Arimoto-Rényi Conditional Entropy and Bayesian M-ary Hypothesis Testing

Igal Sason   Sergio Verdú

Abstract

This paper gives upper and lower bounds on the minimum error probability of Bayesian M-ary hypothesis testing in terms of the Arimoto-Rényi conditional entropy of an arbitrary order α. The improved tightness of these bounds over their specialized versions with the Shannon conditional entropy ($\alpha = 1$) is demonstrated. In particular, in the case where M is finite, we show how to generalize Fano's inequality under both the conventional and list-decision settings. As a counterpart to the generalized Fano's inequality, allowing M to be infinite, a lower bound on the Arimoto-Rényi conditional entropy is derived as a function of the minimum error probability. Explicit upper and lower bounds on the minimum error probability are obtained as a function of the Arimoto-Rényi conditional entropy for both positive and negative α. Furthermore, we give upper bounds on the minimum error probability as functions of the Rényi divergence. In the setup of discrete memoryless channels, we analyze the exponentially vanishing decay of the Arimoto-Rényi conditional entropy of the transmitted codeword given the channel output when averaged over a random-coding ensemble.

Keywords

Arimoto-Rényi conditional entropy, Bayesian minimum probability of error, Chernoff information, Fano's inequality, list decoding, M-ary hypothesis testing, random coding, Rényi entropy, Rényi divergence.

I. Introduction

In Bayesian M-ary hypothesis testing, we have:

- M possible explanations, hypotheses or models for the $\mathcal{Y}$-valued data $\{P_{Y|X=m}, m \in \mathcal{X}\}$ where the set of model indices satisfies $|\mathcal{X}| = M$; and
- a prior distribution P_X on $\mathcal{X}$.

The minimum probability of error of X given Y, denoted by $\varepsilon_{X|Y}$, is achieved by the *maximum-a-posteriori* (MAP) decision rule. Summarized in Sections I-A and I-C, a number of bounds on $\varepsilon_{X|Y}$ involving Shannon information measures or their generalized Rényi measures have been obtained in the literature. As those works attest, there is considerable motivation for the study of the relationships between error probability and information measures. The minimum error probability of Bayesian M-ary hypothesis testing is rarely directly computable, and the best lower and upper bounds are information theoretic. Furthermore, their interplay is crucial in the proof of coding theorems.

I. Sason is with the Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sason@ee.technion.ac.il).

S. Verdú was with the Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA.

Communicated by S. Watanabe, Associate Editor for Shannon Theory.

This manuscript appears in the *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, January 2018. It has been submitted to the journal in September 27, 2016, revised in May 24, 2017, and accepted for publication in September 4, 2017.

This work has been presented in part at the *2017 IEEE International Symposium on Information Theory*, Aachen, Germany, June 25–30, 2017.

This work has been supported by the Israeli Science Foundation (ISF) under Grant 12/12, by ARO-MURI contract number W911NF-15-1-0479 and in part by the Center for Science of Information, an NSF Science and Technology Center under Grant CCF-0939370.

A. Existing bounds involving Shannon information measures

1) Fano's inequality [24] gives an upper bound on the conditional entropy $H(X|Y)$ as a function of $\varepsilon_{X|Y}$ when M is finite.
2) Shannon's inequality [71] (see also [83]) gives an explicit lower bound on $\varepsilon_{X|Y}$ as a function of $H(X|Y)$, also when M is finite.
3) Tightening another bound by Shannon [70], Poor and Verdú [57] gave a lower bound on $\varepsilon_{X|Y}$ as a function of the distribution of the conditional information (whose expected value is $H(X|Y)$). This bound was generalized by Chen and Alajaji [14].
4) Baladová [6], Chu and Chueh [15, (12)], and Hellman and Raviv [38, (41)] showed that

$$\varepsilon_{X|Y} \leq \tfrac{1}{2}\, H(X|Y) \text{ bits} \tag{1}$$

when M is finite. It is also easy to show that (see, e.g., [25, (21)])

$$\varepsilon_{X|Y} \leq 1 - \exp\bigl(-H(X|Y)\bigr). \tag{2}$$

Tighter and generalized upper bounds on $\varepsilon_{X|Y}$ were obtained by Kovalevsky [48], Tebbe and Dwyer [76], and Ho and Verdú [39, (109)].
5) Based on the fundamental tradeoff of a certain auxiliary binary hypothesis test, Polyanskiy *et al.* [55] gave the *meta-converse* implicit lower bound on $\varepsilon_{X|Y}$, which for some choice of auxiliary quantities is shown to be tight in [82].
6) Building up on [38], Kanaya and Han [45] showed that in the case of independent identically distributed (i.i.d.) observations, $\varepsilon_{X|Y^n}$ and $H(X|Y^n)$ vanish exponentially at the same speed, which is governed by the Chernoff information between the closest hypothesis pair. In turn, Leang and Johnson [51] showed that the same exponential decay holds for those cost functions that have zero cost for correct decisions.
7) Birgé [9] gave an implicit lower bound on the minimax error probability ($\varepsilon_{X|Y}$ maximized over all possible priors) as a function of the pairwise relative entropies among the various models.
8) Generalizing Fano's inequality, Han and Verdú [34] gave lower bounds on the mutual information $I(X;Y)$ as a function of $\varepsilon_{X|Y}$.
9) Grigoryan *et al.* [33] and Sason [67] obtained error bounds, in terms of relative entropies, for hypothesis testing with a rejection option. Such bounds recently proved useful in the context of an almost-fixed size hypothesis decision algorithm, which bridges the gap in performance between fixed-sample and sequential hypothesis testing [49].

B. Rényi's information measures

In this paper, we give upper and lower bounds on $\varepsilon_{X|Y}$ not in terms of $H(X|Y)$ but in terms of the *Arimoto-Rényi* conditional entropy $H_\alpha(X|Y)$ of an arbitrary order α. Loosening the axioms given by Shannon [69] as a further buttress for $H(X)$, led Rényi [62] to the introduction of the *Rényi entropy* of order $\alpha \in [0, \infty]$, $H_\alpha(X)$, as well as the *Rényi divergence* $D_\alpha(P\|Q)$. Rényi's entropy and divergence coincide with Shannon's and Kullback-Leibler's measures, respectively, for $\alpha = 1$. Among other applications, $H_\alpha(X)$ serves to analyze the fundamental limits of lossless data compression [13], [16], [19]. Unlike Shannon's entropy, $H_\alpha(X)$ with $\alpha \neq 1$ suffers from the disadvantage that the inequality $H_\alpha(X_1, X_2) \leq H_\alpha(X_1) + H_\alpha(X_2)$ does not hold in general. Moreover, if we mimic the definition of $H(X|Y)$ and define conditional Rényi entropy as $\sum_{y\in\mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y)$, we find the unpleasant property that the conditional version may be larger than $H_\alpha(X)$. To remedy this situation, Arimoto [5] introduced a notion of conditional Rényi entropy, which in this paper we denote as $H_\alpha(X|Y)$, and which is indeed upper bounded by $H_\alpha(X)$. The corresponding α-*capacity* $\max_X \{H_\alpha(X) - H_\alpha(X|Y)\}$ yields a particularly convenient expression for the converse bound found by Arimoto [4] (see also [56]). The Arimoto-Rényi conditional entropy has also found applications in guessing and secrecy problems with side information ([2], [10], [35], [36] and [75]), sequential decoding [2], task encoding with side information available to both the task-describer (encoder) and the task-performer [11], and the list-size capacity of discrete memoryless channels in the presence of noiseless feedback [12]. In [65], the joint range of $H(X|Y)$ and $H_\alpha(X|Y)$ is obtained when the random variable X is restricted to take its values on a finite set with a given cardinality. Although outside the scope of this paper, Csiszár [19] and Sibson [74] proposed implicitly other definitions of conditional Rényi entropy, which lead to the same value of the α-capacity and have found various applications ([19] and [85]). The quantum generalizations of Rényi entropies and their conditional versions ([35], [36], [37], [43], [47], [54], [61], [77]) have recently served to prove strong converse theorems in quantum information theory ([36], [52] and [86]).

C. Existing bounds involving Rényi's information measures and related measures

In continuation to the list of existing bounds involving the conditional entropy and relative entropy in Section I-A, bounds involving Rényi's information measures and related measures include:

10) For Bayesian binary hypothesis testing, Hellman and Raviv [38] gave an upper bound on $\varepsilon_{X|Y}$ as a function of the prior probabilities and the Rényi divergence of order $\alpha \in [0, 1]$ between the two models. The special case of $\alpha = \frac{1}{2}$ yields the *Bhattacharyya bound* [44].
11) In [21] and [81], Devijver and Vajda derived upper and lower bounds on $\varepsilon_{X|Y}$ as a function of the quadratic Arimoto-Rényi conditional entropy $H_2(X|Y)$.
12) In [17], Cover and Hart showed that if M is finite, then
$$\varepsilon_{X|Y} \le H_{\mathrm{o}}(X|Y) \le \varepsilon_{X|Y} \left(2 - \tfrac{M}{M-1}\, \varepsilon_{X|Y}\right) \tag{3}$$
where
$$H_{\mathrm{o}}(X|Y) = \mathbb{E}\left[\sum_{x \in \mathcal{X}} P_{X|Y}(x|Y)\left(1 - P_{X|Y}(x|Y)\right)\right] \tag{4}$$
is referred to as conditional quadratic entropy. The bounds in (3) coincide if X is equiprobable on $\mathcal{X}$, and X and Y are independent.
13) One of the bounds by Han and Verdú [34] in Item 8) was generalized by Polyanskiy and Verdú [56] to give a lower bound on the α-mutual information ([74], [85]).
14) In [73], Shayevitz gave a lower bound, in terms of the Rényi divergence, on the maximal worst-case miss-detection exponent for a binary composite hypothesis testing problem when the false-alarm probability decays to zero with the number of i.i.d. observations.
15) Tomamichel and Hayashi studied optimal exponents of binary composite hypothesis testing, expressed in terms of Rényi's information measures ([37], [78]). A measure of dependence was studied in [78] (see also Lapidoth and Pfister [50]) along with its role in composite hypothesis testing.
16) Fano's inequality was generalized by Toussiant in [80, Theorem 2] (see also [22] and [43, Theorem 6]) not with the Arimoto-Rényi conditional entropy but with the average of the Rényi entropies of the conditional distributions (which, as we mentioned in Section I-B, may exceed the unconditional Rényi entropy).
17) A generalization of the minimax lower bound by Birgé [9] in Item 7) to lower bounds involving f-divergences has been studied by Guntuboyina [32].

D. Main results and paper organization

It is natural to consider generalizing the relationships between $\varepsilon_{X|Y}$ and $H(X|Y)$, itemized in Section I-A, to the Arimoto-Rényi conditional entropy $H_\alpha(X|Y)$. In this paper we find pleasing counterparts to the bounds in Items 1), 4), 6), 10), 11), and 13), resulting in generally tighter bounds. In addition, we enlarge the scope of the problem to consider not only $\varepsilon_{X|Y}$ but the probability that a list decision rule (which is allowed to output a set of L hypotheses) does not include the true one. Previous work on extending Fano's inequality to the setup of list decision rules includes [1, Section 5], [46, Lemma 1], [59, Appendix 3.E] and [84, Chapter 5].

Section II introduces the basic notation and definitions of Rényi information measures. Section III finds bounds for the guessing problem in which there are no observations; those bounds, which prove to be instrumental in the sequel, are tight in the sense that they are attained with equality for certain random variables. Section IV contains the main results in the paper on the interplay between $\varepsilon_{X|Y}$ and $H_\alpha(X|Y)$, giving counterparts to a number of those existing results mentioned in Sections I-A and I-C. In particular:

18) An upper bound on $H_\alpha(X|Y)$ as a function of $\varepsilon_{X|Y}$ is derived for positive α (Theorem 3); it provides an implicit lower bound on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$;
19) Explicit lower bounds on $\varepsilon_{X|Y}$ are given as a function of $H_\alpha(X|Y)$ for both positive and negative α (Theorems 5, 6, and 7);
20) The lower bounds in Items 18) and 19) are generalized to the list-decoding setting (Theorems 8, 9, and 10);
21) As a counterpart to the generalized Fano's inequality, we derive a lower bound on $H_\alpha(X|Y)$ as a function of $\varepsilon_{X|Y}$ capitalizing on the Schur concavity of Rényi entropy (Theorem 11);
22) Explicit upper bounds on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$ are obtained (Theorem 12);

Section V gives explicit upper bounds on $\varepsilon_{X|Y}$ as a function of the error probability of associated binary tests (Theorem 14) and of the Rényi divergence (Theorems 13 and 15) and Chernoff information (Theorem 15).

Section VI analyzes the exponentially vanishing decay of the Arimoto-Rényi conditional entropy of the transmitted codeword given the channel output when averaged over a code ensemble (Theorems 17 and 18). Concluding remarks are given in Section VII.

II. Rényi Information Measures: Definitions and Basic Properties

Definition 1: [62] Let P_X be a probability distribution on a discrete set $\mathcal{X}$. The *Rényi entropy of order* $\alpha \in (0,1) \cup (1,\infty)$ of X, denoted by $H_\alpha(X)$ or $H_\alpha(P_X)$, is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X^\alpha(x) \tag{5}$$

$$= \frac{\alpha}{1-\alpha} \log \|P_X\|_\alpha \tag{6}$$

where $\|P_X\|_\alpha = \left(\sum_{x\in\mathcal{X}} P_X^\alpha(x)\right)^{\frac{1}{\alpha}}$. By its continuous extension,

$$H_0(X) = \log \left|\operatorname{supp} P_X\right|, \tag{7}$$

$$H_1(X) = H(X), \tag{8}$$

$$H_\infty(X) = \log \frac{1}{p_{\max}} \tag{9}$$

where $\operatorname{supp} P_X = \{x \in \mathcal{X} \colon P_X(x) > 0\}$ is the support of P_X, and $p_{\max}$ is the largest of the masses of X.

When guessing the value of X in the absence of an observation of any related quantity, the maximum probability of success is equal to $p_{\max}$; it is achieved by guessing that $X = x_0$ where x_0 is a *mode* of P_X, i.e., $P_X(x_0) = p_{\max}$. Let

$$\varepsilon_X = 1 - p_{\max} \tag{10}$$

denote the minimum probability of error in guessing the value of X. From (9) and (10), the minimal error probability in guessing X without side information is given by

$$\varepsilon_X = 1 - \exp\bigl(-H_\infty(X)\bigr). \tag{11}$$

Definition 2: For $\alpha \in (0,1)\cup(1,\infty)$, the *binary Rényi entropy of order* α is the function $h_\alpha \colon [0,1] \to [0, \log 2]$ that is defined, for $p \in [0,1]$, as

$$h_\alpha(p) = H_\alpha(X_p) = \frac{1}{1-\alpha}\, \log\bigl(p^\alpha + (1-p)^\alpha\bigr), \tag{12}$$

where X_p takes two possible values with probabilities p and $1-p$. The continuous extension of the binary Rényi entropy at $\alpha = 1$ yields the binary entropy function:

$$h(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \tag{13}$$

for $p \in (0,1)$, and $h(0) = h(1) = 0$.

In order to put forth generalizations of Fano's inequality and bounds on the error probability, we consider Arimoto's proposal for the conditional Rényi entropy (named, for short, the Arimoto-Rényi conditional entropy).

