Statistical Hypothesis Testing and Lower Bounds to the Error Probability

Gonzalo Vazquez-Vilar, Member, IEEE, Adrià Tauste Campo, Member, IEEE,
Albert Guillén i Fàbregas, Senior Member, IEEE, Alfonso Martinez, Senior Member, IEEE

Abstract

We prove two alternative expressions for the error probability of Bayesian $M$-ary hypothesis testing. The first expression is related to the error probability of binary hypothesis testing, and the second one to a generalization of the Verdú-Han lower bound. This result is used to characterize the error probability of the main problems in information theory and identify the steps where previous converse results are loose with respect to the actual probability of error.

I. INTRODUCTION

Statistical hypothesis-testing problems appear in areas as diverse as information theory, image processing, signal processing, social sciences or biology. Depending on the field, this problem can be referred to as hypothesis testing, classification, discrimination, signal detection or model selection. The goal of $M$-ary hypothesis testing is to decide among $M$ possible hypotheses based on the observation of a certain random variable. The Bayesian formulation of such problem assumes that there exists a prior distribution over the hypotheses. This allows to formulate the problem in terms of the minimum average error probability criterion, or its generalization, the Bayes risk. When the number of hypotheses is two, this problem reduces to binary hypothesis testing. While a Bayesian approach is still possible, the binary setting allows a simple formulation in terms of the two types of pairwise errors with no prior distribution over the hypotheses. The work of Neyman-Pearson [1] established the optimum binary test in this setting. Thanks to its simplicity and robustness, this has been the most popular approach in the literature.

G. Vazquez-Vilar was with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain. Now he is with the Signal Theory and Communications Department, Universidad Carlos III de Madrid, Leganés, Spain (email: gvazquez@ieee.org). A. Tauste Campo, A. Martinez are with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain (emails: adria.tauste@upf.edu, alfonso.martinez@ieee.org). A. Guillén i Fàbregas is with the Institució Catalana de Recerca i Estudis Avançats (ICREA), the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, and the Department of Engineering, University of Cambridge, CB2 1PZ Cambridge, United Kingdom (email: guillen@ieee.org).

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union’s 7th Framework Programme under grant agreements 303633 and 329837 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03.

This work was presented in part at the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7–12, 2013.
In the context of reliable communication, binary hypothesis testing has been instrumental in the derivation of converse bounds on the error probability. In [2, Sec. III] Shannon, Gallager and Berlekamp derived lower bounds to the error probability in the transmission of \( M \) codewords, including the well known sphere-packing bound [2], [3], by building on their analysis in choosing between two codewords, an instance of binary hypothesis testing. In [4], Forney applied binary hypothesis testing to determine the optimum decision regions in the problem of channel coding with erasures. In [5], Blahut emphasized the fundamental role of binary hypothesis testing in information theory and provided an alternative derivation of the sphere-packing exponent. Inspired by this result, Omura presented in [6] a simple general method for lower-bounding the error probability of channel-coding and source-coding problems. The aforementioned bounds turn out to be exponentially tight in a certain parameter range. More recently, the hypothesis-testing method has been used to obtain accurate finite-length lower bounds to several communication problems (see, e. g., [7] for channel coding or [8], [9] for joint source-channel coding).

The information-spectrum method was proposed in 1993 by Han and Verdú [10] to extend traditional information-theoretic results on memoryless systems to a wide class of sources and channels. The key idea behind the information-spectrum method is to characterize the error probability of a system as a limiting expression of the information spectrum (defined as the asymptotic behavior of a logarithmic likelihood ratio). In the context of reliable communication, the information-spectrum approach leads to both upper and lower bounds to the error probability [11], [12] (see also [13]).

In this work we further develop the connection between information-spectrum, hypothesis-testing and converse bounds in information theory by providing a number of alternative expressions to the error probability of Bayesian \( M \)-ary hypothesis testing. We show that this probability can be equivalently described by the error probability of either a single binary hypothesis test with certain parameters, or a bank of \( M \) binary hypothesis tests. We also provide an explicit alternative expression using information-spectrum measures [13] and illustrate the connection with existing information-spectrum bounds. This framework is then used to study the performance of different communication problems and to identify the weaknesses of several converse results in the literature.

In Section II of this paper we formalize the binary hypothesis-testing problem. In Section III we present the \( M \)-ary hypothesis-testing problem and propose a number of alternative expressions to the average error probability. The hypothesis-testing framework is particularized to several communication problems in Section IV. Proofs of several results are included in the appendices.

II. BINARY HYPOTHESIS TESTING

Consider the binary hypothesis-testing problem of deciding between the two alternatives \( \mathcal{H}_0 \) or \( \mathcal{H}_1 \) based on a realization of the random variable \( Y \) taking values in \( \mathcal{Y} \). The observation \( y \) is assumed to have been generated according to a distribution \( P \) under hypothesis \( \mathcal{H}_0 \) and according to \( Q \) under \( \mathcal{H}_1 \), i. e.,

\[
\mathcal{H}_0 : \ Y \sim P, \\
\mathcal{H}_1 : \ Y \sim Q. 
\]
A binary hypothesis test is defined by a (possibly random) transformation $Y \rightarrow \{H_0, H_1\}$, described in general by the conditional distribution $T \triangleq P_{H|Y}$, where $H$ denotes the random variable associated to the output of the test taking values in $\{H_0, H_1\}$.

