n} P^n_{Y|X}(y^n|x^n) \tag{A13}$$

$$= \frac{\exp\{-\gamma_n\}}{P^n_X(\mathcal{B}^n)}. \tag{A14}$$

Putting everything together, we obtain the following upper bound

$$\mathbb{E}\left[p^{\text{av}}_{\text{e},n}(\boldsymbol{C})\right] \leq \frac{\mathbb{P}[\imath(X^n;Y^n) \leq \gamma_n] + (M_n-1)\exp\{-\gamma_n\}}{\mathbb{P}[X^n \in \mathcal{B}_n]}. \tag{A15}$$

The final result is obtained by loosening $M_n - 1$ to M_n, setting $\gamma_n = n\tau + \log M_n$, and from the fact that there must exist at least one codebook in the ensemble for which the upper bound holds. □

References

1. Liu, F.; Cui, Y.; Masouros, C.; Xu, J.; Han, T.X.; Eldar, Y.C.; Buzzi, S. Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 1728–1767. [CrossRef]
2. Xiong, Y.; Liu, F.; Wan, K.; Yuan, W.; Cui, Y.; Caire, G. From Torch to Projector: Fundamental Tradeoff of Integrated Sensing and Communications. *IEEE BITS Inf. Theory Mag.* **2024**, 1–13. [CrossRef]
3. Wild, T.; Braun, V.; Viswanathan, H. Joint Design of Communication and Sensing for Beyond 5G and 6G Systems. *IEEE Access* **2021**, *9*, 30845–30857. [CrossRef]
4. Sturm, C.; Wiesbeck, W. Waveform Design and Signal Processing Aspects for Fusion of Wireless Communications and Radar Sensing. *Proc. IEEE* **2011**, *99*, 1236–1259. [CrossRef]
5. Ma, D.; Shlezinger, N.; Huang, T.; Liu, Y.; Eldar, Y.C. Joint Radar-Communication Strategies for Autonomous Vehicles: Combining Two Key Automotive Technologies. *IEEE Signal Process. Mag.* **2020**, *37*, 85–97. [CrossRef]
6. Kobayashi, M.; Caire, G.; Kramer, G. Joint State Sensing and Communication: Optimal Tradeoff for a Memoryless Case. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 111–115. [CrossRef]
7. Sutivong, A.; Chiang, M.; Cover, T.; Kim, Y.H. Channel capacity and state estimation for state-dependent Gaussian channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 1486–1495. [CrossRef]
8. Choudhuri, C.; Kim, Y.H.; Mitra, U. Causal State Communication. *IEEE Trans. Inf. Theory* **2013**, *59*, 3709–3719. [CrossRef]
9. Zhang, W.; Vedantam, S.; Mitra, U. Joint Transmission and State Estimation: A Constrained Channel Coding Approach. *IEEE Trans. Inf. Theory* **2011**, *57*, 7084–7095. [CrossRef]
10. Kobayashi, M.; Hamad, H.; Kramer, G.; Caire, G. Joint State Sensing and Communication over Memoryless Multiple Access Channels. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019; pp. 270–274. [CrossRef]
11. Ahmadipour, M.; Kobayashi, M.; Wigger, M.; Caire, G. An Information-Theoretic Approach to Joint Sensing and Communication. *IEEE Trans. Inf. Theory* **2024**, *70*, 1124–1146. [CrossRef]
12. Ahmadipour, M.; Wigger, M. An Information-Theoretic Approach to Collaborative Integrated Sensing and Communication for Two-Transmitter Systems. *IEEE J. Sel. Areas Inf. Theory* **2023**, *4*, 112–127. [CrossRef]
13. Ahmadipour, M.; Wigger, M.; Shamai, S. Integrated Communication and Receiver Sensing with Security Constraints on Message and State. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, 25–30 June 2023; pp. 2738–2743. [CrossRef]