Definition 3: [5] Let P_{XY} be defined on $\mathcal{X} \times \mathcal{Y}$, where X is a discrete random variable. The *Arimoto-Rényi conditional entropy of order* $\alpha \in [0,\infty]$ of X given Y is defined as follows:

- If $\alpha \in (0,1) \cup (1,\infty)$, then

$$H_\alpha(X|Y) = \frac{\alpha}{1-\alpha}\, \log\, \mathbb{E}\left[\left(\sum_{x\in\mathcal{X}} P^{\alpha}_{X|Y}(x|Y)\right)^{\frac{1}{\alpha}}\right] \tag{14}$$

$$= \frac{\alpha}{1-\alpha}\, \log \mathbb{E}\left[\|P_{X|Y}(\cdot|Y)\|_\alpha\right] \tag{15}$$

$$= \frac{\alpha}{1-\alpha}\, \log \sum_{y\in\mathcal{Y}} P_Y(y)\, \exp\left(\frac{1-\alpha}{\alpha}\, H_\alpha(X|Y=y)\right), \tag{16}$$

  where (16) applies if Y is a discrete random variable.

- By its continuous extension, the Arimoto-Rényi conditional entropy of orders 0, 1, and ∞ are defined as

$$H_0(X|Y) = \operatorname{ess\,sup} H_0\left(P_{X|Y}(\cdot|Y)\right) \tag{17}$$

$$= \log \max_{y\in\mathcal{Y}} \bigl|\operatorname{supp} P_{X|Y}(\cdot|y)\bigr| \tag{18}$$

$$= \max_{y\in\mathcal{Y}}\, H_0(X \,|\, Y=y), \tag{19}$$

$$H_1(X|Y) = H(X|Y), \tag{20}$$

$$H_\infty(X|Y) = \log \frac{1}{\mathbb{E}\Bigl[\max_{x\in\mathcal{X}} P_{X|Y}(x|Y)\Bigr]} \tag{21}$$

  where $\operatorname{ess\,sup}$ in (17) denotes the essential supremum, and (18) and (19) apply if Y is a discrete random variable.

Although not nearly as important, sometimes in the context of finitely valued random variables, it is useful to consider the unconditional and conditional Rényi entropies of negative orders $\alpha \in (-\infty, 0)$ in (5) and (14) respectively. By continuous extension

$$H_{-\infty}(X) = \log \frac{1}{p_{\min}}, \tag{22}$$

$$H_{-\infty}(X|Y) = \log \mathbb{E}^{-1}\left[\min_{x \in \mathcal{X}\colon P_{X|Y}(x|Y)>0} P_{X|Y}(x|Y)\right] \tag{23}$$

where $p_{\min}$ in (22) is the smallest of the nonzero masses of X, and $\mathbb{E}^{-1}[Z]$ in (23) denotes the reciprocal of the expected value of Z. In particular, note that if $|\mathcal{X}| = 2$, then

$$H_{-\infty}(X|Y) = \log \frac{1}{1 - \exp\left(-H_{\infty}(X|Y)\right)}. \tag{24}$$

Basic properties of $H_\alpha(X|Y)$ appear in [27], and in [43], [77] in the context of quantum information theory. Next, we provide two additional properties, which are used in the sequel.

Proposition 1: $H_\alpha(X|Y)$ is monotonically decreasing in α throughout the real line.

The proof of Proposition 1 is given in Appendix A. In the special case of $\alpha \in [1, \infty]$ and discrete Y, a proof can be found in [8, Proposition 4.6].

We will also have opportunity to use the following monotonicity result, which holds as a direct consequence of the monotonicity of the norm of positive orders.

Proposition 2: $\frac{\alpha-1}{\alpha}\, H_\alpha(X|Y)$ is monotonically increasing in α on $(0, \infty)$ and $(-\infty, 0)$.

Remark 1: Unless X is a deterministic function of Y, it follows from (17) that $H_0(X|Y) > 0$; consequently,

$$\lim_{\alpha \uparrow 0} \frac{\alpha-1}{\alpha}\, H_\alpha(X|Y) = +\infty, \tag{25}$$

$$\lim_{\alpha \downarrow 0} \frac{\alpha-1}{\alpha}\, H_\alpha(X|Y) = -\infty. \tag{26}$$

It follows from Proposition 1 that if $\beta > 0$, then

$$H_\beta(X|Y) \leq H_0(X|Y) \leq H_{-\beta}(X|Y), \tag{27}$$

and, from Proposition 2,

$$\frac{\alpha-1}{\alpha}\, H_\alpha(X|Y) \leq \frac{\beta-1}{\beta}\, H_\beta(X|Y) \tag{28}$$

if $0 < \alpha < \beta$ or $\alpha < \beta < 0$.

The third Rényi information measure used in this paper is the Rényi divergence. Properties of the Rényi divergence are studied in [23], [68, Section 8] and [73].

Definition 4: [62] Let P and Q be probability measures on $\mathcal{X}$ dominated by R, and let their densities be respectively denoted by $p = \frac{\mathrm{d}P}{\mathrm{d}R}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}R}$. The *Rényi divergence of order* $\alpha \in [0, \infty]$ is defined as follows:

- If $\alpha \in (0,1) \cup (1, \infty)$, then

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\, \log \mathbb{E}\left[p^\alpha(Z)\, q^{1-\alpha}(Z)\right] \tag{29}$$

$$= \frac{1}{\alpha-1}\, \log \sum_{x \in \mathcal{X}} P^\alpha(x)\, Q^{1-\alpha}(x) \tag{30}$$

  where $Z \sim R$ in (29), and (30) holds if $\mathcal{X}$ is a discrete set.

- By the continuous extension of $D_\alpha(P\|Q)$,

$$D_0(P\|Q) = \max_{\mathcal{A}:P(\mathcal{A})=1} \log \frac{1}{Q(\mathcal{A})}, \tag{31}$$

$$D_1(P\|Q) = D(P\|Q), \tag{32}$$

$$D_\infty(P\|Q) = \log \operatorname{ess\,sup} \frac{p(Z)}{q(Z)} \tag{33}$$

with $Z \sim R$.

Definition 5: For all $\alpha \in (0,1) \cup (1,\infty)$, the *binary Rényi divergence of order* α, denoted by $d_\alpha(p\|q)$ for $(p,q) \in [0,1]^2$, is defined as $D_\alpha([p\ 1-p]\|[q\ 1-q])$. It is the continuous extension to $[0,1]^2$ of

$$d_\alpha(p\|q) = \frac{1}{\alpha-1}\ \log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Big). \tag{34}$$

Used several times in this paper, it is easy to verify the following identity satisfied by the binary Rényi divergence. If $t \in [0,1]$ and $s \in [0,\theta]$ for $\theta > 0$, then

$$\log\theta - d_\alpha\big(t\|\tfrac{s}{\theta}\big) = \frac{1}{1-\alpha}\ \log\left(t^\alpha\, s^{1-\alpha} + (1-t)^\alpha\,(\theta-s)^{1-\alpha}\right). \tag{35}$$

By analytic continuation in α, for $(p,q) \in (0,1)^2$,

$$d_0(p\|q) = 0, \tag{36}$$

$$d_1(p\|q) = d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \tag{37}$$

$$d_\infty(p\|q) = \log\max\left\{\frac{p}{q}, \frac{1-p}{1-q}\right\} \tag{38}$$

where $d(\cdot\|\cdot)$ denotes the binary relative entropy. Note that if $t \in [0,1]$ and $M > 1$, then

$$\log M - d(t\,\|1-\tfrac{1}{M}) = t\,\log(M-1) + h(t). \tag{39}$$

A simple nexus between the Rényi entropy and the Rényi divergence is

$$D_\alpha(X\|U) = \log M - H_\alpha(X) \tag{40}$$

when X takes values on a set of M elements on which U is equiprobable.

Definition 6: The *Chernoff information* of a pair of probability measures defined on the same measurable space is equal to

$$C(P\|Q) = \sup_{\alpha\in(0,1)} (1-\alpha) D_\alpha(P\|Q). \tag{41}$$

Definition 7: ([74], [85]) Given the probability distributions P_X and $P_{Y|X}$, the α*-mutual information* is defined for $\alpha > 0$ as

$$I_\alpha(X;Y) = \min_{Q_Y} D_\alpha(P_{XY}\|P_X \times Q_Y). \tag{42}$$

As for the conventional mutual information, which refers to $\alpha = 1$, sometimes it is useful to employ the alternative notation

$$I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y). \tag{43}$$

In the discrete case, for $\alpha \in (0,1) \cup (1,\infty)$,

$$I_\alpha(X;Y) = \frac{\alpha}{1-\alpha}\, E_0\left(\frac{1}{\alpha} - 1, P_X\right), \tag{44}$$

where

$$E_0(\rho, P_X) = -\log \sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} P_X(x)\, P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho} \tag{45}$$

denotes Gallager's error exponent function [29]. Note that, in general, $I_\alpha(X;Y)$ does not correspond to the difference $H_\alpha(X) - H_\alpha(X|Y)$, which is equal to

$$H_\alpha(X) - H_\alpha(X|Y) = \frac{\alpha}{1-\alpha}\, E_0\left(\frac{1}{\alpha} - 1, P_{X_\alpha}\right) \tag{46}$$

where P_{X_α} is the scaled version of P_X, namely the normalized version of P_X^α. Since the equiprobable distribution is equal to its scaled version for any $\alpha \geq 0$, if X is equiprobable on a set of M elements, then

$$I_\alpha(X;Y) = \log M - H_\alpha(X|Y). \tag{47}$$

III. Upper and Lower Bounds on $H_\alpha(X)$

In this section, we obtain upper and lower bounds on the unconditional Rényi entropy of order α which are tight in the sense that they are attained with equality for certain random variables.

A. Upper bounds

In this subsection we limit ourselves to finite alphabets.

Theorem 1: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and let X be a random variable taking values on $\mathcal{X}$. Let $f\colon \mathcal{X} \to \mathcal{Y}$ be a deterministic function, and denote the cardinality of the inverse image by $L_y = |f^{-1}(y)|$ for every $y \in \mathcal{Y}$. Then, for every $\alpha \in [0,1) \cup (1,\infty)$, the Rényi entropy of X satisfies

$$H_\alpha(X) \leq \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} L_y^{1-\alpha}\, \mathbb{P}^\alpha[f(X) = y] \tag{48}$$

which holds with equality if and only if either $\alpha = 0$ or $\alpha \in (0,1) \cup (1,\infty)$ and P_X is equiprobable on $f^{-1}(y)$ for every $y \in \mathcal{Y}$ such that $P_X(f^{-1}(y)) > 0$.

Proof: It can be verified from (7) that (48) holds with equality for $\alpha = 0$ (by assumption $\operatorname{supp} P_X = \mathcal{X}$, and $\sum_{y\in\mathcal{Y}} L_y = |\mathcal{X}|$).

Suppose $\alpha \in (0,1) \cup (1,\infty)$. Let $|\mathcal{X}| = M < \infty$, and let U be equiprobable on $\mathcal{X}$. Let $V = f(X)$ and $W = f(U)$, so $(V, W) \in \mathcal{Y}^2$ and

$$P_V(y) = \mathbb{P}\big[f(X) = y\big], \tag{49}$$

$$P_W(y) = \frac{L_y}{M} \tag{50}$$

for all $y \in \mathcal{Y}$. To show (48), note that

$$H_\alpha(X) = \log M - D_\alpha(X\|U) \tag{51}$$
$$\leq \log M - D_\alpha(V\|W) \tag{52}$$
$$= \log M - \frac{1}{\alpha-1} \log \sum_{y\in\mathcal{Y}} \mathbb{P}^\alpha\big[f(X)=y\big] \left(\frac{L_y}{M}\right)^{1-\alpha} \tag{53}$$
$$= \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} L_y^{1-\alpha}\, \mathbb{P}^\alpha[f(X)=y] \tag{54}$$

where (51) is (40); (52) holds due to the data processing inequality for the Rényi divergence (see [23, Theorem 9 and Example 2]); and (53) follows from (30), (49) and (50).

From (51)–(54), the upper bound in (48) is attained with equality if and only if the data processing inequality (52) holds with equality. For all $\alpha \in (0,1)\cup(1,\infty)$,

$$D_\alpha(X\|U) = \frac{1}{\alpha-1} \log \sum_{x\in\mathcal{X}} P_X^\alpha(x) M^{\alpha-1} \tag{55}$$
$$= \frac{1}{\alpha-1} \log \left(\sum_{y\in\mathcal{Y}} L_y M^{\alpha-1} \sum_{x\in f^{-1}(y)} \frac{1}{L_y} P_X^\alpha(x) \right) \tag{56}$$
$$\geq \frac{1}{\alpha-1} \log \left(\sum_{y\in\mathcal{Y}} L_y M^{\alpha-1} \left(\frac{1}{L_y} \sum_{x\in f^{-1}(y)} P_X(x) \right)^\alpha \right) \tag{57}$$
$$= \frac{1}{\alpha-1} \log \sum_{y\in\mathcal{Y}} P_V^\alpha(y) \left(\frac{L_y}{M}\right)^{1-\alpha} \tag{58}$$
$$= D_\alpha(V\|W) \tag{59}$$

where (55) is satisfied by the definition in (30) with U being equiprobable on $\mathcal{X}$ and $|\mathcal{X}| = M$; (56) holds by expressing the sum over $\mathcal{X}$ as a double sum over $\mathcal{Y}$ and the elements of $\mathcal{X}$ that are mapped by f to the same element in $\mathcal{Y}$; inequality (57) holds since $|f^{-1}(y)| = L_y$ for all $y \in \mathcal{Y}$, and for every random variable Z we have $\mathbb{E}[Z^\alpha] \geq \mathbb{E}^\alpha[Z]$ if $\alpha \in (1,\infty)$ or the opposite inequality if $\alpha \in [0,1)$; finally, (58)–(59) are due to (30), (49) and (50). Note that (57) holds with equality if and only if P_X is equiprobable on $f^{-1}(y)$ for every $y \in \mathcal{Y}$ such that $P_X(f^{-1}(y)) > 0$. Hence, for $\alpha \in (0,1)\cup(1,\infty)$, (52) holds with equality if and only if the condition given in the theorem statement is satisfied. ∎

Corollary 1: Let X be a random variable taking values on a finite set $\mathcal{X}$ with $|\mathcal{X}| = M$, and let $\mathcal{L} \subseteq \mathcal{X}$ with $|\mathcal{L}| = L$. Then, for all $\alpha \in [0,1)\cup(1,\infty)$,

$$H_\alpha(X) \leq \frac{1}{1-\alpha} \log\Big(L^{1-\alpha}\, \mathbb{P}^\alpha[X\in\mathcal{L}] + (M-L)^{1-\alpha}\, \mathbb{P}^\alpha[X\notin\mathcal{L}] \Big) \tag{60}$$
$$= \log M - d_\alpha\left(\mathbb{P}[X\in\mathcal{L}] \,\|\, \tfrac{L}{M}\right) \tag{61}$$

with equality in (60) if and only if X is equiprobable on both $\mathcal{L}$ and $\mathcal{L}^{\mathrm{c}}$.

Proof: Inequality (60) is a specialization of Theorem 1 to a binary-valued function $f\colon \mathcal{X} \to \{0,1\}$ such that $f^{-1}(0) = \mathcal{L}$ and $f^{-1}(1) = \mathcal{L}^{\mathrm{c}}$. Equality (61) holds by setting $\theta = M$, $t = \mathbb{P}[X\in\mathcal{L}]$, and $s = L$ in (35). ∎

Moreover, specializing Corollary 1 to $L = 1$, we obtain:

Corollary 2: Let X be a random variable taking values on a finite set $\mathcal{X}$ with $|\mathcal{X}| = M$. Then, for all $\alpha \in [0,1) \cup (1,\infty)$,

$$H_\alpha(X) \le \min_{x\in\mathcal{X}} \frac{1}{1-\alpha} \log\Big(P_X^\alpha(x) + (M-1)^{1-\alpha} \left(1 - P_X(x)\right)^\alpha\Big) \tag{62}$$

$$= \log M - \max_{x\in\mathcal{X}} d_\alpha\big(P_X(x) \| \tfrac{1}{M}\big) \tag{63}$$

with equality in (62) if and only if P_X is equiprobable on $\mathcal{X} \setminus \{x^*\}$ where $x^* \in \mathcal{X}$ attains the maximum in (63).