In the Bayesian setting, with prior probabilities $\pi_i \triangleq P_{H}(H_i)$, $i = 0, 1$, the average error probability of a test $T$ is given by

$$\bar{\epsilon}(T) = \pi_0 \epsilon_0(P, T) + \pi_1 \epsilon_1(Q, T),$$

(3)

with the type-0 and type-1 errors defined, respectively, as

$$\epsilon_0(P, T) = \sum_y P(y) T(H_1|y),$$

(4)

$$\epsilon_1(Q, T) = \sum_y Q(y) T(H_0|y).$$

(5)

The type-$i$ error denotes the probability of making an error under the $i$-th hypothesis, that is, $\epsilon_i$ denotes the probability of choosing $H_j$, $j \neq i$, when the true hypothesis is $H_i$, $i = 0, 1$. Substituting (4)-(5) into (3), and optimizing over tests $T$, we find that the best average error probability achievable by any binary test is given by

$$\min_T \bar{\epsilon}(T) = \min_T \left\{ \sum_y \left( \pi_0 P(y) + (\pi_1 Q(y) - \pi_0 P(y)) T(H_0|y) \right) \right\}.$$  

(6)

Since the argument of the minimization in (6) is linear in $T \in [0, 1]$, the optimal test is such that

$$T_{NP}(H_0|y) = \begin{cases} 1, & \text{if } \frac{P(y)}{Q(y)} > \gamma, \\ p_0, & \text{if } \frac{P(y)}{Q(y)} = \gamma, \\ 0, & \text{otherwise}, \end{cases}$$

(7)

where $\gamma = \frac{\pi_1}{\pi_0}$ and the value of $p_0 \in [0, 1]$ is irrelevant since it does not affect the objective in (6).

In the non-Bayesian setting, the priors $\pi_i$, $i = 0, 1$, are unknown and the quantity $\bar{\epsilon}$ cannot be used as performance measure. Instead, one focuses on the optimal trade-off between $\epsilon_0(\cdot)$ and $\epsilon_1(\cdot)$. We define the smallest type-0 error among all tests $T$ with a type-1 error at most $\beta$ as

$$\alpha_\beta(P, Q) \triangleq \min_{T: \epsilon_1(Q, T) \leq \beta} \left\{ \epsilon_0(P, T) \right\}.$$  

(8)

It follows that the optimal tests in both the Bayesian and the non-Bayesian formulations coincide. By introducing the constraint in the minimization in (8) into the objective by means of a Lagrange multiplier $\lambda$, we obtain

$$\alpha_\beta(P, Q) = \max_{\lambda \geq 0} \min_T \left\{ \epsilon_0(P, T) + \lambda (\epsilon_1(Q, T) - \beta) \right\}$$

(9)

$$= \max_{\lambda \geq 0} \left\{ \min_T \left\{ \sum_y (P(y) + (\lambda Q(y) - P(y)) T(H_0|y) \right\} - \lambda \beta \right\}.$$  

(10)

For fixed $\lambda$, the minimization over $T$ in (10) is equivalent to that in (6) with $\frac{\pi_1}{\pi_0} = \lambda$. The optimal test is thus given by (7) with $\gamma = \lambda$. Moreover, the maximization in (10) forces the constraint $\epsilon_1(Q, T) \leq \beta$ to hold with
equality. As a result, a test optimizing (9) is \( T_{NP} \) in (7) with the threshold \( \gamma \) and the probability \( p_0 \) chosen such that \( \epsilon_1(Q, T_{NP}) = \beta \).

The above formulation corresponds to the Neyman-Pearson (NP) lemma \([1]\) and gives an explicit form of a test \( T \) achieving the optimum trade-off (8). Any test achieving (8) is said to be optimal in the Neyman-Pearson sense.

### III. \( M \)-ary Hypothesis Testing

Consider now the general \( M \)-ary hypothesis-testing problem. We define two random variables \( V \) and \( Y \), where \( V \) takes values in a finite set \( \mathcal{V} \) of cardinality \( |\mathcal{V}| = M \), and \( Y \) is arbitrary. The joint distribution of these two random variables is described by \( P_{VY} \). The problem of estimating \( V \) from an observation of \( Y \) is an \( M \)-ary hypothesis-testing problem. Since the joint distribution \( P_{VY} \) defines a prior distribution \( P_V \) over the alternatives, the problem is naturally cast within the Bayesian framework.

An \( M \)-ary hypothesis test is defined by a (possibly random) transformation \( Y \rightarrow V \) described by the conditional distribution \( P_{\hat{V}|Y} \). We denote the average error probability of a test \( P_{\hat{V}|Y} \) as

\[
\bar{\epsilon}(P_{\hat{V}|Y}) \triangleq \Pr\{\hat{V} \neq V\}
\]

Minimizing over all possible conditional distributions \( P_{\hat{V}|Y} \) gives the smallest average error probability, namely

\[
\bar{\epsilon} \triangleq \min_{P_{\hat{V}|Y}} \bar{\epsilon}(P_{\hat{V}|Y}). \tag{13}
\]

The minimum in (13) is achieved by the test choosing the hypothesis \( v \) with largest maximum a posteriori (MAP) metric given the observation \( y \). How ties are resolved does not affect the error probability. The MAP test that breaks ties randomly with equal probability is

\[
P_{\hat{V}|Y}^{MAP}(v|y) = \begin{cases} \frac{1}{|\mathcal{S}(y)|}, & \text{if } v \in \mathcal{S}(y), \\ 0, & \text{otherwise}, \end{cases} \tag{14}
\]

where the set \( \mathcal{S}(y) \) is defined as

\[
\mathcal{S}(y) \triangleq \left\{ v \mid P_{VY}(v, y) = \max_{v'} \{P_{VY}(v', y)\} \right\}. \tag{15}
\]

Substituting (14) in (12) gives the minimum possible error probability,

\[
\bar{\epsilon} = 1 - \sum_{v,y} P_{VY}(v, y) P_{\hat{V}|Y}^{MAP}(v|y)
= 1 - \sum_{y} \max_{v'} P_{VY}(v', y). \tag{17}
\]

The next theorem presents two alternative, equivalent expressions for the minimum error probability \( \bar{\epsilon} \).