14. Günlü, O.; Bloch, M.R.; Schaefer, R.F.; Yener, A. Secure Integrated Sensing and Communication. *IEEE J. Sel. Areas Inf. Theory* **2023**, *4*, 40–53. [CrossRef]
15. Xiong, Y.; Liu, F.; Cui, Y.; Yuan, W.; Han, T.X.; Caire, G. On the Fundamental Tradeoff of Integrated Sensing and Communications Under Gaussian Channels. *IEEE Trans. Inf. Theory* **2023**, *69*, 5723–5751. [CrossRef]
16. Joudeh, H.; Caire, G. Joint Communication and State Sensing under Logarithmic Loss. In Proceedings of the 4th IEEE International Symposium on Joint Communications & Sensing (JC&S), Leuven, Belgium, 19–21 March 2024; pp. 1–6. [CrossRef]
17. Joudeh, H.; Willems, F.M.J. Joint Communication and Binary State Detection. *IEEE J. Sel. Areas Inf. Theory* **2022**, *3*, 113–124. [CrossRef]
18. Wu, H.; Joudeh, H. On Joint Communication and Channel Discrimination. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Espoo, Finland, 26 June–1 July 2022; pp. 3321–3326. [CrossRef]
19. Wu, H.; Joudeh, H. Joint Communication and Channel Discrimination. *arXiv* **2022**, arXiv:2208.07450v1.
20. Chang, M.C.; Erdogan, T.; Wang, S.Y.; Bloch, M.R. Rate and Detection Error-Exponent Tradeoffs of Joint Communication and Sensing. In Proceedings of the 2nd IEEE International Symposium on Joint Communications & Sensing (JC&S), Seefeld, Austria, 9–10 March 2022; pp. 1–6. [CrossRef]
21. Chang, M.C.; Wang, S.Y.; Erdoğan, T.; Bloch, M.R. Rate and Detection-Error Exponent Tradeoff for Joint Communication and Sensing of Fixed Channel States. *IEEE J. Sel. Areas Inf. Theory* **2023**, *4*, 245–259. [CrossRef]
22. Joudeh, H. Joint communication and target detection with multiple antennas. In Proceedings of the 26th International ITG Workshop on Smart Antennas and 13th Conference on Systems, Communications, and Coding, Braunschweig, Germany, 27 February–3 March 2023; pp. 1–6.
23. Ahmadipour, M.; Wigger, M.; Shamai, S. Strong Converses for Memoryless Bi-Static ISAC. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, 25–30 June 2023; pp. 1818–1823. [CrossRef]
24. Ahmadipour, M.; Wigger, M.; Shamai, S. Strong Converse for Bi-Static ISAC with Two Detection-Error Exponents. In Proceedings of the International Zurich Seminar on Information and Communication (IZS), Zurich, Switzerland, 6–8 March 2024; p. 45. [CrossRef]
25. Weinberger, N.; Merhav, N. Codeword or Noise? Exact Random Coding Exponents for Joint Detection and Decoding. *IEEE Trans. Inf. Theory* **2014**, *60*, 5077–5094. [CrossRef]
26. Weinberger, N.; Merhav, N. Channel Detection in Coded Communication. *IEEE Trans. Inf. Theory* **2017**, *63*, 6364–6392. [CrossRef]
27. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507. [CrossRef]
28. Hoeffding, W. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **1965**, 369–401. [CrossRef]
29. Shannon, C.E.; Gallager, R.G.; Berlekamp, E.R. Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels. I. *Inf. Control.* **1967**, *10*, 65–103. [CrossRef]
30. Csiszár, I.; Longo, G. On the error exponent for source coding and for testing simple statistical hypotheses. *Studia Sci. Math. Hungar.* **1971**, *6*, 181–191.
31. Blahut, R. Hypothesis Testing and Information Theory. *IEEE Trans. Inf. Theory* **1974**, *20*, 405–417. [CrossRef]
32. Hayashi, M. Discrimination of Two Channels by Adaptive Methods and Its Application to Quantum System. *IEEE Trans. Inf. Theory* **2009**, *55*, 3807–3820. [CrossRef]
33. Polyanskiy, Y.; Verdú, S. Binary hypothesis testing with feedback. In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 6–11 February 2011.
34. Nitinawarat, S.; Atia, G.K.; Veeravalli, V.V. Controlled Sensing for Multihypothesis Testing. *IEEE Trans. Autom. Control* **2013**, *58*, 2451–2464. [CrossRef]
35. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011. [CrossRef]
36. El Gamal, A.; Kim, Y.H. *Network Information Theory*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2011.
37. Shannon, C.E. Certain results in coding theory for noisy channels. *Inf. Control* **1957**, *1*, 6–25. [CrossRef]
38. Polyanskiy, Y.; Wu, Y. *Information Theory: From Coding to Learning*; Cambridge University Press: Cambridge, UK, 2024. [CrossRef]
39. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
40. Nikbakht, H.; Wigger, M.; Shamai, S.; Poor, H.V. Integrated Sensing and Communication in the Finite Blocklength Regime. *arXiv* **2024**, arXiv:2401.15752.
41. Chang, M.C.; Wang, S.Y.; Bloch, M.R. Sequential joint communication and sensing of fixed channel states. In Proceedings of the IEEE Information Theory Workshop (ITW), Saint-Malo, France, 23–28 April 2023; pp. 462–467.