Remark 2: Taking the limit $\alpha \to 1$ in the right side of (62) yields

$$H(X) \le \min_{x\in\mathcal{X}} \Big\{ \big(1 - P_X(x)\big) \log(M-1) + h\big(P_X(x)\big) \Big\} \tag{64}$$

with the same necessary and sufficient condition for equality in Corollary 2. Furthermore, choosing $x \in \mathcal{X}$ to be a mode of P_X gives

$$H(X) \le \big(1 - p_{\max}\big) \log(M-1) + h\big(p_{\max}\big) \tag{65}$$

$$= \varepsilon_X \log(M-1) + h(\varepsilon_X) \tag{66}$$

where (66) follows from (10), and the symmetry of the binary entropy function around $\frac{1}{2}$.

If we loosen Corollary 2 by, instead of minimizing the right side of (62), choosing $x \in \mathcal{X}$ to be a mode of P_X, we recover [7, Theorem 6]:

Corollary 3: Let X be a random variable taking M possible values, and assume that its largest mass is $p_{\max}$. Then, for all $\alpha \in [0,1) \cup (1,\infty)$,

$$H_\alpha(X) \le \frac{1}{1-\alpha} \log\Big(p_{\max}^\alpha + (M-1)^{1-\alpha} \left(1 - p_{\max}\right)^\alpha\Big) \tag{67}$$

$$= \log M - d_\alpha\big(p_{\max} \| \tfrac{1}{M}\big) \tag{68}$$

with equality in (67) under the condition in the statement of Corollary 2 with x^* being equal to a mode of P_X.

Remark 3: In view of the necessary and sufficient condition for equality in (62), the minimization of the upper bound in the right side of (62) for a given P_X does not necessarily imply that the best choice of $x \in \mathcal{X}$ is a mode of P_X. For example, let $\mathcal{X} = \{0, 1, 2\}$, $P_X(0) = 0.2$, and $P_X(1) = P_X(2) = 0.4$. Then, $H_2(X) = 1.474$ bits, which coincides with its upper bound in (62) by choosing $x = 0$; on the other hand, (67) yields $H_2(X) \le 1.556$ bits.

B. Lower bound

The following lower bound provides a counterpart to the upper bound in Corollary 3 without restricting to finitely valued random variables.

Theorem 2: Let X be a discrete random variable attaining a maximal mass $p_{\max}$. Then, for $\alpha \in (0,1) \cup (1,\infty)$,

$$H_\alpha(X) \ge \frac{1}{1-\alpha} \log\left(\left\lfloor \frac{1}{p_{\max}} \right\rfloor p_{\max}^\alpha + \left(1 - p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor \right)^\alpha \right) \tag{69}$$

$$= \log\left(1 + \left\lfloor \frac{1}{p_{\max}} \right\rfloor\right) - d_\alpha \left(p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor \middle\| \frac{\left\lfloor \frac{1}{p_{\max}} \right\rfloor}{1 + \left\lfloor \frac{1}{p_{\max}} \right\rfloor} \right) \tag{70}$$

where, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to x. Equality in (69) holds if and only if P_X has $\left\lfloor \frac{1}{p_{\max}} \right\rfloor$ masses equal to $p_{\max}$, and an additional mass equal to $1 - p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor$ whenever $\frac{1}{p_{\max}}$ is not an integer.

Proof: For $\alpha \in (0,1) \cup (1,\infty)$, the Rényi entropy of the distribution identified in the statement of Theorem 2 as attaining the condition for equality is equal to the right side of (69). Furthermore, that distribution majorizes any distribution whose maximum mass is $p_{\max}$. The result in (69) follows from the Schur-concavity of the Rényi entropy in the general case of a countable alphabet (see [40, Theorem 2]). In view of Lemma 1 and Theorem 2 of [40], the Schur concavity of the Rényi entropy is strict for any $\alpha \in (0,1) \cup (1,\infty)$ and therefore (69) holds with strict inequality for any distribution other than the one specified in the statement of this result. To get (70), let $s = \left\lfloor \frac{1}{p_{\max}} \right\rfloor$, $t = s\, p_{\max}$ and $\theta = s+1$ in (35). ■

Remark 4: Let X be a discrete random variable, and consider the case of $\alpha = 0$. Unless X is finitely valued, (7) yields $H_0(X) = \infty$. If X is finitely valued, then (7) and the inequality $p_{\max} \, |\operatorname{supp} P_X| \geq 1$ yield

$$H_0(X) \geq \log \left\lceil \frac{1}{p_{\max}} \right\rceil \tag{71}$$

where, for $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer that is larger than or equal to x. The bound in (69) therefore holds for $\alpha = 0$ with the convention that $0^0 = 0$, and it then coincides with (71). Equality in (71) holds if and only if $|\operatorname{supp} P_X| = \left\lceil \frac{1}{p_{\max}} \right\rceil$ (e.g., if X is equiprobable).

Remark 5: Taking the limit $\alpha \to 1$ in Theorem 2 yields

$$H(X) \geq h\left(p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor\right) + p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor \log \left\lfloor \frac{1}{p_{\max}} \right\rfloor \tag{72}$$

$$= \log\left(1 + \left\lfloor \frac{1}{p_{\max}} \right\rfloor\right) - d\left(p_{\max} \left\lfloor \frac{1}{p_{\max}} \right\rfloor \,\middle\|\, \frac{\left\lfloor \frac{1}{p_{\max}} \right\rfloor}{1 + \left\lfloor \frac{1}{p_{\max}} \right\rfloor}\right) \tag{73}$$

with equality in (72) if and only if the condition for equality in Theorem 2 holds. Hence, the result in Theorem 2 generalizes the bound in (72), which is due to Kovalevsky [48] and Tebbe and Dwyer [76] in the special case of a finitely-valued X (rediscovered in [25]), and to Ho and Verdú [39, Theorem 10] in the general setting of a countable alphabet.

IV. Arimoto-Rényi Conditional Entropy and Error Probability

Section IV forms the main part of this paper, and its results are outlined in Section I-D (see Items 18)–22)).

A. Upper bound on the Arimoto-Rényi conditional entropy: Generalized Fano's inequality

The minimum error probability $\varepsilon_{X|Y}$ can be achieved by a deterministic function (*maximum-a-posteriori* decision rule) $\mathcal{L}^*\colon \mathcal{Y} \to \mathcal{X}$:

$$\varepsilon_{X|Y} = \min_{\mathcal{L}\colon \mathcal{Y} \to \mathcal{X}} \mathbb{P}[X \neq \mathcal{L}(Y)] \tag{74}$$

$$= \mathbb{P}[X \neq \mathcal{L}^*(Y)] \tag{75}$$

$$= 1 - \mathbb{E}\left[\max_{x \in \mathcal{X}} P_{X|Y}(x|Y)\right] \tag{76}$$

$$\leq 1 - p_{\max} \tag{77}$$

$$\leq 1 - \frac{1}{M} \tag{78}$$

where (77) is the minimum error probability achievable among blind decision rules that disregard the observations (see (10)).

Fano's inequality links the decision-theoretic uncertainty $\varepsilon_{X|Y}$ and the information-theoretic uncertainty $H(X|Y)$ through

$$H(X|Y) \leq \log M - d\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big) \tag{79}$$

$$= h(\varepsilon_{X|Y}) + \varepsilon_{X|Y} \log(M-1) \tag{80}$$

where the identity in (80) is (39) with $t = \varepsilon_{X|Y}$. Although the form of Fano's inequality in (79) is not nearly as popular as (80), it turns out to be the version that admits an elegant (although not immediate) generalization to the Arimoto-Rényi conditional entropy. It is straightforward to obtain (80) by averaging a conditional version of (66) with respect to the observation. This simple route to the desired result is not viable in the case of $H_\alpha(X|Y)$ since it is not an average of Rényi entropies of conditional distributions. The conventional proof of Fano's inequality (e.g., [18, pp. 38–39]), based on the use of the chain rule for entropy, is also doomed to failure for the Arimoto-Rényi conditional entropy of order $\alpha \neq 1$ since it does not satisfy the chain rule.

Before we generalize Fano's inequality by linking $\varepsilon_{X|Y}$ with $H_\alpha(X|Y)$ for $\alpha \in (0, \infty)$, note that for $\alpha = \infty$, the following generalization of (11) holds in view of (21) and (76):

$$\varepsilon_{X|Y} = 1 - \exp\big(-H_\infty(X|Y)\big). \tag{81}$$

Theorem 3: Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M < \infty$. For all $\alpha \in (0, \infty)$,

$$H_\alpha(X|Y) \leq \log M - d_\alpha\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big). \tag{82}$$

Let $\mathcal{L}^* \colon \mathcal{Y} \to \mathcal{X}$ be a deterministic MAP decision rule. Equality holds in (82) if and only if, for all $y \in \mathcal{Y}$,

$$P_{X|Y}(x|y) = \begin{cases} \dfrac{\varepsilon_{X|Y}}{M-1}, & x \neq \mathcal{L}^*(y), \\ 1 - \varepsilon_{X|Y}, & x = \mathcal{L}^*(y). \end{cases} \tag{83}$$

Proof: If in Corollary 3 we replace P_X by $P_{X|Y=y}$, then $p_{\max}$ is replaced by $\max_{x \in \mathcal{X}} P_{X|Y}(x|y)$ and we obtain

$$H_\alpha(X|Y=y) \leq \frac{1}{1-\alpha} \log\Big(\big(1 - \varepsilon_{X|Y}(y)\big)^\alpha + (M-1)^{1-\alpha} \, \varepsilon^\alpha_{X|Y}(y)\Big) \tag{84}$$

where we have defined the conditional error probability given the observation:

$$\varepsilon_{X|Y}(y) = 1 - \max_{x \in \mathcal{X}} P_{X|Y}(x|y), \tag{85}$$

which satisfies with $Y \sim P_Y$,

$$\varepsilon_{X|Y} = \mathbb{E}[\varepsilon_{X|Y}(Y)]. \tag{86}$$

For $\alpha \in (0,1) \cup (1, \infty)$, $(\beta, \gamma) \in (0, \infty)^2$, define the function $f_{\alpha,\beta,\gamma} \colon [0,1] \to (0, \infty)$:

$$f_{\alpha,\beta,\gamma}(u) = \left(\gamma(1-u)^\alpha + \beta u^\alpha\right)^{\frac{1}{\alpha}}. \tag{87}$$

If $\alpha > 1$, then (16) allows us to bound a monotonically decreasing function of $H_\alpha(X|Y)$:

$$\exp\left(\frac{1-\alpha}{\alpha}\, H_\alpha(X|Y)\right) \geq \mathbb{E}\left[\left(\big(1-\varepsilon_{X|Y}(Y)\big)^\alpha + \beta\, \varepsilon_{X|Y}^\alpha(Y)\right)^{\frac{1}{\alpha}}\right] \tag{88}$$

$$= \mathbb{E}\left[f_{\alpha,\beta,1}\big(\varepsilon_{X|Y}(Y)\big)\right] \tag{89}$$

$$\geq f_{\alpha,\beta,1}\big(\varepsilon_{X|Y}\big) \tag{90}$$

$$= \left(\big(1-\varepsilon_{X|Y}\big)^\alpha + (M-1)^{1-\alpha}\, \varepsilon_{X|Y}^\alpha\right)^{\frac{1}{\alpha}} \tag{91}$$

where $\beta = (M-1)^{1-\alpha}$ in (88)–(90); (88) follows from (16) and (84); (90) follows from (86) and Jensen's inequality, due to the convexity of $f_{\alpha,\beta,\gamma}\colon [0,1] \to (0,\infty)$ for $\alpha \in (1,\infty)$ and $(\beta,\gamma) \in (0,\infty)^2$ (see Lemma 1 following this proof); finally, (91) follows from (87). For $\alpha \in (1,\infty)$, (82) follows from (88)–(91) and identity (35) with $(\theta, s, t) = (M,\, M-1,\, \varepsilon_{X|Y})$.

For $\alpha \in (0,1)$, inequality (88) is reversed; and due to the concavity of $f_{\alpha,\beta,\gamma}(\cdot)$ on $[0,1]$ (Lemma 1), inequality (90) is also reversed. The rest of the proof proceeds as in the case where $\alpha > 1$.

The necessary and sufficient condition for equality in (82) follows from the condition for equality in the statement of Corollary 3 when P_X is replaced by $P_{X|Y=y}$ for $y \in \mathcal{Y}$. Under (83), it follows from (74)–(78) and (85) that $\varepsilon_{X|Y}(y) = \varepsilon_{X|Y}$ for all $y \in \mathcal{Y}$; this implies that inequalities (88) and (90) for $\alpha \in (1,\infty)$, and the opposite inequalities for $\alpha \in (0,1)$, hold with equalities. ■

Lemma 1: Let $\alpha \in (0,1) \cup (1,\infty)$ and $(\beta,\gamma) \in (0,\infty)^2$. The function $f_{\alpha,\beta,\gamma}\colon [0,1] \to (0,\infty)$ defined in (87) is strictly convex for $\alpha \in (1,\infty)$, and strictly concave for $\alpha \in (0,1)$.

Proof: The second derivative of $f_{\alpha,\beta,\gamma}(\cdot)$ in (87) is given by

$$f''_{\alpha,\beta,\gamma}(u) = (\alpha-1)\beta\gamma\Big(\gamma(1-u)^\alpha + \beta u^\alpha\Big)^{\frac{1}{\alpha}-2}\big(u(1-u)\big)^{\alpha-2} \tag{92}$$

which is strictly negative if $\alpha \in (0,1)$, and strictly positive if $\alpha \in (1,\infty)$ for any $u \in [0,1]$. ■

Remark 6: From (37), Fano's inequality (see (79)–(80)) is recovered by taking the limit $\alpha \to 1$ in (82).

Remark 7: A pleasing feature of Theorem 3 is that as $\alpha \to \infty$, the bound becomes tight. To see this, we rewrite the identity in (81) as

$$H_\infty(X|Y) = \log M - \log \frac{1-\varepsilon_{X|Y}}{\frac{1}{M}} \tag{93}$$

$$= \log M - d_\infty\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big) \tag{94}$$

where (94) follows from (38) since $\varepsilon_{X|Y} \leq 1 - \frac{1}{M}$ (see (74)–(78)).

Remark 8: For $\alpha = 0$, (82) also holds. To see this, it is useful to distinguish two cases:

- $\varepsilon_{X|Y} = 0$. Then, $H_0(X|Y) = 0$, and the right side of (82) is also equal to zero since $d_0(0\|q) = -\log(1-q)$ for all $q \in [0,1)$.
- $\varepsilon_{X|Y} > 0$. Then, the right side of (82) is equal to $\log M$ (see (36)) which is indeed an upper bound to $H_0(X|Y)$. The condition for equality in this case is that there exists $y \in \mathcal{Y}$ such that $P_{X|Y}(x|y) > 0$ for all $x \in \mathcal{X}$.