---

\(^1\)While both binary and \( M \)-ary hypothesis tests are defined by conditional distributions, for clarity of exposition, we use different notation to refer to binary tests, denoted in this manuscript by \( T \), and \( M \)-ary tests, denoted by \( P_{\hat{V}|Y} \).
Theorem 1: The average error probability of an $M$-ary hypothesis-testing problem satisfies
\[
\bar{\epsilon} = \max_{Q_V} \alpha_{1/\mu} \left(\bar{P}_{V|Y}, Q_V^* \times Q_Y\right) = \max_{Q_V} \sup_{\gamma \geq 0} \left\{ \Pr \left[ \frac{P_{V|Y}(V,Y)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\}
\]
(18)
where $Q_V^*(v) \triangleq \frac{1}{M}$ for all $v$. Moreover, a maximizing distribution $Q_Y$ in both expressions is
\[
Q_Y^*(y) \triangleq \frac{1}{\mu} \max_{v'} P_{V|Y}(v',y)
\]
(20)
where $\mu \triangleq \sum_{V} \max_{v'} P_{V|Y}(v',y)$ is a normalizing constant.

This result shows that the error probability of Bayesian $M$-ary hypothesis testing can be expressed as the best type-0 error probability of a binary hypothesis test discriminating between the original distribution $P_{V|Y}$ and an alternative distribution $Q_V^* \times Q_Y^*$ with type-1-error constraint equal to $\frac{1}{\mu}$. In the proof of Theorem 1 this binary test is related to the MAP test of the original $M$-ary hypothesis testing problem. We conclude that the MAP criterion, that minimizes the average error probability of an $M$-ary hypothesis test, can be alternatively used to solve a binary hypothesis-testing problem upon appropriately defining the alternative distribution.

Theorem 1 also provides an alternative characterization based on information-spectrum measures. In particular, the probability term in (19) corresponds to the tail probability of the information density $\log h$ hypothesis-testing problem upon appropriately defining the alternative distribution. Accordingly, i.e., this test emulates the MAP test yielding the exact error probability.

Before proving Theorem 1, we briefly discuss several possible extensions and connections to relevant results in information theory.

a) Extension to countable alphabets: The function $\alpha_{\sigma}(\cdot)$ in (8) can be defined for arbitrary $\sigma$-finite measures, not necessarily probabilities. If we choose $\bar{Q}_V$ to be the counting measure, i.e., $\bar{Q}_V(v) = 1$ for all $v$, the type-1 error probability becomes $\epsilon_1(\bar{Q}_V \times Q_Y, T) = 1$ and (18) can be written as
\[
\bar{\epsilon} = \max_{Q_V} \alpha_1 \left(\bar{P}_{V|Y}, \bar{Q}_V \times Q_Y\right).
\]
(21)
This expression applies to $V$ being either finite or countably infinite. Furthermore, since (19) directly applies to finite or countably infinite source alphabets, so does the result from Theorem 1.

b) A single test versus a bank of $M$ tests: Eq. (18) relates the error probability $\bar{\epsilon}$ to the type-0 error probability of a binary test between distributions $P_{V|Y}$ and $Q_V^* \times Q_Y$ where the optimal distribution $Q_Y$ is given by (20). Instead of a single binary test, it is also possible to consider a bank of $M$ binary hypothesis tests between distributions $P_{Y|V=v}$ and $Q_Y$. In this case, we can also express the average error probability of $M$-ary hypothesis testing as
\[
\bar{\epsilon} = \max_{Q_Y} \left\{ \sum_{v} P_Y(v) \alpha_{Q_Y(v)}(P_{Y|V=v}, Q_Y) \right\}
\]
(22)
where $Q_Y(v) \triangleq \sum_{y} Q_Y(y) P_{V|Y}^\text{MAP}(v|y)$. See Appendix A for details.
c) Connection with the metaconverse: Fixing an arbitrary $Q_Y$ in (18) gives a lower bound to the minimum error probability, that is

$$\bar{\epsilon} \geq \alpha_{\frac{1}{M}} (P_{VY}, Q^*_V \times Q_Y).$$

(23)

In channel coding, the decoder has to decide between $M$ possible messages based on the channel output. Hence, the channel-coding problem with a fixed codebook is precisely an $M$-ary hypothesis-testing problem. In this setting, the bound (23) coincides with the metaconverse bound [7, Th. 26] for the choice $Q_{VY} = Q^*_V \times Q_Y$. Theorem 1 thus shows that the metaconverse is tight after optimization over the auxiliary distribution $Q_{VY}$.

d) Connection with Verdú-Han’s lower bound: Weakening the identity in (19) for an arbitrary $Q_Y$ we obtain

$$\bar{\epsilon} \geq \sup_{\gamma \geq 0} \left\{ \Pr \left[ \frac{P_{Y|V}(Y|V)P_V(V)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\}. \tag{24}$$

By choosing $Q_Y = P_Y$ in (24) we recover the Verdú-Han lower bound in the channel [11, Th. 4] and joint source-channel coding settings [14, Lem. 3.2]. The bound (24) with arbitrary $Q_Y$ was also known in the information theory community, and can be recovered from Hayashi-Nagaoka’s lemma [15, Lem. 4], an extension of the Verdú-Han lower bound [11, Th. 4] to classical-quantum channels.

e) Connection with Wolfowitz’s strong converse: If we consider the best source message in (19), i.e.,

$$\bar{\epsilon} = \sup_{Q_Y} \sup_{\gamma \geq 0} \left\{ \sum_v P_V(v) \Pr \left[ \frac{P_{Y|V}(Y|v)P_V(v)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\}, \tag{25}$$

we recover Wolfowitz’s channel coding strong converse [16].

f) Connection with Poor-Verdú’s lower bound: A tightening of the Verdú-Han lower bound is the Poor-Verdú bound [12], derived for the $M$-ary hypothesis-testing problem. The result in Theorem 1 also recovers this bound, as we show by means of the following lemma, which lower-bounds the best attainable performance of a binary test. Let us denote by $P[E]$ (resp. $Q[E]$) the probability of the event $E$ with respect to the underlying distribution $P$ (resp. $Q$).