Remark 9: Since $d_\alpha\big(\cdot \| 1 - \frac{1}{M}\big)$ is monotonically decreasing in $[0, 1-\frac{1}{M}]$, Theorem 3 gives an implicit lower bound on $\varepsilon_{X|Y}$. Although, currently, there is no counterpart to Shannon's explicit lower bound as a function of $H(X|Y)$ [71], we do have explicit lower bounds as a function of $H_\alpha(X|Y)$ in Section IV-B.

Remark 10: If X and Y are random variables taking values on a set of M elements and X is equiprobable, then [56, Theorem 5.3] shows that

$$I_\alpha(X;Y) \geq d_\alpha(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}), \tag{95}$$

which, together with (47), yields (82). However, note that in Theorem 3 we do not restrict X to be equiprobable.

In information-theoretic problems, it is common to encounter the case in which X and Y are actually vectors of dimension n. Fano's inequality ensures that vanishing error probability implies vanishing normalized conditional entropy as $n \to \infty$. As we see next, the picture with the Arimoto-Rényi conditional entropy is more nuanced.

Theorem 4: Let $\{X_n\}$ be a sequence of random variables, with X_n taking values on $\mathcal{X}_n$ for $n \in \{1, 2, 3, \ldots\}$, and assume that there exists an integer $M \geq 2$ such that $|\mathcal{X}_n| \leq M^n$ for all n.[1] Let $\{Y_n\}$ be an arbitrary sequence of random variables, for which $\varepsilon_{X_n|Y_n} \to 0$ as $n \to \infty$.

a) If $\alpha \in (1, \infty]$, then $H_\alpha(X_n|Y_n) \to 0$;
b) If $\alpha = 1$, then $\frac{1}{n} H(X_n|Y_n) \to 0$;
c) If $\alpha \in [0, 1)$, then $\frac{1}{n} H_\alpha(X_n|Y_n)$ is upper bounded by $\log M$; nevertheless, it does not necessarily tend to 0.

Proof:

a) For $\alpha \in (1, \infty)$,

$$H_\alpha(X_n|Y_n) \leq n \log M - d_\alpha\big(\varepsilon_{X_n|Y_n} \,\|\, 1 - M^{-n}\big) \tag{96}$$

$$= \frac{1}{1-\alpha} \log \left(\varepsilon^{\alpha}_{X^n|Y^n} (M^n - 1)^{1-\alpha} + \big(1 - \varepsilon_{X_n|Y_n}\big)^{\alpha} \right), \tag{97}$$

where (96) follows from (82) and $|\mathcal{X}_n| \leq M^n$; (97) holds due to (35) by setting the parameters $\theta = M^n$, $s = M^n - 1$ and $t = \varepsilon_{X^n|Y^n}$; hence, $H_\alpha(X_n|Y_n) \to 0$ since $\alpha > 1$ and $\varepsilon_{X_n|Y_n} \to 0$. Item a) also holds for $\alpha = \infty$ since $H_\alpha(X_n|Y_n)$ is monotonically decreasing in α throughout the real line (Proposition 1).

b) For $\alpha = 1$, from Fano's inequality,

$$\tfrac{1}{n} H(X_n|Y_n) \leq \varepsilon_{X_n|Y_n} \log M + \tfrac{1}{n} h\big(\varepsilon_{X_n|Y_n}\big) \to 0. \tag{98}$$

Hence, not only does $\frac{1}{n} H(X_n|Y_n) \to 0$ if $\varepsilon_{X_n|Y_n} = o(1)$ but also $H(X_n|Y_n) \to 0$ if $\varepsilon_{X_n|Y_n} = o\left(\frac{1}{n}\right)$.

c) Proposition 1 implies that if $\alpha > 0$, then

$$H_\alpha(X_n|Y_n) \leq H_\alpha(X_n) \tag{99}$$

$$\leq \log |\mathcal{X}_n| \tag{100}$$

$$\leq n \log M. \tag{101}$$

■

A counterexample where $\varepsilon_{X_n|Y_n} \to 0$ exponentially fast, and yet $\frac{1}{n} H_\alpha(X_n|Y_n) \not\to 0$ for all $\alpha \in [0, 1)$ is given as follows.

Remark 11: In contrast to the conventional case of $\alpha = 1$, it is not possible to strengthen Theorem 4c) to claim that $\varepsilon_{X_n|Y_n} \to 0$ implies that $\frac{1}{n} H_\alpha(X_n|Y_n) \to 0$ for $\alpha \in [0, 1)$. By Proposition 1, it is sufficient to consider the following counterexample: fix any $\alpha \in (0, 1)$, and let Y_n be deterministic, $\mathcal{X}_n = \{1, \ldots, M^n\}$, and $X_n \sim P_{X_n}$ where

$$P_{X_n} = \left[1 - \beta^{-n},\ \frac{\beta^{-n}}{M^n - 1}, \ldots, \frac{\beta^{-n}}{M^n - 1} \right] \tag{102}$$

[1] Note that this encompasses the conventional setting in which $\mathcal{X}_n = \mathcal{A}^n$.

with

$$\beta = M^{\frac{1-\alpha}{2\alpha}} > 1. \tag{103}$$

Then,

$$\varepsilon_{X_n|Y_n} = \beta^{-n} \to 0, \tag{104}$$

and

$$\begin{align}
H_\alpha(X_n|Y_n) &= H_\alpha(X_n) \tag{105}\\
&= \frac{1}{1-\alpha} \log\left(\left(1-\beta^{-n}\right)^\alpha + (M^n-1)\left(\frac{\beta^{-n}}{M^n-1}\right)^\alpha \right) \tag{106}\\
&= \frac{1}{1-\alpha} \log\left(\left(1-\beta^{-n}\right)^\alpha + (M^n-1)^{1-\alpha} M^{-\frac{n(1-\alpha)}{2}} \right) \tag{107}\\
&= \tfrac{1}{2}\, n \log M + \frac{1}{1-\alpha} \log\Big((1-\beta^{-n})^\alpha M^{-\frac{n(1-\alpha)}{2}} + (1-M^{-n})^{1-\alpha} \Big) \tag{108}
\end{align}$$

where (105) holds since Y_n is deterministic; (106) follows from (5) and (102); (107) holds due to (103). Consequently, since $\alpha \in (0,1)$, $M \geq 2$, and $\beta > 1$, the second term in the right side of (108) tends to 0. In conclusion, normalizing (105)–(108) by n, and letting $n \to \infty$ yields

$$\lim_{n\to\infty} \frac{1}{n} H_\alpha(X_n|Y_n) = \tfrac{1}{2} \log M. \tag{109}$$

Remark 12: Theorem 4b) is due to [39, Theorem 15]. Furthermore, [39, Example 2] shows that Theorem 4b) cannot be strengthened to $H(X_n|Y_n) \to 0$, in contrast to the case where $\alpha \in (1, \infty]$ in Theorem 4a).

B. *Explicit lower bounds on $\varepsilon_{X|Y}$*

The results in Section IV-A yield implicit lower bounds on $\varepsilon_{X|Y}$ as a function of the Arimoto-Rényi conditional entropy. In this section, we obtain several explicit bounds. As the following result shows, Theorem 3 readily results in explicit lower bounds on $\varepsilon_{X|Y}$ as a function of $H_{\frac{1}{2}}(X|Y)$ and of $H_2(X|Y)$.

Theorem 5: Let X be a discrete random variable taking $M \geq 2$ possible values. Then,

$$\varepsilon_{X|Y} \geq \left(1 - \frac{1}{M}\right) \frac{1}{\xi_1} \left(1 - \sqrt{\frac{\xi_1 - 1}{M-1}}\right)^2, \tag{110}$$

$$\varepsilon_{X|Y} \geq \left(1 - \frac{1}{M}\right) \left(1 - \sqrt{\frac{\xi_2 - 1}{M-1}}\right) \tag{111}$$

where

$$\xi_1 = M \exp\big(-H_{\frac{1}{2}}(X|Y)\big), \tag{112}$$

$$\xi_2 = M \exp\big(-H_2(X|Y)\big). \tag{113}$$

Proof: Since $0 \leq H_2(X|Y) \leq H_{\frac{1}{2}}(X|Y) \leq \log M$, (112)–(113) imply that $1 \leq \xi_1 \leq \xi_2 \leq M$. To prove (110), setting $\alpha = \frac{1}{2}$ in (82) and using (35) with $(\theta, s, t) = (M, M-1, \varepsilon_{X|Y})$ we obtain

$$\sqrt{1 - \varepsilon_{X|Y}} + \sqrt{(M-1)\varepsilon_{X|Y}} \geq \exp\left(\tfrac{1}{2}\, H_{\frac{1}{2}}(X|Y)\right). \tag{114}$$

Substituting

$$v = \sqrt{\varepsilon_{X|Y}}, \tag{115}$$

$$z = \exp\left(\tfrac{1}{2}\, H_{\frac{1}{2}}(X|Y)\right) \tag{116}$$

yields the inequality

$$\sqrt{1-v^2} \geq z - \sqrt{M-1}\, v. \tag{117}$$

If the right side of (117) is non-negative, then (117) is transformed to the following quadratic inequality in v:

$$Mv^2 - 2\sqrt{M-1}\, zv + (z^2 - 1) \leq 0 \tag{118}$$

which yields

$$v \geq \frac{1}{M}\left(\sqrt{M-1}\, z - \sqrt{M - z^2}\right). \tag{119}$$

If, however, the right side of (117) is negative then

$$v > \frac{z}{\sqrt{M-1}} \tag{120}$$

which implies the satisfiability of (119) also in the latter case. Hence, (119) always holds. In view of (112), (115), (116), and since $\xi_1 \in [1, M]$, it can be verified that the right side of (119) is non-negative. Squaring both sides of (119), and using (112), (115) and (116) give (110).

Similarly, setting $\alpha = 2$ in (82) and using (35) with $(\theta, s, t) = (M, M-1, \varepsilon_{X|Y})$ yield a quadratic inequality in $\varepsilon_{X|Y}$, from which (111) follows. ■

Remark 13: Following up on this work, Renes [60] has generalized (110) to the quantum setting.

Remark 14: The corollary to Theorem 3 in (111) is equivalent to [21, Theorem 3].

Remark 15: Consider the special case where X is an equiprobable binary random variable, and Y is a discrete random variable which takes values on a set $\mathcal{Y}$. Following the notation in [44, (7)], let $\rho \in [0, 1]$ denote the Bhattacharyya coefficient

$$\rho = \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|0) P_{Y|X}(y|1)} \tag{121}$$

$$= 2 \sum_{y \in \mathcal{Y}} P_Y(y) \sqrt{P_{X|Y}(0|y)\, P_{X|Y}(1|y)} \tag{122}$$

where (122) holds due to Bayes' rule which implies that $P_{Y|X}(y|x) = 2P_{X|Y}(x|y)P_Y(y)$ for all $x \in \{0, 1\}$ and $y \in \mathcal{Y}$. From (14) and (122), we obtain

$$H_{\frac{1}{2}}(X|Y) = \log(1 + \rho). \tag{123}$$

Since X is a binary random variable, it follows from (112) and (123) that $\xi_1 = \frac{2}{1+\rho}$; hence, the lower bound on the minimal error probability in (110) is given by

$$\varepsilon_{X|Y} \geq \tfrac{1}{2}\left(1 - \sqrt{1-\rho^2}\right) \tag{124}$$

recovering the bound in [44, (49)] (see also [79]).

Remark 16: The lower bounds on $\varepsilon_{X|Y}$ in (110) and (111) depend on $H_\alpha(X|Y)$ with $\alpha = \frac{1}{2}$ and $\alpha = 2$, respectively; due to their dependence on different orders α, none of these bounds is superseded by the other, and they both prove to be useful in view of their relation to the conditional Bayesian distance in [21, Definition 2] and the Bhattacharyya coefficient (see Remarks 14 and 15).

Remark 17: Taking the limit $M \to \infty$ in the right side of (111) yields

$$\varepsilon_{X|Y} \geq 1 - \exp\left(-\tfrac{1}{2}\, H_2(X|Y)\right) \tag{125}$$

which is equivalent to [81, (8)] (see also [21, Theorem 2]), and its loosening yields [21, Corollary 1]. Theorem 7 (see also Theorem 10) tightens (125).

Remark 18: Arimoto [3] introduced a different generalization of entropy and conditional entropy parameterized by a continuously differentiable function $\mathtt{f}\colon (0,1] \to [0,\infty)$ satisfying $\mathtt{f}(1) = 0$:

$$H_{\mathtt{f}}(X) = \inf_{Y} \mathbb{E}\left[\mathtt{f}(P_Y(X))\right], \tag{126}$$

$$H_{\mathtt{f}}(X|Y) = \mathbb{E}\left[H_{\mathtt{f}}(P_{X|Y}(\cdot|Y))\right], \tag{127}$$

where the infimum is over all the distributions defined on the same set as X. Arimoto [3, Theorem 3] went on to show the following generalization of Fano's inequality:

$$H_{\mathtt{f}}(X|Y) \leq \min_{\theta \in (0,1)} \left\{ \mathbb{P}[X = Y]\, \mathtt{f}(1-\theta) + \mathbb{P}[X \neq Y]\, \mathtt{f}\left(\frac{\theta}{M-1}\right) \right\}. \tag{128}$$

A functional dependence can be established between the Rényi entropy and $H_{\mathtt{f}}(X)$ for a certain choice of $\mathtt{f}$ (see [3, Example 2]); in view of (16), the Arimoto-Rényi conditional entropy can be expressed in terms of $H_{\mathtt{f}}(X|Y)$, although the analysis of generalizing Fano's inequality with the Arimoto-Rényi conditional entropy becomes rather convoluted following this approach.

For convenience, we assume throughout the rest of this subsection that

$$P_{X|Y}(x|y) > 0, \quad (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{129}$$

The following bound, which is a special case of Theorem 9, involves the Arimoto-Rényi conditional entropy of negative orders, and bears some similarity to Arimoto's converse for channel coding [4].

Theorem 6: Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M < \infty$, which satisfies (129). For all $\alpha \in (-\infty, 0)$,

$$\varepsilon_{X|Y} \geq \exp\left(\frac{1-\alpha}{\alpha}\Big[H_\alpha(X|Y) - \log(M-1)\Big]\right). \tag{130}$$

It can be verified that the bound in [64, (23)] is equivalent to (130). A different approach can be found in the proof of Theorem 9.

Remark 19: By the assumption in (129), it follows from (27) that, for $\alpha \in (-\infty, 0)$, the quantity in the exponent in the right side of (130) satisfies

$$H_\alpha(X|Y) - \log(M-1) \geq H_0(X|Y) - \log(M-1) \tag{131}$$

$$= \log \frac{M}{M-1}. \tag{132}$$

Hence, by letting $\alpha \to 0$, the bound in (130) is trivial; while by letting $\alpha \to -\infty$, it follows from (23), (129) and (130) that

$$\varepsilon_{X|Y} \geq (M-1)\, \mathbb{E}\Big[\min_{x \in \mathcal{X}} P_{X|Y}(x|Y)\Big]. \tag{133}$$

The $\alpha \in (-\infty, 0)$ that results in the tightest bound in (130) is examined numerically in Example 2, which illustrates the utility of Arimoto-Rényi conditional entropies of negative orders.

Remark 20: For binary hypothesis testing, the lower bound on $\varepsilon_{X|Y}$ in (130) is asymptotically tight by letting $\alpha \to -\infty$ since, in view of (24) and (81),

$$\varepsilon_{X|Y} = \exp\bigl(-H_{-\infty}(X|Y)\bigr). \tag{134}$$

Next, we provide a lower bound depending on the Arimoto-Rényi conditional entropy of orders greater than 1. A more general version of this result is given in Theorem 10 by relying on our generalization of Fano's inequality for list decoding.