**Lemma 1:** For a pair of distributions $\{P, Q\}$ and any $\gamma' \geq 0$, it holds

$$\alpha_{\beta}(P, Q) \geq P \left[ \frac{dP}{dQ} < \gamma' \right] - \gamma' \beta. \tag{27}$$

Moreover, provided that

$$0 \leq \beta \leq \frac{Q \left[ \frac{dP}{dQ} \geq \gamma' \right]}{P \left[ \frac{dP}{dQ} \geq \gamma' \right]}, \tag{28}$$

the following result holds,

$$\alpha_{\beta}(P, Q) \geq (1 - \gamma' \beta)P \left[ \frac{dP}{dQ} < \gamma' \right]. \tag{29}$$

**Proof:** See Appendix B.
Using (29) from Lemma 1 in (18) we obtain
\[ \bar{\varepsilon} \geq (1 - \gamma) \Pr \left[ \frac{P_{Y|V}(Y|V)P_Y(V)}{Q_Y(Y)} \leq \gamma \right], \]  
provided that \( Q_Y \) and \( \gamma \geq 0 \) satisfy
\[ \sum_{v,y} P_{V|Y}(v,y) \mathbb{I} \left\{ \frac{P_{V|Y}(v,y)}{Q_Y(y)} \geq \gamma \right\} \leq \sum_{v,y} Q_Y(y) \mathbb{I} \left\{ \frac{P_{V|Y}(v,y)}{Q_Y(y)} \geq \gamma \right\}. \]  
This condition is fulfilled for every \( \gamma \geq 0 \), when \( Q_Y = P_Y \) or \( Q_Y = Q^*_Y \). However, it is possible to find counterexamples in which (31) does not hold for some \( \gamma \) and \( Q_Y \) pairs. For \( Q_Y = P_Y \), and optimizing over \( \gamma \geq 0 \), (30) recovers the Poor-Verdú bound [12]. For \( Q_Y = Q^*_Y \), optimizing over \( \gamma \geq 0 \), (30) provides an alternative characterization of the actual probability of error.
\[ \bar{\varepsilon} = \max_{\gamma \geq 0} \left\{ (1 - \gamma) \Pr \left[ \frac{P_{Y|V}(Y|V)P_Y(V)}{Q^*_Y(Y)} \leq \gamma \right] \right\}. \]  

A. Proof of Theorem 1

We first prove the equality between the left- and right-hand-side of (18) by showing that the optimization problems (13) and (18) are equivalent. From (13) we have that
\[ \bar{\varepsilon} = \min_{P_{V|Y}: \sum_{v} P_{V|Y}(v|y) \leq 1, y \in \mathcal{Y}} \left\{ \sum_{v,y} P_{V|Y}(v,y) \left( 1 - P_{V|Y}(v|y) \right) \right\} \]  
\[ = \max_{\{\lambda_y \geq 0\}} \min_{P_{V|Y}} \left\{ \sum_{v,y} P_{V|Y}(v,y) \left( 1 - P_{V|Y}(v|y) \right) + \sum_y \lambda_y \left( \sum_v P_{V|Y}(v|y) - 1 \right) \right\}, \]  
where in (33) we wrote explicitly the (active) constraints resulting from \( P_{V|Y} \) being a conditional distribution; and (34) follows from introducing the constraints into the objective via the Lagrange multipliers \( \lambda_y, y \in \mathcal{Y} \). Similarly, we write (18) as
\[ \max_{Q_Y} \alpha_{\beta}(P_{V|Y}, Q^*_V \times Q_Y) \]  
\[ = \max_{Q_Y} \max_{T: \sum_{v,y} \frac{1}{\mathbb{P} Q_Y(y) T(H_0|v,y)} \leq \frac{1}{\mathbb{P}} \left\{ \sum_{v,y} P_{V|Y}(v,y) T(H_1|v,y) \right\} \]  
\[ = \max_{\lambda \geq 0} \min_{Q_Y} \left\{ \sum_{v,y} P_{V|Y}(v,y) \left( 1 - T(H_0|v,y) \right) + \lambda \left( \sum_{v,y} Q_Y(y) T(H_0|v,y) - 1 \right) \right\}, \]  
where in (35) we used the definitions of \( Q^*_V \) and \( \alpha_{\beta}(\cdot) \); (36) follows from introducing the constraint into the objective via the Lagrange multiplier \( \lambda \); and (37) follows by noting that \( \lambda \) and \( Q_Y \) only appear in the objective function of (36) as \( \lambda Q_Y(y), y \in \mathcal{Y} \), hence we may optimize (36) over \( \lambda_y \triangleq \lambda Q_Y(y) \) instead. As the constraints \( P_{V|Y} \in [0, 1] \) and \( T \in [0, 1] \) are inactive in (34) and (37), respectively, we have chosen not to write them explicitly.

Comparing (34) and (37), it is readily seen that the optimization problems (13) and (18) are equivalent.
In order to prove identity (19), we first use (27) from Lemma 1 into (18) to obtain
\[
\tilde{\epsilon} \geq \max_{Q_Y} \sup_{\gamma \geq 0} \left\{ \Pr \left[ \frac{P_{Y|V}(Y|V)P_V(V)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\},
\]  
(38)
By using the distribution \(Q_Y = Q'_Y\) in (20) and by choosing \(\gamma = \mu\), the probability term in (38) becomes
\[
\Pr \left[ \frac{P_{Y|V}(Y|V)P_V(V)}{Q'_Y(Y)} \leq \mu \right] = \Pr \left[ P_Y(Y)P_{V|Y}(V|Y) \leq P_Y(Y) \max_{v'} P_{V|Y}(v'|Y) \right] = 1.
\]  
(39)
Substituting \(Q_Y = Q'_Y, \gamma = \mu\), and using (39) in (38) we obtain
\[
\tilde{\epsilon} \geq \max_{Q_Y} \sup_{\gamma \geq 0} \left\{ \Pr \left[ \frac{P_{Y|V}(Y|V)P_V(V)}{Q_Y(Y)} \leq \gamma \right] - \gamma \right\}
\]
\[= 1 - \mu, \]
\[= 1 - \sum_{v} \max_{y} P_{V|Y}(v', y)
\]
\[= \bar{\epsilon}, \]
(42)
where in (42) we used the definition of \(\mu\) and (43) follows from (17). The identity (19) in the theorem is due to (40)-(42), where it is readily seen that \(Q_Y = Q'_Y\) is a maximizer of (19). Moreover, since \(Q'_Y\) is a maximizer of (19), and since Lemma 1 applies for a fixed \(Q_Y\), it follows that \(Q'_Y\) is also an optimal solution to (18).

Remark: The optimality of \(Q'_Y\) in (18) can also be proved using a (more tedious) constructive argument. To see this, let us define \(Q^*_Y \triangleq Q'_Y \times Q'_Y\) and the binary test \(T_{MAP}(H_0|v, y) \triangleq P_{MAP}(v|y).\) Using (14) and (15), this test can be written as
\[
T_{MAP}(H_0|v, y) \triangleq \begin{cases} \frac{1}{|S(y)|}, & \text{if } v \in S(y), \\ 0, & \text{otherwise}, \end{cases}
\]
(44)
where
\[
S(y) = \left\{ v \mid \frac{P_{V|Y}(v, y)}{Q^*_Y(v, y)} = \mu|V| \right\}.
\]
(45)
Using similar arguments as in [17, Th. 1] it can be shown that \(T_{MAP}\) achieves the exact same performance as a the Neyman-Pearson test \(T_{NP}\) in (7) with parameters \(\gamma = \mu|V|\) and
\[
\rho_0 = \frac{\sum_{v} \sum_{y} S(y) P_{V|Y}(v, y)}{\sum_{y} \sum_{v} S(y) P_{V|Y}(v, y)}.
\]
(46)
Then it holds that
\[
\alpha \frac{1}{\rho_0} (P_{V|Y}, Q^*_Y) = \epsilon_0(P_{V|Y}, T_{NP})
\]
\[= \epsilon_0(P_{V|Y}, T_{MAP})
\]
\[= \bar{\epsilon},
\]
(47)
(48)
(49)
where in the last step we used the definition of \(\epsilon_0(\cdot)\) and the identity \(T_{MAP}(H_0|v, y) = P_{V|Y}(v|y).\)

While both \(T_{MAP}\) and \(T_{NP}\) attain the Neyman-Pearson performance, in general they are not the same test, as they may differ in the set of points \(y\) where there exists a tie among different hypotheses according to the MAP criterion.
IV. Applications

We now apply and extend the framework presented in previous sections to study the error probability of different communication problems. We particularize our results to almost-lossless source coding, channel coding and lossy source coding. For a fixed codebook, the almost-lossless source-channel coding problem is an $M$-ary hypothesis-testing problem, and the results from Section III directly apply; for details we refer to the reader to [17].

A. Almost-lossless source coding

Consider fixed-length almost-lossless compression, where a discrete source with finite alphabet $V^n$ and source distribution $P_{V^n}$ is compressed into a set of $M$ indexes. A decoder generates the reconstructed message $\hat{v}$ according to a codebook $C = \{\hat{v}_1, \ldots, \hat{v}_M\}$ and the error probability is given by $\epsilon(C) = \Pr\{\hat{V} \neq V\}$.

Consider a binary hypothesis test deciding if an observation $v$ is generated according to the distribution $P_V$ ($H_0$) or according to a uniform distribution $Q_*V$ ($H_1$). For a given source code $C$, we let $T_{SC}$ denote the test $T_{SC}(H_0|v) \doteq 1\{\hat{v} = v\}$, (50) that decides $H_0$, if $\hat{v} = v$, or $Q_*V$, otherwise, for a given observation $v$.

Then, $\epsilon_0(P_V, T_{SC}) = \epsilon(C)$ and, as the number of distinct messages $v$ belonging to $C$ is at most $M$, $\epsilon_1(Q_*V, T_{SC}) \leq M|V|^{-n}$. The performance of this test is lower bounded by the performance of the Neyman-Pearson test. Hence,

$$\epsilon(C) \geq \alpha_{M|V|^{-n}}(P_V, Q_*V).$$

(51)

Since the lower bound (51) does not depend on an specific code, it is a general converse result on the error probability of source coding. In order to compute the right-hand-side of (51) we may resort to (7), which upon particularization yields

$$T_{NP}(H_0|v) = \begin{cases} 
1, & \text{if } P_V(v) > \gamma, \\
p_0, & \text{if } P_V(v) = \gamma, \\
0, & \text{otherwise},
\end{cases}$$

(52)

where $p_0, \gamma \in [0, 1]$ are parameters such that the type-1 error of this test is $\epsilon_1(Q_*V, T_{NP}) = M|V|^{-n}$.