Theorem 7: Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ which satisfies (129), with $\mathcal{X}$ being finite or countably infinite. For all $\alpha \in (1, \infty)$

$$\varepsilon_{X|Y} \geq 1 - \exp\left(\frac{1-\alpha}{\alpha}\, H_\alpha(X|Y)\right). \tag{135}$$

Proof: In view of the monotonicity property for positive orders in (28), we obtain that

$$H_\infty(X|Y) \geq \frac{\alpha - 1}{\alpha}\, H_\alpha(X|Y). \tag{136}$$

and the desired result follows in view of (81). (Note that for $\alpha \in (0, 1]$, the right side of (135) is nonpositive.) ■

Remark 21: The implicit lower bound on $\varepsilon_{X|Y}$ given in (91) (same as (82)) is tighter than the explicit lower bound in (135).

Remark 22: The lower bounds on $\varepsilon_{X|Y}$ in Theorems 3 and 7, which hold for positive orders of $H_\alpha(X|Y)$, are both asymptotically tight by letting $\alpha \to \infty$. In contrast, the lower bound on $\varepsilon_{X|Y}$ of Theorem 6, which holds for negative orders of $H_\alpha(X|Y)$, is not asymptotically tight by letting $\alpha \to -\infty$ (unless X is a binary random variable).

In the following example, the lower bounds on $\varepsilon_{X|Y}$ in Theorems 3 and 7 are examined numerically in their common range of $\alpha \in (1, \infty)$.

Example 1: Let X and Y be random variables defined on the set $\mathcal{X} = \{1, 2, 3\}$, and let

$$\bigl[P_{XY}(x,y)\bigr]_{(x,y)\in\mathcal{X}^2} = \frac{1}{45}\begin{pmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{pmatrix}. \tag{137}$$

It can be verified that $\varepsilon_{X|Y} = \frac{21}{45} \approx 0.4667$. Note that although in this example the bound in (135) is only slightly looser than the bound in (82) for moderate values of $\alpha > 1$ (see Table I)[2], both are indeed asymptotically tight as $\alpha \to \infty$; furthermore, (135) has the advantage of providing a closed-form lower bound on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$ for $\alpha \in (1, \infty)$.

Example 2: Let X and Y be random variables defined on the set $\mathcal{X} = \{1, 2, 3, 4\}$, and let

$$\bigl[P_{XY}(x,y)\bigr]_{(x,y)\in\mathcal{X}^2} = \frac{1}{400}\begin{pmatrix} 10 & 38 & 10 & 26 \\ 32 & 20 & 44 & 20 \\ 10 & 29 & 10 & 35 \\ 41 & 20 & 35 & 20 \end{pmatrix}. \tag{138}$$

In this case $\varepsilon_{X|Y} = \frac{121}{200} = 0.6050$, and the tightest lower bound in (130) for $\alpha \in (-\infty, 0)$ is equal to 0.4877 (obtained at $\alpha = -2.531$). Although $\varepsilon_{X|Y}$ can be calculated exactly when $H_\infty(X|Y)$ is known (see (81)), this example illustrates that conditional Arimoto-Rényi entropies of negative orders are useful in the sense that (130) gives an informative lower bound on $\varepsilon_{X|Y}$ by knowing $H_\alpha(X|Y)$ for a negative α.

[2] Recall that for $\alpha = 2$, (82) admits the explicit expression in (111).

α	(82)	(135)
2	0.4247	0.3508
4	0.4480	0.4406
6	0.4573	0.4562
8	0.4620	0.4613
10	0.4640	0.4635
50	0.4667	0.4667

TABLE I. LOWER BOUNDS ON $\varepsilon_{X|Y}$ IN (82) AND (135) FOR EXAMPLE 1.

C. *List decoding*

In this section we consider the case where the decision rule outputs a list of choices. The extension of Fano's inequality to list decoding was initiated in [1, Section 5] (see also [59, Appendix 3.E]). It is useful for proving converse results in conjunction with the blowing-up lemma [20, Lemma 1.5.4]. The main idea of the successful combination of these two tools is that, given an arbitrary code, one can blow-up the decoding sets in such a way that the probability of decoding error can be as small as desired for sufficiently large blocklengths; since the blown-up decoding sets are no longer disjoint, the resulting setup is a list decoder with subexponential list size.

A generalization of Fano's inequality for list decoding of size L is [84][3]

$$H(X|Y) \leq \log M - d\big(P_{\mathcal{L}} \| 1 - \tfrac{L}{M}\big), \tag{139}$$

where $P_{\mathcal{L}}$ denotes the probability of X not being in the list. As we noted before, averaging a conditional version of (48) with respect to the observation is not viable in the case of $H_\alpha(X|Y)$ with $\alpha \neq 1$ (see (14)). A pleasing generalization of (139) to the Arimoto-Rényi conditional entropy does indeed hold as the following result shows.

Theorem 8: Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M$. Consider a decision rule[4] $\mathcal{L} \colon \mathcal{Y} \to \binom{\mathcal{X}}{L}$, and denote the decoding error probability by

$$P_{\mathcal{L}} = \mathbb{P}\big[X \notin \mathcal{L}(Y)\big]. \tag{140}$$

Then, for all $\alpha \in (0,1) \cup (1,\infty)$,

$$H_\alpha(X|Y) \leq \log M - d_\alpha\big(P_{\mathcal{L}} \| 1 - \tfrac{L}{M}\big) \tag{141}$$

$$= \frac{1}{1-\alpha} \, \log\Big(L^{1-\alpha}\left(1 - P_{\mathcal{L}}\right)^\alpha + (M-L)^{1-\alpha} \, P_{\mathcal{L}}^\alpha\Big) \tag{142}$$

with equality in (141) if and only if

$$P_{X|Y}(x|y) = \begin{cases} \dfrac{P_{\mathcal{L}}}{M-L}, & x \notin \mathcal{L}(y) \\[2ex] \dfrac{1-P_{\mathcal{L}}}{L}, & x \in \mathcal{L}(y). \end{cases} \tag{143}$$

Proof: Instead of giving a standalone proof, for brevity, we explain the differences between this and the proof of Theorem 3:

---

[3]See [46, Lemma 1] for a weaker version of (139).

[4]$\binom{\mathcal{X}}{L}$ stands for the set of all the subsets of $\mathcal{X}$ with cardinality L, with $L \leq |\mathcal{X}|$.

- Instead of the conditional version of (67), we use a conditional version of (60) given the observation $Y = y$;
- The choices of the arguments of the function $f_{\alpha,\beta,\gamma}$ in (88)–(90) namely, $\beta = (M-1)^{1-\alpha}$ and $\gamma = 1$, are replaced by $\beta = (M-L)^{1-\alpha}$ and $\gamma = L^{1-\alpha}$;
- (142) follows from (35) with $(\theta, s, t) = (M, M-L, P_{\mathcal{L}})$;
- Equality in (141) holds if and only if the condition for equality in the statement of Corollary 1, conditioned on the observation, is satisfied; the latter implies that, given $Y = y$, X is equiprobable on both $\mathcal{L}(y)$ and $\mathcal{L}^{\mathrm{c}}(y)$ for all $y \in \mathcal{Y}$.

■

Next, we give the fixed list-size generalization of Theorem 6.

Theorem 9: Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M < \infty$, which satisfies (129), and let $\mathcal{L} \colon \mathcal{Y} \to \binom{\mathcal{X}}{L}$. Then, for all $\alpha \in (-\infty, 0)$, the probability that the decoding list does not include the correct decision satisfies

$$P_{\mathcal{L}} \geq \exp\left(\frac{1-\alpha}{\alpha}\left[H_\alpha(X|Y) - \log(M-L)\right]\right). \tag{144}$$

Proof: Let $\xi \colon \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be given by the indicator function

$$\xi(x,y) = 1\big\{x \notin \mathcal{L}(y)\big\}. \tag{145}$$

Let $u \colon \mathcal{X} \times \mathcal{Y} \to [0,\infty)$, and $\beta > 1$. By Hölder's inequality,

$$\mathbb{E}\big[\xi(X,Y)\, u(X,Y)\big] \leq \mathbb{E}^{\frac{1}{\beta}}\left[\xi^\beta(X,Y)\right] \mathbb{E}^{\frac{\beta-1}{\beta}}\left[u^{\frac{\beta}{\beta-1}}(X,Y)\right]. \tag{146}$$

Therefore,

$$\begin{aligned} P_{\mathcal{L}} &= \mathbb{E}\big[\xi^\beta(X,Y)\big] && (147)\\ &\geq \mathbb{E}^\beta\big[\xi(X,Y)\, u(X,Y)\big]\; \mathbb{E}^{1-\beta}\big[u^{\frac{\beta}{\beta-1}}(X,Y)\big]. && (148) \end{aligned}$$

Under the assumption in (129), we specialize (148) to

$$u(x,y) = \frac{1}{P_{X|Y}(x|y)} \left(\sum_{x' \in \mathcal{X}} P_{X|Y}^{\frac{1}{1-\beta}}(x'|y)\right)^{1-\beta} \tag{149}$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. From (14), (145) and (149), we have

$$\begin{aligned} \mathbb{E}\big[\xi(X,Y)\, u(X,Y)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\xi(X,Y)\, u(X,Y) \,\big|\, Y\big]\Big] && (150)\\ &= \mathbb{E}\left[\sum_{x \in \mathcal{X}} P_{X|Y}(x|Y)\, u(x,Y)\, \xi(x,Y)\right] && (151)\\ &= \mathbb{E}\left[\left(\sum_{x' \in \mathcal{X}} P_{X|Y}^{\frac{1}{1-\beta}}(x'|Y)\right)^{1-\beta} \sum_{x \in \mathcal{X}} \xi(x,Y)\right] && (152)\\ &= \mathbb{E}\left[\left(\sum_{x' \in \mathcal{X}} P_{X|Y}^{\frac{1}{1-\beta}}(x'|Y)\right)^{1-\beta} |\mathcal{L}^{\mathrm{c}}(Y)|\right] && (153)\\ &= (M-L) \exp\Big(-\beta\, H_{\frac{1}{1-\beta}}(X|Y)\Big) && (154) \end{aligned}$$

and

$$\mathbb{E}\left[u^{\frac{\beta}{\beta-1}}(X,Y)\right] = \mathbb{E}\Big[\mathbb{E}\left[u^{\frac{\beta}{\beta-1}}(X,Y)\,\big|\,Y\right]\Big] \tag{155}$$

$$= \mathbb{E}\left[\sum_{x\in\mathcal{X}} P_{X|Y}(x|Y)\, u^{\frac{\beta}{\beta-1}}(x,Y)\right] \tag{156}$$

$$= \mathbb{E}\left[\left(\sum_{x\in\mathcal{X}} P_{X|Y}^{\frac{1}{1-\beta}}(x|Y)\right)^{1-\beta}\right] \tag{157}$$

$$= \exp\bigl(-\beta\, H_{\frac{1}{1-\beta}}(X|Y)\bigr). \tag{158}$$

Assembling (148), (154) and (158) and substituting $\alpha = \frac{1}{1-\beta}$ results in (144). ■

The fixed list-size generalization of Theorem 7 is the following.

Theorem 10: Let P_{XY} be a probability measure defined on $\mathcal{X}\times\mathcal{Y}$ which satisfies (129), with $\mathcal{X}$ being finite or countably infinite, and let $\mathcal{L}\colon \mathcal{Y}\to\binom{\mathcal{X}}{L}$. Then, for all $\alpha\in(1,\infty)$,

$$P_{\mathcal{L}} \geq 1-\exp\left(\frac{1-\alpha}{\alpha}\,\Big[H_\alpha(X|Y)-\log L\Big]\right). \tag{159}$$

Proof: From (142), for all $\alpha\in(1,\infty)$,

$$H_\alpha(X|Y) \leq \frac{1}{1-\alpha}\,\log\bigl(L^{1-\alpha}\,(1-P_{\mathcal{L}})^{\alpha}\bigr) \tag{160}$$

$$= \log L + \frac{\alpha}{1-\alpha}\,\log(1-P_{\mathcal{L}}) \tag{161}$$

which gives the bound in (159). ■

Remark 23: The implicit lower bound on $\varepsilon_{X|Y}$ given by the generalized Fano's inequality in (141) is tighter than the explicit lower bound in (159) (this can be verified from the proof of Theorem 10).

D. Lower bounds on the Arimoto-Rényi conditional entropy

The major existing lower bounds on the Shannon conditional entropy $H(X|Y)$ as a function of the minimum error probability $\varepsilon_{X|Y}$ are:

1) In view of [39, Theorem 11], (1) (shown in [6, Theorem 1], [15, (12)] and [38, (41)] for finite alphabets) holds for a general discrete random variable X. As an example where (1) holds with equality, consider a Z-channel with input and output alphabets $\mathcal{X}=\mathcal{Y}=\{0,1\}$, and assume that $P_X(0)=1-P_X(1)=p\in\bigl(0,\frac{1}{2}\bigr]$, and

$$P_{Y|X}(0|0)=1, \quad P_{Y|X}(0|1)=\frac{p}{1-p}. \tag{162}$$

It can be verified that $\varepsilon_{X|Y}=p$, and $H(X|Y)=2p$ bits.

2) Due to Kovalevsky [48], Tebbe and Dwyer [76] (see also [25]) in the finite alphabet case, and to Ho and Verdú [39, (109)] in the general case,

$$\phi\bigl(\varepsilon_{X|Y}\bigr) \leq H(X|Y) \tag{163}$$

where $\phi\colon[0,1)\to[0,\infty)$ is the piecewise linear function that is defined on the interval $t\in\bigl[1-\frac{1}{k},1-\frac{1}{k+1}\bigr)$ as

$$\phi(t) = t\,k(k+1)\log\left(\frac{k+1}{k}\right) + (1-k^2)\log(k+1) + k^2\log k \tag{164}$$

where k is an arbitrary positive integer. Note that (163) is tighter than (1) since $\phi(t) \geq 2t \log 2$.

In view of (81), since $H_\alpha(X|Y)$ is monotonically decreasing in α, one can readily obtain the following counterpart to Theorem 7:

$$H_\alpha(X|Y) \geq \log \frac{1}{1-\varepsilon_{X|Y}} \tag{165}$$

for $\alpha \in [0, \infty]$ with equality if $\alpha = \infty$.

Remark 24: As a consequence of (165), it follows that if $\alpha \in [0, \infty)$ and $\{X_n\}$ is a sequence of discrete random variables, then $H_\alpha(X_n|Y_n) \to 0$ implies that $\varepsilon_{X_n|Y_n} \to 0$, thereby generalizing the case $\alpha = 1$ shown in [39, Theorem 14].

Remark 25: Setting $\alpha = 2$ in (165) yields [21, Theorem 1], and setting $\alpha = 1$ in (165) yields (2). Recently, Prasad [58, Section 2.C] improved the latter bound in terms of the entropy of tilted probability measures, instead of $H_\alpha(X|Y)$.

The next result gives a counterpart to Theorem 3, and a generalization of (163).

Theorem 11: If $\alpha \in (0,1) \cup (1,\infty)$, then

$$\frac{\alpha}{1-\alpha} \log g_\alpha(\varepsilon_{X|Y}) \leq H_\alpha(X|Y), \tag{166}$$

where the piecewise linear function $g_\alpha \colon [0,1) \to \mathcal{D}_\alpha$, with $\mathcal{D}_\alpha = [1,\infty)$ for $\alpha \in (0,1)$ and $\mathcal{D}_\alpha = (0,1]$ for $\alpha \in (1,\infty)$, is defined by

$$g_\alpha(t) = \Big(k(k+1)^{\frac{1}{\alpha}} - k^{\frac{1}{\alpha}}(k+1)\Big)\, t + k^{\frac{1}{\alpha}+1} - (k-1)(k+1)^{\frac{1}{\alpha}} \tag{167}$$

on the interval $t \in \big[1-\frac{1}{k}, 1-\frac{1}{k+1}\big)$ for an arbitrary positive integer k.

Proof: For every $y \in \mathcal{Y}$ such that $P_Y(y) > 0$, Theorem 2 and (85) yield

$$H_\alpha(X \,|\, Y = y) \geq s_\alpha\big(\varepsilon_{X|Y}(y)\big) \tag{168}$$

where $s_\alpha \colon [0,1) \to [0,\infty)$ is given by

$$s_\alpha(t) = \frac{1}{1-\alpha} \log \left(\left\lfloor \frac{1}{1-t} \right\rfloor (1-t)^\alpha + \left(1 - (1-t) \left\lfloor \frac{1}{1-t} \right\rfloor \right)^\alpha \right). \tag{169}$$

In the remainder of the proof, we consider the cases $\alpha \in (0,1)$ and $\alpha \in (1,\infty)$ separately.

- $\alpha \in (0,1)$. Define the function $f_\alpha \colon [0,1) \to [1,\infty)$ as

$$f_\alpha(t) = \exp\left(\frac{1-\alpha}{\alpha}\, s_\alpha(t)\right). \tag{170}$$

  The piecewise linear function $g_\alpha(\cdot)$ in (167) coincides with the monotonically increasing function $f_\alpha(\cdot)$ in (170) at $t_k = 1 - \frac{1}{k}$ for every positive integer k (note that $f_\alpha(t_k) = g_\alpha(t_k) = k^{\frac{1}{\alpha}-1}$). It can be also verified that $g_\alpha(\cdot)$ is the lower convex envelope of $f_\alpha(\cdot)$ (i.e., $g_\alpha(\cdot)$ is the largest convex function for which $g_\alpha(t) \leq f_\alpha(t)$ for all $t \in [0,1)$). Hence,

$$\begin{aligned} H_\alpha(X|Y) &\geq \frac{\alpha}{1-\alpha} \log \mathbb{E}\left[f_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] && (171)\\ &\geq \frac{\alpha}{1-\alpha} \log \mathbb{E}\left[g_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] && (172)\\ &\geq \frac{\alpha}{1-\alpha} \log\, g_\alpha(\varepsilon_{X|Y}) && (173)\end{aligned}$$

where (171) follows from (16), (168), (170), and since $f_\alpha\colon [0,1) \to [1,\infty)$ is a monotonically increasing function for $\alpha \in (0,1)$; (172) follows from $f_\alpha \geq g_\alpha$; (86) and Jensen's inequality for the convex function $g_\alpha(\cdot)$ result in (173).

- $\alpha \in (1,\infty)$. The function $f_\alpha(\cdot)$ in (170) is monotonically decreasing in $[0,1)$, and the piecewise linear function $g_\alpha(\cdot)$ in (167) coincides with $f_\alpha(\cdot)$ at $t_k = 1 - \frac{1}{k}$ for every positive integer k. Furthermore, $g_\alpha(\cdot)$ is the smallest concave function that is not below $f_\alpha(\cdot)$ on the interval $[0,1)$. The proof now proceeds as in the previous case, and (171)–(173) continue to hold for $\alpha > 1$.

■

Remark 26: The most useful domain of applicability of Theorem 11 is $\varepsilon_{X|Y} \in [0, \frac{1}{2}]$, in which case the lower bound specializes to ($k = 1$)

$$\frac{\alpha}{1-\alpha} \log\Big(1 + \big(2^{\frac{1}{\alpha}} - 2\big)\varepsilon_{X|Y}\Big) \leq H_\alpha(X|Y) \tag{174}$$

which yields (1) as $\alpha \to 1$.

Remark 27: Theorem 11 is indeed a generalization of (163) since for all $\tau \in [0,1]$,

$$\lim_{\alpha\to 1} \frac{\alpha}{1-\alpha} \log g_\alpha(\tau) = \phi(\tau), \tag{175}$$

with ϕ defined in (164).

Remark 28: As $\alpha \to \infty$, (166) is asymptotically tight: from (167),

$$\lim_{\alpha\to\infty} g_\alpha(t) = 1 - t \tag{176}$$

so the right side of (166) converges to $\log \frac{1}{1-\varepsilon_{X|Y}}$ which, recalling (81), proves the claim.

Remark 29: It can be shown that the function in (167) satisfies

$$g_\alpha(t) \begin{cases} \leq (1-t)^{1-\frac{1}{\alpha}}, & \alpha \in (1,\infty] \\ \geq (1-t)^{1-\frac{1}{\alpha}}, & \alpha \in (0,1). \end{cases} \tag{177}$$

Moreover, if $t = 1 - \frac{1}{k}$, for $k \in \{1,2,3,\ldots\}$, then

$$g_\alpha(t) = (1-t)^{1-\frac{1}{\alpha}}. \tag{178}$$

Therefore, Theorem 11 gives a tighter bound than (165), unless $\varepsilon_{X|Y} \in \big\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots \frac{M-1}{M}\big\}$ (M is allowed to be ∞ here) in which case they are identical, and independent of α as it is illustrated in Figure 1.

The following proposition applies when M is finite, but does not depend on M. It is supported by Figure 1.

Proposition 3: Let $M \in \{2,3,\ldots\}$ be finite, and let the upper and lower bounds on $H_\alpha(X|Y)$ as a function of $\varepsilon_{X|Y}$, as given in Theorems 3 and 11, be denoted by $u_{\alpha,M}(\cdot)$ and $l_\alpha(\cdot)$, respectively. Then,

a) these bounds coincide if and only if X is a deterministic function of the observation Y or X is equiprobable on the set $\mathcal{X}$ and independent of Y;

b) the limit of the ratio of the upper-to-lower bounds when $\varepsilon_{X|Y} \to 0$ is given by

$$\lim_{\varepsilon_{X|Y}\to 0} \frac{u_{\alpha,M}(\varepsilon_{X|Y})}{l_\alpha(\varepsilon_{X|Y})} = \begin{cases} \infty, & \alpha \in (0,1), \\ \dfrac{1}{2 - 2^{\frac{1}{\alpha}}}, & \alpha \in (1,\infty). \end{cases} \tag{179}$$

Proof: See Appendix B. ■

Remark 30: The low error probability limit of the ratio of upper-to-lower bounds in (179) decreases monotonically with $\alpha \in (1,\infty)$ from ∞ to 1.

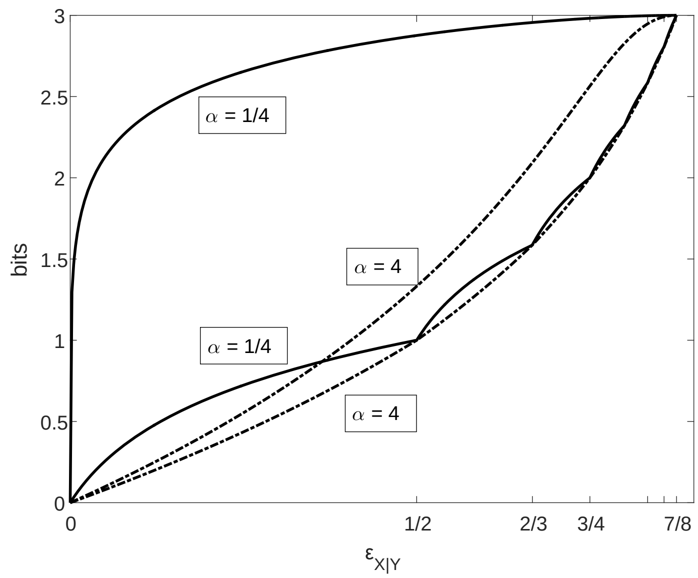

Fig. 1. Upper and lower bounds on $H_\alpha(X|Y)$ in Theorems 3 and 11, respectively, as a function of $\varepsilon_{X|Y} \in [0, 1 - \frac{1}{M}]$ for $\alpha = \frac{1}{4}$ (solid lines) and $\alpha = 4$ (dash-dotted lines) with $M = 8$.

E. Explicit upper bounds on $\varepsilon_{X|Y}$

In this section, we give counterparts to the explicit lower bounds on $\varepsilon_{X|Y}$ given in Section IV-B by capitalizing on the bounds in Section IV-D.

The following result is a consequence of Theorem 11:

Theorem 12: Let $k \in \{1, 2, 3, \dots\}$, and $\alpha \in (0, 1) \cup (1, \infty)$. If $\log k \le H_\alpha(X|Y) < \log(k + 1)$, then

$$\varepsilon_{X|Y} \le \frac{\exp\left(\frac{1-\alpha}{\alpha} H_\alpha(X|Y)\right) - k^{\frac{1}{\alpha}+1} + (k-1)(k+1)^{\frac{1}{\alpha}}}{k(k+1)^{\frac{1}{\alpha}} - k^{\frac{1}{\alpha}}(k+1)}. \tag{180}$$

Furthermore, the upper bound on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$ is asymptotically tight in the limit where $\alpha \to \infty$.

Proof: The left side of (166) is equal to $\log k$ when $\varepsilon_{X|Y} = 1 - \frac{1}{k}$ for $k \in \{1, 2, 3, \dots\}$, and it is also monotonically increasing in $\varepsilon_{X|Y}$. Hence, if $\varepsilon_{X|Y} \in \left[1 - \frac{1}{k}, 1 - \frac{1}{k+1}\right)$, then the lower bound on $H_\alpha(X|Y)$ in the left side of (166) lies in the interval $[\log k, \log(k + 1))$. The bound in (180) can be therefore verified to be equivalent to Theorem 11. Finally, the asymptotic tightness of (180) in the limit where $\alpha \to \infty$ is inherited by the same asymptotic tightness of the equivalent bound in Theorem 11. ■

Remark 31: By letting $\alpha \to 1$ in the right side of (180), we recover the bound by Ho and Verdú [39, (109)]:

$$\varepsilon_{X|Y} \le \frac{H(X|Y) + (k^2 - 1)\log(k + 1) - k^2 \log k}{k(k + 1)\log\left(\frac{k+1}{k}\right)} \tag{181}$$

if $\log k \le H(X|Y) < \log(k + 1)$ for an arbitrary $k \in \{1, 2, 3, \dots\}$. In the special case of finite alphabets, this result was obtained in [25], [48] and [76]. In the finite alphabet setting, Prasad [58, Section 5] recently refined the bound in (181) by lower bounding $H(X|Y)$ subject to the knowledge of the first two largest posterior probabilities rather than only the largest one; following the same approach, [58, Section 6] gives a refinement of Fano's inequality.

Example 1 *(cont.)* Table II compares the asymptotically tight bounds in Theorems 3 and 12 for three values of α.

α	lower bound	$\varepsilon_{X\|Y}$	upper bound
1	0.4013	0.4667	0.6061
10	0.4640	0.4667	0.4994
100	0.4667	0.4667	0.4699

TABLE II. UPPER AND LOWER BOUNDS ON $\varepsilon_{X|Y}$ IN (82) AND (180) FOR EXAMPLE 1.

As the following example illustrates, the bounds in (82) and (180) are in general not monotonic in α.

Example 3: Suppose X and Y are binary random variables with joint distribution

$$\left[P_{XY}(x,y)\right]_{(x,y)\in\{0,1\}^2} = \begin{pmatrix} 0.1906 & 0.3737 \\ 0.4319 & 0.0038 \end{pmatrix}. \tag{182}$$

In this case $\varepsilon_{X|Y} = 0.1944$, and its upper and lower bounds are shown in Figure 2 as a function of α.

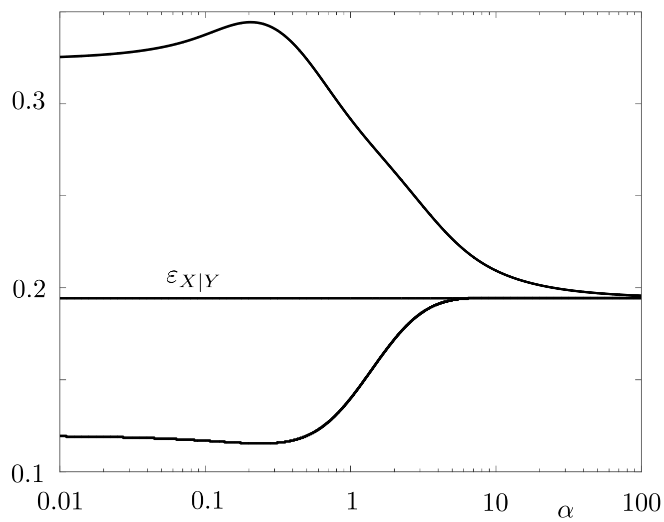

Fig. 2. Upper (Theorem 12) and lower (Theorem 3) bounds on $\varepsilon_{X|Y}$ as a function of α in Example 3.

V. UPPER BOUNDS ON $\varepsilon_{X|Y}$ BASED ON BINARY HYPOTHESIS TESTING

In this section, we give explicit upper bounds on $\varepsilon_{X|Y}$ by connecting the M-ary hypothesis testing problem for finite M with the associated $\binom{M}{2}$ binary hypothesis testing problems. The latter bounds generalize the binary hypothesis testing upper bound on $\varepsilon_{X|Y}$ by Hellman and Raviv [38, Theorem 1] to any number of hypotheses, including the tightening of the upper bound by Kanaya and Han [45], and Leang and Johnson [51].