The NP test (52) gives not only a converse result, but describes the structure of the best achievable scheme. As expected, any practical source code achieving (51) with equality needs to recover without error the messages $v$ with $P_V(v) > \gamma$ and a fraction $p_0$ of the messages such that $P_V(v) = \gamma$. That is, the best scheme encodes the most probable messages and discards the rest. Also, the condition $\epsilon_1(Q_*V, T_{NP}) = M|V|^{-n}$ indicates that the total number of messages encoded without error must be equal to $M$, i.e., the codebook size. The exact error probability of this construction equals the right-hand-side of (51) and, therefore, (51) holds with equality.
Figure 1. Channel coding error probability bounds for the BSC with parameters $P_{Y|X}(1|0) = 0.1$, $M = 4$.

B. Channel coding

Consider a length-$n$ channel coding scenario, where one of $M$ equiprobable messages is transmitted over a channel $P_{Y|X}$, with input $x \in \mathcal{X}^n$ and output $y \in \mathcal{Y}^n$. The encoder selects codeword $x$ corresponding to the message $v \in \{1, \ldots, M\}$ within a codebook $C = \{x_1, \ldots, x_M\}$. We let $P_X(C)$ denote the input distribution induced by the code $C$. The decoder uses the maximum likelihood (ML) rule to determine $\hat{v} \in \{1, \ldots, M\}$ and the error probability is given by $\epsilon(C) = \text{Pr}\{\hat{V} \neq V\}$.

Consider a binary hypothesis test deciding if an observation $(x, y)$ is generated according to the distribution $P_X(C) \times P_{Y|X}$ or according to $P_X(C) \times Q_Y$. Under this setting, Theorem 1 yields

$$\min_C \epsilon(C) = \min_C \max_{Q_Y} \left\{ \alpha \frac{1}{n} \left( P_X(C) \times P_{Y|X}, P_X(C) \times Q_Y \right) \right\} \geq \inf_{P_X} \sup_{Q_Y} \left\{ \alpha \frac{1}{n} \left( P_X \times P_{Y|X}, P_X \times Q_Y \right) \right\},$$

where in (54) we enlarged the set of distributions over which the minimization is carried out.

The bound in (54) coincides with the hypothesis-testing bound from [7, Thm. 27]. For memoryless channels, the exponential decay of (54) in the block length $n$ has recently been shown [18, Sec. VI.E] to be equal to the sphere-packing exponent [2]. However, below the critical rate of the channel, the reliability function is in general bounded away [3] from the sphere-packing exponent and thus, the gap between (53) and (54) may grow exponentially with $n$.

Fig. 1 shows different bounds on the error probability for the transmission of $M = 4$ messages and a binary symmetric channel (BSC). For very low rates, the best code can be obtained explicitly [19] and its ML decoding error probability can be explicitly computed. As upper bound, we show the exact random coding error probability
when the ties are decoded in error (RCE) for a random-coding ensemble with equiprobable codebook entries. As lower bounds, we depict (53) (computed for the best code) and (54). At this low rate, the reliability function of the error probability does not coincide neither with the random-coding nor with the sphere-packing error exponents, given respectively by the exponential decay of the bounds RCE and (54) in the figure. As a result, the weakening (54) yields a looser bound that incurs in a loss in exponent.

C. Lossy source coding

Consider a fixed-length lossy compression scenario, where the output of a general source with alphabet $\mathcal{V}^n$ and source distribution $P_V$ is mapped to a codeword $w \in \mathcal{W}^n$ according to a codebook $C = \{w_1, w_2, \ldots, w_M\}$. We define a distortion measure $d: \mathcal{V}^n \times \mathcal{W}^n \to [0, \infty)$ and, for a maximum allowed distortion $D$, an excess distortion error probability given by

$$\epsilon_d(C, D) \doteq \Pr\{d(V, W) > D\}. \quad (55)$$

As it happened in almost-lossless source coding, the best encoder simply maps the length-$k$ source message $v$ to the closest (in terms of distortion) codeword $w$ in the codebook $C = \{w_1, w_2, \ldots, w_M\}$. We denote this encoder by $P^{(C)}_{W|V}$. Then, given a codebook $C$, the minimum achievable distortion for the source message $v$ is

$$d(v, C) \doteq \min_{w \in C} d(v, w), \quad (56)$$

and, consequently, the excess distortion probability of $P^{(C)}_{W|V}$ is given by

$$\epsilon_d(C, D) = \sum_v P_V(v) \mathbb{1}\{d(v, C) > D\}. \quad (57)$$

Given the possible overlap between the covering regions, there is no straightforward equivalence between the excess distortion probability and an $M$-ary hypothesis-testing error probability. We may yet define an alternative binary hypothesis test as follows. Given an observation $v$, we choose $H_0$ if the encoder meets the maximum allowed distortion and $H_1$ otherwise, i.e. the test is defined as

$$T_{LSC}(H_0|v) = \mathbb{1}\{d(v, C) \leq D\}. \quad (58)$$

Particularizing (4) and (5) with this test, yields

$$\epsilon_0(P_V, T_{LSC}) = \sum_v P_V(v) \mathbb{1}\{d(v, C) > D\}, \quad (59)$$

and

$$\epsilon_1(Q_V, T_{LSC}) = \sum_v Q_V(v) \mathbb{1}\{d(v, C) \leq D\}. \quad (60)$$

As (57) and (59) coincide, $\epsilon_d(C, D)$ can be lower-bounded by the type-0 error of a Neyman-Pearson test, i.e.,

$$\epsilon_d(C, D) \geq \max_{Q_V} \left\{ \alpha_{\epsilon_1(Q_V, T_{LSC})}(P_V, Q_V) \right\}. \quad (61)$$

Moreover, the bound (61) holds with equality, as the next result shows.
Theorem 2: The average error probability of a given codebook \( C \) under MAP decoding satisfies

\[
\epsilon_d(C, D) = \max_{Q_V} \left\{ \alpha_{\epsilon_1(Q_V, T_{ISC})}(P_V, Q_V) \right\}
\]

\[
\geq \max_{Q_V} \left\{ \alpha_{M_{\sup w} \mathbb{P}[d(V, w) \leq D]}(P_V, Q_V) \right\},
\]

where \( Q[\mathcal{E}] \) denotes the probability of the event \( \mathcal{E} \) with respect to the underlying distribution \( Q_V \).