Hellman and Raviv [38, Theorem 1] generalized the Bhattacharyya bound to give an upper bound on the error probability for *binary* hypothesis testing in terms of the Rényi divergence between both models P_0 and P_1 (with prior probabilities $\mathbb{P}[\mathsf{H}_0]$, and $\mathbb{P}[\mathsf{H}_1]$ respectively):

$$\varepsilon_{X|Y} \leq \inf_{\alpha\in(0,1)} \mathbb{P}^{\alpha}[\mathsf{H}_1]\mathbb{P}^{1-\alpha}[\mathsf{H}_0] \exp\big((\alpha-1)D_{\alpha}(P_1\|P_0)\big). \tag{183}$$

To generalize (183) to any number of hypotheses (including infinity), it is convenient to keep the same notation denoting the models by

$$P_i = P_{Y|X=i} \tag{184}$$

with prior probabilities $\mathbb{P}[\mathsf{H}_i]$ for $i \in \{1, \ldots, M\}$, and let

$$\bar{\mathbb{P}}[\mathsf{H}_i] = 1 - \mathbb{P}[\mathsf{H}_i], \tag{185}$$

$$\overline{P}_i = \sum_{j \neq i} \frac{\mathbb{P}[\mathsf{H}_j]}{\bar{\mathbb{P}}[\mathsf{H}_i]} P_j. \tag{186}$$

Theorem 13:

$$\varepsilon_{X|Y} \leq \min_{i \neq k} \inf_{\alpha \in (0,1)} \bar{\mathbb{P}}^{\alpha}[\mathsf{H}_i] \bar{\mathbb{P}}^{1-\alpha}[\mathsf{H}_k] \exp\left((\alpha - 1) D_\alpha \left(\overline{P}_i \,\|\, \overline{P}_k\right)\right). \tag{187}$$

Proof: Let

$$Y \sim P_Y = \sum_{m=1}^{M} \mathbb{P}[\mathsf{H}_m] \, P_m \tag{188}$$

and, for $m \in \{1, \ldots, M\}$, denote the densities

$$p_m = \frac{\mathrm{d}P_m}{\mathrm{d}P_Y}, \quad \bar{p}_m = \frac{\mathrm{d}\overline{P}_m}{\mathrm{d}P_Y}. \tag{189}$$

For $\alpha \in (0,1)$ and for all $i, k \in \{1, \ldots, M\}$ with $i \neq k$,

$$\varepsilon_{X|Y} = 1 - \mathbb{E}\left[\max_{m \in \{1,\ldots,M\}} \mathbb{P}[\mathsf{H}_m | Y]\right] \tag{190}$$

$$= \mathbb{E}\left[\min_{m \in \{1,\ldots,M\}} \sum_{j \neq m} \mathbb{P}[\mathsf{H}_j] \, p_j(Y)\right] \tag{191}$$

$$= \mathbb{E}\left[\min_{m \in \{1,\ldots,M\}} \bar{\mathbb{P}}[\mathsf{H}_m] \, \bar{p}_m(Y)\right] \tag{192}$$

$$\leq \mathbb{E}\left[\min\left\{\bar{\mathbb{P}}[\mathsf{H}_i] \, \bar{p}_i(Y), \, \bar{\mathbb{P}}[\mathsf{H}_k] \, \bar{p}_k(Y)\right\}\right] \tag{193}$$

$$\leq \mathbb{E}\left[\bar{\mathbb{P}}^{\alpha}[\mathsf{H}_i] \; \bar{p}_i^{\alpha}(Y) \; \bar{\mathbb{P}}^{1-\alpha}[\mathsf{H}_k] \; \bar{p}_k^{1-\alpha}(Y)\right] \tag{194}$$

$$= \bar{\mathbb{P}}^{\alpha}[\mathsf{H}_i] \, \bar{\mathbb{P}}^{1-\alpha}[\mathsf{H}_k] \exp\left((\alpha - 1) D_\alpha \left(\overline{P}_i \,\|\, \overline{P}_k\right)\right) \tag{195}$$

where

- (191) follows from Bayes' rule:

$$1 - \mathbb{P}[\mathsf{H}_m | Y] = \sum_{j \neq m} \mathbb{P}[\mathsf{H}_j | Y] = \sum_{j \neq m} \mathbb{P}[\mathsf{H}_j] \, p_j(Y), \tag{196}$$

- (192) follows from the definition in (186),
- (194) follows from $\min\{t, s\} \leq t^\alpha s^{1-\alpha}$ if $\alpha \in [0,1]$ and $t, s \geq 0$,
- (195) follows from (29).

Minimizing the right side of (195) over all (i, k) with $i \neq k$ gives (187). ■

We proceed to generalize Theorem 13 along a different direction. We can obtain upper and lower bounds on the Bayesian M-ary minimal error probability in terms of the minimal error probabilities of the $\binom{M}{2}$ associated binary hypothesis testing problems. For that end, denote the random variable X restricted to $\{i,j\} \subseteq \mathcal{X}$, $i \neq j$, by X_{ij}, namely,

$$\mathbb{P}[X_{ij} = i] = \frac{\mathbb{P}[\mathsf{H}_i]}{\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]}, \tag{197}$$

$$\mathbb{P}[X_{ij} = j] = \frac{\mathbb{P}[\mathsf{H}_j]}{\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]}. \tag{198}$$

Guessing the hypothesis without the benefit of observations yields the error probability

$$\varepsilon_{X_{ij}} = \min\{\mathbb{P}[X_{ij} = i], \mathbb{P}[X_{ij} = j]\} \tag{199}$$

$$= \frac{\min\{\mathbb{P}[\mathsf{H}_i], \mathbb{P}[\mathsf{H}_j]\}}{\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]}. \tag{200}$$

The minimal error probability of the binary hypothesis test

$$\mathsf{H}_i : Y_{ij} \sim P_{Y|X_{ij}=i} = P_{Y|X=i} = P_i, \tag{201}$$

$$\mathsf{H}_j : Y_{ij} \sim P_{Y|X_{ij}=j} = P_{Y|X=j} = P_j \tag{202}$$

with the prior probabilities in (197) and (198) is denoted by $\varepsilon_{X_{ij}|Y_{ij}}$, and it is given by

$$\varepsilon_{X_{ij}|Y_{ij}} = \mathbb{E}\left[\min\{\mathbb{P}[\mathsf{H}_i|Y_{ij}],\, \mathbb{P}[\mathsf{H}_j|Y_{ij}]\}\right], \tag{203}$$

with

$$Y_{ij} \sim P_{Y_{ij}} = \mathbb{P}[X_{ij} = i]P_i + \mathbb{P}[X_{ij} = j]P_j, \tag{204}$$

and

$$\mathbb{P}[\mathsf{H}_i|Y_{ij} = y] = \mathbb{P}[X_{ij} = i]\, \frac{\mathrm{d}P_i}{\mathrm{d}P_{Y_{ij}}}(y). \tag{205}$$

The error probability achieved by the Bayesian M-ary MAP decision rule can be upper bounded in terms of the error probabilities in (203) of the $\frac{1}{2}\, M(M-1)$ binary hypothesis tests in (201)–(202), with $1 \leq i < j \leq M$, by means of the following result.

Theorem 14:

$$\varepsilon_{X|Y} \leq \sum_{1 \leq i < j \leq M} \mathbb{E}\left[\min\{\mathbb{P}[\mathsf{H}_i|Y], \mathbb{P}[\mathsf{H}_j|Y]\}\right] \tag{206}$$

$$= \sum_{1 \leq i < j \leq M} \left(\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]\right)\, \varepsilon_{X_{ij}|Y_{ij}} \tag{207}$$

where $Y \sim P_Y$ with P_Y in (188), and $Y_{ij} \sim P_{Y_{ij}}$ in (204).

Proof: Every probability mass function $(q_1, \ldots, q_M)$ satisfies

$$1 - \max\{q_1, \ldots, q_M\} \leq \sum_{1 \leq i < j \leq M} \min\{q_i, q_j\} \tag{208}$$

because, due to symmetry, if without any loss of generality q_1 attains the maximum in the left side of (208), then the partial sum in the right side of (208) over the $M-1$ terms that involve q_1 is equal $q_2+\ldots+q_M = 1-q_1$ which is equal to the left side of (208). Hence,

$$\varepsilon_{X|Y} = \int \varepsilon_{X|Y=y}\, \mathrm{d}P_Y(y) \tag{209}$$

$$= \int \left(1 - \max\left\{\mathbb{P}[\mathsf{H}_1|Y=y], \ldots, \mathbb{P}[\mathsf{H}_M|Y=y]\right\}\right) \mathrm{d}P_Y(y) \tag{210}$$

$$\leq \int \sum_{1\leq i<j\leq M} \min\{\mathbb{P}[\mathsf{H}_i|Y=y],\, \mathbb{P}[\mathsf{H}_j|Y=y]\}\, \mathrm{d}P_Y(y) \tag{211}$$

$$= \sum_{1\leq i<j\leq M} \mathbb{E}\left[\min\{\mathbb{P}[\mathsf{H}_i|Y], \mathbb{P}[\mathsf{H}_j|Y]\}\right] \tag{212}$$

where (211) follows from (208) with $q_i \leftarrow \mathbb{P}[\mathsf{H}_i|Y=y]$ for $i \in \{1,\ldots,M\}$ and $y \in \mathcal{Y}$. This proves (206).

To show (207), note that for all $i, j \in \{1,\ldots,M\}$ with $i \neq j$,

$$(\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j])\, \varepsilon_{X_{ij}|Y_{ij}}$$

$$= (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) \int \min\big\{\mathbb{P}[\mathsf{H}_i|Y_{ij}=y], \mathbb{P}[\mathsf{H}_j|Y_{ij}=y]\big\}\, \mathrm{d}P_{Y_{ij}}(y) \tag{213}$$

$$= (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) \int \min\big\{\mathbb{P}[X_{ij}=i]\, \mathrm{d}P_i(y),\, \mathbb{P}[X_{ij}=j]\, \mathrm{d}P_j(y)\big\} \tag{214}$$

$$= \int \min\big\{\mathbb{P}[\mathsf{H}_i]\, \mathrm{d}P_i(y),\, \mathbb{P}[\mathsf{H}_j]\, \mathrm{d}P_j(y)\big\} \tag{215}$$

$$= \mathbb{E}\left[\min\{\mathbb{P}[\mathsf{H}_i|Y], \mathbb{P}[\mathsf{H}_j|Y]\}\right] \tag{216}$$

where (213) and (214) hold due to (203) and (205), respectively; (215) follows from (197) and (198); (216) follows from the equality $\mathbb{P}[\mathsf{H}_i|Y=y] = \mathbb{P}[\mathsf{H}_i]\, \frac{\mathrm{d}P_i}{\mathrm{d}P_Y}(Y=y)$. Finally, (207) follows from (206) and (213)–(216). ■

The following immediate consequence of Theorem 14 is the bound obtained in [51, (3)].

Corollary 4:

$$\varepsilon_{X|Y} \leq \tfrac{1}{2} M(M-1) \max_{i\neq j} \varepsilon_{X_{ij}|Y_{ij}}. \tag{217}$$

We proceed to give upper bounds on $\varepsilon_{X|Y}$ based on Rényi divergence and Chernoff information.

Theorem 15:

$$\varepsilon_{X|Y} \leq \sum_{1\leq i<j\leq M} \inf_{\alpha\in(0,1)} \mathbb{P}^{\alpha}[\mathsf{H}_i]\, \mathbb{P}^{1-\alpha}[\mathsf{H}_j]\, \exp\big((\alpha-1) D_\alpha(P_i\|P_j)\big) \tag{218}$$

$$\leq (M-1) \exp\Big(-\min_{i\neq j} C(P_i\|P_j)\Big). \tag{219}$$

Proof: Consider the binary hypothesis test in (201)–(202) with its minimal error probability $\varepsilon_{X_{ij}|Y_{ij}}$. Substituting the prior probabilities in (197)–(198) into the Hellman-Raviv bound (183), we obtain

$$(\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j])\, \varepsilon_{X_{ij}|Y_{ij}}$$

$$\leq (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) \inf_{\alpha\in(0,1)} \Big\{\mathbb{P}^{\alpha}[X_{ij}=i]\, \mathbb{P}^{1-\alpha}[X_{ij}=j]\, \exp\big((\alpha-1) D_\alpha(P_i\|P_j)\big)\Big\} \tag{220}$$

$$= \inf_{\alpha\in(0,1)} \mathbb{P}^{\alpha}[\mathsf{H}_i]\, \mathbb{P}^{1-\alpha}[\mathsf{H}_j] \exp\big((\alpha-1) D_\alpha(P_i\|P_j)\big) \tag{221}$$

which, upon substitution in (207), yields (218). To show (219), note that for all $t, s \geq 0$ and $\alpha \in (0,1)$

$$t^\alpha \, s^{1-\alpha} \leq \max\{t, s\} \leq t + s. \tag{222}$$

Hence, we get

$$\begin{aligned}
\varepsilon_{X|Y} &\leq \sum_{1 \leq i < j \leq M} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) \inf_{\alpha \in (0,1)} \exp\bigl((\alpha - 1) D_\alpha(P_i \| P_j)\bigr) && (223)\\
&= \sum_{1 \leq i < j \leq M} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) \exp\bigl(-C(P_i \| P_j)\bigr) && (224)\\
&\leq \exp\Bigl(-\min_{i \neq j} C(P_i \| P_j)\Bigr) \sum_{1 \leq i < j \leq M} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) && (225)
\end{aligned}$$

where (223) holds due to (218) and (222) with $t \leftarrow \mathbb{P}[\mathsf{H}_i]$ and $s \leftarrow \mathbb{P}[\mathsf{H}_j]$; (224) holds by the definition in (41) with $P \leftarrow P_i$ and $Q \leftarrow P_j$. The sum in the right side of (225) satisfies

$$\begin{aligned}
\sum_{1 \leq i < j \leq M} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) &= \tfrac{1}{2} \sum_{i \neq j} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) && (226)\\
&= \tfrac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} (\mathbb{P}[\mathsf{H}_i] + \mathbb{P}[\mathsf{H}_j]) - \sum_{i=1}^{M} \mathbb{P}[H_i] && (227)\\
&= M - 1. && (228)
\end{aligned}$$

The result in (219) follows from (223)–(228). ■

Remark 32: The bound on $\varepsilon_{X|Y}$ in (219) can be improved by relying on the middle term in (222), which then gives that for $t, s \geq 0$ and $\alpha \in (0,1)$

$$t^\alpha \, s^{1-\alpha} \leq \max\{t, s\} = \tfrac{1}{2}(t + s + |t - s|). \tag{229}$$

Following the analysis in (223)–(228), we get

$$\varepsilon_{X|Y} \leq \left(\frac{M-1}{2} + \frac{1}{2} \sum_{1 \leq i < j \leq M} \bigl| \mathbb{P}[\mathsf{H}_i] - \mathbb{P}[\mathsf{H}_j] \bigr| \right) \exp\Bigl(-\min_{i \neq j} C(P_i \| P_j)\Bigr), \tag{230}$$

which yields an improvement over (219) by a factor of at most $\frac{1}{2}$, attained when X is equiprobable.

Remark 33: Using the identity

$$(1 - \alpha) D_\alpha(P \| Q) = \alpha D_{1-\alpha}(Q \| P) \tag{231}$$

for all $\alpha \in (0,1)$ and probability measures P and Q, it is easy to check that Theorems 13 and 15 result in the same bound for binary hypothesis testing.

Remark 34: Kanaya and Han [45] and Leang and Johnson [51] give a weaker bound with $\frac{1}{2} M(M-1)$ in lieu of $M - 1$ in (219).

Remark 35: In the absence of observations, the bound in (187) is tight since $\varepsilon_{X|Y} = 1 - \max_i \mathbb{P}[\mathsf{H}_i]$. On the other hand, (217) and (218) need not be tight in that case, and in fact may be strictly larger than 1.

Numerical experimentation with Examples 1–3 and others shows that, in general, the Arimoto-Rényi conditional entropy bounds are tighter than (187) and (218).

VI. Arimoto-Rényi Conditional Entropy Averaged over Codebook Ensembles

In this section we consider the channel coding setup with a code ensemble $\mathcal{C}$, over which we are interested in averaging the Arimoto-Rényi conditional entropy of the channel input given the channel output. We denote such averaged quantity by $\mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big]$ where $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$. Some motivation for this study arises from the fact that the normalized equivocation $\frac{1}{n} H(X^n|Y^n)$ as a reliability measure was used by Shannon [69] in his proof that reliable communication is impossible at rates above channel capacity; furthermore, the asymptotic convergence to zero of the equivocation $H(X^n|Y^n)$ at rates below capacity was studied by Feinstein [26, Section 3].

We can capitalize on the convexity of $d_\alpha(\cdot\|q)$ for $\alpha \in [0,1]$ (see [23, Theorem 11] and [23, Section 6.A]) to claim that in view of Theorem 3, for any ensemble $\mathcal{C}$ of size-M codes and blocklength n,

$$\mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big] \leq \log M - d_\alpha\big(\bar{\varepsilon}\,\|1 - \tfrac{1}{M}\big) \tag{232}$$

where $\alpha \in [0,1]$ and $\bar{\varepsilon}$ is the error probability of the maximum-likelihood decoder averaged over the code ensemble, or an upper bound thereof which does not exceed $1 - \frac{1}{M}$ (for such upper bounds see, e.g., [66, Chapters 1–4]). Analogously, in view of Theorem 8 for list decoding with a fixed list size L, the same convexity argument implies that

$$\mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big] \leq \log M - d_\alpha\big(\overline{P_{\mathcal{L}}}\,\|1 - \tfrac{L}{M}\big) \tag{233}$$

where $\overline{P_{\mathcal{L}}}$ denotes the minimum list decoding error probability averaged over the ensemble, or an upper bound thereof which does not exceed $1 - \frac{L}{M}$ (for such bounds see, e.g., [28], [30, Problem 5.20], [42, Section 5] and [53]).