Proof: See Appendix C.

In contrast with the almost lossless setup, right-hand-side of (62) still depends on the code \( C \) through \( \epsilon_1(\cdot) \). This dependence disappears in the relaxation (63), that recovers the converse bound [20, Th. 8]. The weakness of (63) comes from relaxing the type-1 error in the bound to \( M \) times the type-1-error contribution of the worst codeword belonging to the reconstruction alphabet.

In the almost-lossless scenario, \( D = 0 \), and the errors corresponding to different codewords are non-overlapping. Moreover, when \( Q_V \) is assumed uniform we have that \( \mathbb{P}[d(V, w) \leq 0] = \mathbb{P}[V = w] = \frac{1}{|V|} \) independently of \( w \) and, as a result, (63) holds with equality.

ACKNOWLEDGEMENT

The authors would like to thank Sergio Verdú for several discussions. We would also thank Te Sun Han for pointing out the classical version of the Hayashi-Nagaoka’s lemma.

APPENDIX A

ONE TEST VERSUS MULTIPLE TESTS

In this appendix, we prove the equivalence between the optimization problems in (18) and (22).

First note that the argument of the maximization in (22) can be written as

\[
\sum_v P_V(v) \alpha_{\hat{Q}_V(v)}(P_{Y|V=v}, Q_Y)
\]

\[
= \sum_v P_V(v) \min_{T_v: \epsilon_1(Q_Y, T_v) \leq \hat{Q}_V(v)} \left\{ \epsilon_0(P_{Y|V=v}, T_v) \right\} \quad (64)
\]

\[
= \sum_v P_V(v) \max_{\lambda_0 \geq 0} \min_{T_v} \left\{ \sum_y P_{Y|V}(y|v) T_v(H_1|y) - \lambda_0 \left( \sum_{y'} Q_Y(y') T_v(H_0|y') - Q_V(v) \right) \right\} \quad (65)
\]

where (64) follows from the definition of \( \alpha(\cdot) \), and in (65) we used the definitions of the type-0 and type-1 errors and introduced the constraints into the objective by means of the Lagrange multipliers \( \lambda_v \).
Similarly, from (18) we have that
\[
\max_{Q_Y} \alpha_{\frac{1}{\alpha}} (P_{VY}, Q^*_V \times Q_Y) = \max_{Q_Y} \alpha_{\epsilon_1(Q_Y \times Q_Y, T_{\mathrm{MAP}})} (P_{VY}, Q_Y \times Q_Y)
\]
\[
= \max_{Q_Y} \max_{\lambda \geq 0} \max_{Q_V} \min_{v, y} \left\{ \sum_{v, y} P_{VY}(v, y) T(H_1|v, y) + \lambda \left( \sum_{v', y'} Q_V(v') Q_Y(y') \left( T(H_0|v', y') - P_{\mathrm{MAP}}(v'|y') \right) \right) \right\}
\]
\[
= \max_{Q_Y} \sum_{v} P_V(v) \max_{\lambda_0 \geq 0} \min_{v, y} \left\{ \sum_{y} P_{Y|V}(y|v) T(H_1|v, y) + \lambda_0 \left( \sum_{y'} Q_Y(y') T(H_0|v, y') - Q_Y(v) \right) \right\},
\]
where (66) follows as $Q^*_V$ is a maximizer of the RHS of (66); in (67) used the definition of $\alpha_{\epsilon_1}(\cdot, \cdot)$, and introduced the constraint into the objective by means of the Lagrange multiplier $\lambda$; and in (68) we rearranged terms and defined
\[
\lambda_0 \triangleq \frac{\lambda Q_V(v)}{P_V(v)}.
\]

The result follows from (65) and (68) by optimizing (65) over $Q_Y$ and identifying $T(H_i|v, y) \equiv T_v(H_i|y)$.

**APPENDIX B**

**PROOF OF LEMMA 1**

Consider a binary hypothesis test between distributions $P$ and $Q$. For clarity of exposition let us assume that, for a given type-1 error $\beta$, the term $p_0$ in (7) is equal to zero. In fact, the proof can be easily extended to consider an arbitrary $p_0$, but with heavier notation. Then, there exists $\gamma^*$ such that
\[
\beta = Q \left[ \frac{dP}{dQ} > \gamma^* \right],
\]
and the NP lemma yields
\[
\alpha_{\beta}(P, Q) = P \left[ \frac{dP}{dQ} \leq \gamma^* \right].
\]

For $0 \leq \gamma' < \gamma^*$, $P \left[ \frac{dP}{dQ} < \gamma' \right] \leq P \left[ \frac{dP}{dQ} < \gamma^* \right] = \alpha_{\beta}(P, Q)$ and Lemma 1 holds trivially.

For $\gamma' \geq \gamma^*$ it holds that
\[
\alpha_{\beta}(P, Q) = P \left[ \frac{dP}{dQ} < \gamma' \right] - \left[ \gamma^* < \frac{dP}{dQ} < \gamma' \right] \geq P \left[ \frac{dP}{dQ} < \gamma' \right] - \gamma Q \left[ \frac{dP}{dQ} < \gamma^* \right] \geq P \left[ \frac{dP}{dQ} < \gamma' \right] - \gamma' \left( Q \left[ \gamma^* < \frac{dP}{dQ} \right] - Q \left[ \frac{dP}{dQ} \geq \gamma' \right] \right),
\]
where (73) follows by noting that in the interval considered $dP < \gamma'dQ$. The first part of the lemma follows by lower bounding $Q \left[ \frac{dP}{dQ} \geq \gamma' \right] \geq 0$. For the second part we use the tighter lower bound
\[
Q \left[ \frac{dP}{dQ} \geq \gamma' \right] \geq \beta P \left[ \frac{dP}{dQ} \geq \gamma' \right],
\]

November 13, 2014
where $\beta$ by assumption (28) is such that
\[
\beta \leq \frac{Q \left[ \frac{dP}{dQ} \geq \gamma' \right]}{P \left[ \frac{dP}{dQ} \geq \gamma' \right]}.
\]

**APPENDIX C**

**PROOF OF THEOREM 2**

Let us choose $Q_V = Q_V^{(C)}$ where
\[
Q_V^{(C)}(v) = \frac{1}{\mu'} \mathbb{1} \{ d(v, C) > D \},
\]
with $\mu'$ a normalization constant. Then, we have that
\[
\epsilon_{II}(Q_V^{(C)}, T_{LSC}) = \sum_v Q_V^{(C)}(v) \mathbb{1} \{ d(v, C) \leq D \} = 0.
\]

Consider the NP test (7) with $\gamma = \mu'$, $p_0 = 1$. The NP test (7) particularizes to
\[
T_{NP}(H_0|v) = \begin{cases} 1, & \text{if } P_V(v) \geq \mathbb{1} \{ d(v, C) > D \}, \\ 0, & \text{otherwise}. \end{cases}
\]

Assuming that $P_V(v) < 1$ for all $v$, eq. (81) reduces to
\[
T_{NP}(H_0|v) = \mathbb{1} \{ d(v, C) \leq D \}
\]

That is, for $Q_V = Q_V^{(C)}$ the test $T_{LSC}$ is optimal in the Newman-Pearson sense. Then it holds that
\[
\max_{Q_V} \{ \alpha_{\epsilon_{II}(Q_V, T_{LSC})}(P_V, Q_V) \} \geq \alpha_{\epsilon_{II}(Q_V^{(C)}, T_{LSC})}(P_V, Q_V^{(C)}) = \alpha_{\epsilon_{d}(C, D)}.
\]

From the inequalities (61) and (84)-(86) the equality (62) follows.
The relaxation (63) comes from the fact that

$$\epsilon_{\mathcal{U}}(Q_V, T_{LSC}) = \sum_v Q_V(v) \mathbb{1}\{d(v, \mathcal{C}) \leq D\}$$

(87)

$$= \sum_v Q_V(v) \sum_w T_{LSC}(w|v) \mathbb{1}\{d(v, w) \leq D\}$$

(88)

$$\leq \sum_w \sum_v Q_V(v) \mathbb{1}\{d(v, w) \leq D\}$$

(89)

$$\leq M \sup_{w \in \mathcal{C}} \sum_v Q_V(v) \mathbb{1}\{d(v, w) \leq D\}$$

(90)

$$\leq M \sup_{w \in \mathcal{C}} \sum_v Q_V(v) \mathbb{1}\{d(v, w) \leq D\},$$

(91)

where in (89) we used that $T_{LSC}(w|v) = 0$ for $w \notin \mathcal{C}$ and that $T_{LSC}(w|v) \leq 1$ for $w \in \mathcal{C}$; (90) follows from considering the worst codeword in the sum, and (91) follows from relaxing the set over which the maximization is performed.

REFERENCES

[1] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Phil. Trans. R. Soc. Lond. A*, vol. 231, no. 694-706, p. 289, 1933.

[2] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels. I,” *Inf. Contr.*, vol. 10, no. 1, pp. 65–103, Jan. 1967.

[3] ——, “Lower bounds to error probability for coding on discrete memoryless channels. II,” *Inf. Contr.*, vol. 10, no. 5, pp. 522–552, May 1967.

[4] J. Forney, G.D., “Exponential error bounds for erasure, list, and decision feedback schemes,” *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, March 1968.

[5] R. E. Blahut, “Hypothesis testing and information theory,” *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 405–417, 1974.

[6] J. K. Omura, “A lower bounding method for channel and source coding probabilities,” *Inf. and Control*, vol. 27, no. 2, pp. 148–177, Feb. 1975.

[7] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[8] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, and A. Martinez, “Converse bounds for finite-length joint source-channel coding,” in *50th Annual Allerton Conference on Communication, Control and Computing*, Monticello, Illinois, USA, Oct. 2012.

[9] V. Kostina and S. Verdú, “Lossy joint source-channel coding in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.

[10] T. S. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[11] S. Verdú and T. S. Han, “A general formula for channel capacity,” *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[12] V. Poor and S. Verdú, “A lower bound on the probability of error in multihypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1992–1994, Nov. 1995.

[13] T. S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, 2003.

[14] ——, “Joint source-channel coding revisited: Information-spectrum approach,” *arxiv preprint arXiv:0712.2959v1*, 2007.

[15] M. Hayashi and H. Nagaoka, “General formulas for capacity of classical-quantum channels,” *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1753–1768, July 2003.

[16] J. Wolfowitz, “Notes on a general strong converse,” *Information and Control*, vol. 12, no. 1, pp. 1–4, Jan. 1968.

[17] G. Vazquez-Vilar, A. Tauste Campo, A. Guillén i Fábregas, and A. Martinez, “The meta-converse bound is tight,” in *2013 IEEE Int. Symp. on Inf. Theory*, Istanbul, Turkey, July 2013.
[18] Y. Polyanskiy, “Saddle point in the minimax converse for channel coding.” *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2576–2595, May 2013.

[19] P.-N. Chen, H.-Y. Lin, and S. Moser, “Optimal ultrasmall block-codes for binary discrete memoryless channels,” *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7346–7378, Nov. 2013.

[20] V. Kostina and S. Verdú, “Fixed-length lossy compression in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.