In the remainder of this section, we consider the discrete memoryless channel (DMC) model

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i), \tag{234}$$

an ensemble $\mathcal{C}$ of size-M codes and blocklength n such that the M messages are assigned independent codewords, drawn i.i.d. with per-letter distribution P_X on the input alphabet

$$P_{X^n}(x^n) = \prod_{i=1}^{n} P_X(x_i). \tag{235}$$

In this setting, Feder and Merhav show in [25] the following result for the Shannon conditional entropy.

Theorem 16: [25, Theorem 3] For a DMC with transition probability matrix $P_{Y|X}$, the conditional entropy of the transmitted codeword given the channel output, averaged over a random coding selection with per-letter distribution P_X such that $I(P_X, P_{Y|X}) > 0$, is bounded by

$$\mathbb{E}_{\mathcal{C}}\big[H(X^n|Y^n)\big] \leq \inf_{\rho\in(0,1]} \left(1 + \frac{1}{\rho}\right) \exp\left(-n\rho\left(I_{\frac{1}{1+\rho}}(P_X, P_{Y|X}) - R\right)\right) \log \mathrm{e} \tag{236}$$

$$\leq \left(1 + \frac{1}{\rho^*(R, P_X)}\right) \exp\big(-nE_{\mathrm{r}}(R, P_X)\big) \log \mathrm{e} \tag{237}$$

with

$$R = \tfrac{1}{n} \log M \leq I(P_X, P_{Y|X}), \tag{238}$$

where E_{r} in the right side of (237) denotes the random-coding error exponent, given by (recall (44))

$$E_{\mathrm{r}}(R, P_X) = \max_{\rho \in [0,1]} \rho \left(I_{\frac{1}{1+\rho}}(P_X, P_{Y|X}) - R \right), \tag{239}$$

and the argument that maximizes (239) is denoted in (237) by $\rho^*(R, P_X)$.

Remark 36: Since $H_\alpha(X^n|Y^n)$ is monotonically decreasing in α (see Proposition 1), the upper bounds in (236)–(237) also apply to $\mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big]$ for $\alpha \in [1, \infty]$.

We are now ready to state and prove the main result in this section.

Theorem 17: The following results hold under the setting in Theorem 16:

1) For all $\alpha > 0$, and rates R below the channel capacity C,

$$\limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big] \leq E_{\mathrm{sp}}(R), \tag{240}$$

where $E_{\mathrm{sp}}(\cdot)$ denotes the sphere-packing error exponent

$$E_{\mathrm{sp}}(R) = \sup_{\rho \geq 0} \rho \left(\max_{Q_X} I_{\frac{1}{1+\rho}}(Q_X, P_{Y|X}) - R \right) \tag{241}$$

with the maximization in the right side of (241) over all single-letter distributions Q_X defined on the input alphabet.

2) For all $\alpha \in (0, 1)$,

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big] \geq \alpha E_{\mathrm{r}}(R, P_X) - (1-\alpha)R, \tag{242}$$

provided that

$$R < R_\alpha(P_X, P_{Y|X}) \tag{243}$$

where $R_\alpha(P_X, P_{Y|X})$ is the unique solution $r \in (0, I(P_X, P_{Y|X}))$ to

$$E_{\mathrm{r}}(r, P_X) = \left(\frac{1}{\alpha} - 1\right) r. \tag{244}$$

3) The rate $R_\alpha(P_X, P_{Y|X})$ is monotonically increasing and continuous in $\alpha \in (0, 1)$, and

$$\lim_{\alpha \downarrow 0} R_\alpha(P_X, P_{Y|X}) = 0, \tag{245}$$

$$\lim_{\alpha \uparrow 1} R_\alpha(P_X, P_{Y|X}) = I(P_X, P_{Y|X}). \tag{246}$$

Proof: In proving the first two items we actually give an asymptotic lower bound and a non-asymptotic upper bound, which yield the respective results upon taking $\limsup_{n\to\infty} -\frac{1}{n}\log(\cdot)$ and $\liminf_{n\to\infty} -\frac{1}{n}\log(\cdot)$, respectively.

1) From (174) and the sphere-packing lower bound [72, Theorem 2] on the decoding error probability, the following inequality holds for all rates $R \in [0, C)$ and $\alpha \in (0,1) \cup (1, \infty)$,

$$\mathbb{E}_{\mathcal{C}}\big[H_\alpha(X^n|Y^n)\big] \geq \frac{\alpha}{1-\alpha} \log\Big(1 + \big(2^{\frac{1}{\alpha}} - 2\big) \exp\big(-n[E_{\mathrm{sp}}(R - o_1(n)) + o_2(n)]\big)\Big) \tag{247}$$

where $o_1(n)$ and $o_2(n)$ tend to zero as $n \to \infty$. Note that (174) holds when the error probability of the optimal code lies in $[0, \frac{1}{2}]$, but this certainly holds for all $R < C$ and sufficiently large n (since this error probability decays exponentially to zero).

2) Note that $E_{\mathrm{r}}(R, P_X) > 0$ since, for $\alpha \in (0, 1)$,

$$R < R_\alpha(P_X, P_{Y|X}) < I(P_X, P_{Y|X}), \tag{248}$$

where the left inequality in (248) is the condition in (243), and the right inequality in (248) holds by the definition of $R_\alpha(P_X, P_{Y|X})$ in (244) (see its equivalent form in (255)). In view of the random coding bound [29], let

$$\bar{\varepsilon}_n = \bar{\varepsilon}_n(R, P_X) \triangleq \min\Bigl\{\exp\bigl(-nE_{\mathrm{r}}(R, P_X)\bigr),\, 1 - \exp(-nR)\Bigr\} \tag{249}$$

be an upper bound on the decoding error probability of the code ensemble $\mathcal{C}$, which does not exceed $1 - \frac{1}{M}$ with $M = \exp(nR)$. From (232), for all $\alpha \in (0,1)$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\bigl[H_\alpha(X^n|Y^n)\bigr] \\
&\leq nR - d_\alpha\bigl(\bar{\varepsilon}_n \,\|\, 1 - \exp(-nR)\bigr) && (250)\\
&= \tfrac{1}{1-\alpha}\, \log\Bigl((1-\bar{\varepsilon}_n)^\alpha + \bigl(\exp(nR) - 1\bigr)^{1-\alpha}\, \bar{\varepsilon}_n^{\,\alpha}\Bigr) && (251)\\
&\leq \tfrac{1}{1-\alpha}\, \log\Bigl((1-\bar{\varepsilon}_n)^\alpha + \exp\bigl(n(1-\alpha)R\bigr)\, \bar{\varepsilon}_n^{\,\alpha}\Bigr) && (252)\\
&\leq \tfrac{1}{1-\alpha}\, \log\biggl(\Bigl(1 - \exp\bigl(-nE_{\mathrm{r}}(R, P_X)\bigr)\Bigr)^\alpha + \exp\Bigl(-n\bigl[\alpha E_{\mathrm{r}}(R, P_X) - (1-\alpha)R\bigr]\Bigr)\biggr), && (253)
\end{aligned}$$

where equality (251) follows from (35) by setting $(\theta, s, t) = (\exp(nR),\, \exp(nR) - 1,\, \bar{\varepsilon}_n)$; the inequality in (253) holds with equality for all sufficiently large n such that $\bar{\varepsilon}_n = \exp\bigl(-nE_{\mathrm{r}}(R, P_X)\bigr)$ (see (249)). The result in (242) follows consequently by taking $\liminf\limits_{n\to\infty} -\frac{1}{n}\log(\cdot)$ of (250)–(253). To that end, note that the second exponent in the right side of (253) decays to zero whenever

$$\begin{aligned}
R &< R_\alpha(P_X, P_{Y|X}) && (254)\\
&= \sup\Bigl\{r \in [0, I(P_X, P_{Y|X})) : E_{\mathrm{r}}(r, P_X) > \bigl(\tfrac{1}{\alpha} - 1\bigr)\, r\Bigr\}. && (255)
\end{aligned}$$

It can be verified that the representation of $R_\alpha(P_X, P_{Y|X})$ in (255) is indeed equivalent to the way it is defined in (244) because $E_{\mathrm{r}}(r, P_X)$ is monotonically decreasing and continuous in r (see [30, (5.6.32)]), the right side of (244) is strictly monotonically increasing and continuous in r for $\alpha \in (0,1)$, and the left and right sides of (244) vanish, respectively, at $r = I(P_X, P_{Y|X})$ and $r = 0$. The equivalent representation of $R_\alpha(P_X, P_{Y|X})$ in (255) yields the uniqueness of the solution $r = R_\alpha(P_X, P_{Y|X})$ of (244) in the interval $(0, I(P_X, P_{Y|X}))$ for all $\alpha \in (0,1)$, and it justifies in particular the right inequality in (248).

3) From (244), for $\alpha \in (0,1)$, $r = R_\alpha(P_X, P_{Y|X})$ is the unique solution $r \in (0, I(P_X, P_{Y|X}))$ to

$$\tfrac{1}{r}\, E_{\mathrm{r}}(r, P_X) = \tfrac{1}{\alpha} - 1. \tag{256}$$

Since $E_{\mathrm{r}}(r, P_X)$ is positive and monotonically decreasing in r on the interval $(0, I(P_X, P_{Y|X}))$, so is the left side of (256) as a function of r. Since also the right side of (256) is positive and monotonically decreasing in $\alpha \in (0,1)$, it follows from (256) that $r = R_\alpha(P_X, P_{Y|X})$ is monotonically increasing in α. The continuity of $R_\alpha(P_X, P_{Y|X})$ in α on the interval $(0,1)$ holds due to (256) and the continuity of $E_{\mathrm{r}}(\cdot, P_X)$ (see [30, p. 143]). The limits in (245) and (246) follow from (244): (245) holds since we have $\lim\limits_{r\downarrow 0} E_{\mathrm{r}}(r, P_X) = E_{\mathrm{r}}(0, P_X) \in (0, \infty)$, whereas $\lim\limits_{\alpha\downarrow 0} \bigl(\frac{1}{\alpha} - 1\bigr) = +\infty$; (246) holds since $E_{\mathrm{r}}(R, P_X)$ is equal to zero at $R = I(P_X, P_{Y|X})$, and it is positive if $R < I(P_X, P_{Y|X})$ (see [30, p. 142]).

■

Remark 37: For $\alpha = 1$, Theorem 17 strengthens the result in Theorem 16 by also giving the converse in Item 1). This result enables to conclude that if P_X is the input distribution which maximizes the random-coding error exponent

$$E_{\mathrm{r}}(R) = \max_{Q} E_{\mathrm{r}}(R, Q), \tag{257}$$

then

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{E}_{\mathcal{C}}\big[H(X^n|Y^n)\big] = E_{\mathrm{r}}(R) \tag{258}$$

at all rates between the critical rate R_{c} and channel capacity C of the DMC, since for all $R \in [R_{\mathrm{c}}, C]$ [30, Section 5.8]

$$E_{\mathrm{r}}(R) = E_{\mathrm{sp}}(R) \tag{259}$$

where $E_{\mathrm{sp}}(\cdot)$ is given in (241).

Remark 38: With respect to (259), recall that although the random-coding error exponent is tight for the average code [31], it coincides with the sphere-packing error exponent (for the optimal code) at the high-rate region between the critical rate and channel capacity. As shown by Gallager in [31], the fact that the random coding error exponent is not tight for optimal codes at rates below the critical rate of the DMC stems from the poor performance of the bad codes in this ensemble rather than a weakness of the bounding technique in [29].

Remark 39: The result in Theorem 17 can be extended to list decoding with a fixed size L by relying on (233), and the upper bound on the list decoding error probability in [30, Problem 5.20] where

$$\overline{P_{\mathcal{L}}} = \min\left\{\exp\big(-nE_L(R, P_X)\big), 1 - \frac{L}{M}\right\} \tag{260}$$

with $M = \exp(nR)$, and

$$E_L(R, P_X) = \max_{\rho\in[0,L]} \rho \left(I_{\frac{1}{1+\rho}}(P_X, P_{Y|X}) - R\right). \tag{261}$$

This result can be further generalized to structured code ensembles by relying on upper bounds on the list decoding error probability in [42, Section 5].

The following result is obtained by applying Theorem 17 in the setup of communication over a memoryless binary-input output-symmetric channel with a symmetric input distribution. The following rates, specialized to the considered setup, are required for the presentation of our next result.

1) The cutoff rate is given by

$$R_0 = E_0(1, P_X^*) \tag{262}$$

$$= 1 - \log\big(1 + \rho_{Y|X}\big) \text{ bits} \tag{263}$$

where in the right side of (262), $E_0(\rho, P_X^*)$ is given in (45) with the symmetric binary input distribution $P_X^* = \left[\frac{1}{2}\ \frac{1}{2}\right]$, and $\rho_{Y|X} \in [0, 1]$ in the right side of (263) denotes the Bhattacharyya constant given by

$$\rho_{Y|X} = \sum_{y\in\mathcal{Y}} \sqrt{P_{Y|X}(y|0)\, P_{Y|X}(y|1)} \tag{264}$$

with an integral replacing the sum in the right side of (264) if $\mathcal{Y}$ is a non-discrete set.

2) The critical rate is given by

$$R_{\mathrm{c}} = E_0'(1, P_X^*) \tag{265}$$

where the differentiation of E_0 in the right side of (265) is with respect to ρ at $\rho = 1$.

3) The channel capacity is given by

$$C = I(P_X^*, P_{Y|X}). \tag{266}$$

In general, the random coding error exponent is given by (see [30, Section 5.8])

$$E_{\mathrm{r}}(R) = \begin{cases} R_0 - R, & 0 \le R \le R_{\mathrm{c}} \\ E_{\mathrm{sp}}(R), & R_{\mathrm{c}} \le R \le C, \end{cases} \tag{267}$$

and

$$0 \le R_{\mathrm{c}} \le R_0 \le C. \tag{268}$$

Consequently, $E_{\mathrm{r}}(R)$ in (267) is composed of two parts: a straight line at rates $R \in [0, R_{\mathrm{c}}]$, starting from R_0 at zero rate with slope -1, and the sphere-packing error exponent at rates $R \in [R_{\mathrm{c}}, C]$.

Theorem 18: Let $P_{Y|X}$ be the transition probability matrix of a memoryless binary-input output-symmetric channel, and let $P_X^* = \left[\frac{1}{2}\ \frac{1}{2}\right]$. Let R_{c}, R_0, and C denote the critical and cutoff rates and the channel capacity, respectively, and let[5]

$$\alpha_{\mathrm{c}} = \frac{R_{\mathrm{c}}}{R_0} \in (0, 1). \tag{269}$$

The rate $R_\alpha = R_\alpha(P_X^*, P_{Y|X})$, as introduced in Theorem 17–2) with the symmetric input distribution P_X^*, can be expressed as follows:

a) for $\alpha \in (0, \alpha_{\mathrm{c}}]$,

$$R_\alpha = \alpha R_0; \tag{270}$$

b) for $\alpha \in (\alpha_{\mathrm{c}}, 1)$, $R_\alpha \in (R_{\mathrm{c}}, C)$ is the solution to

$$E_{\mathrm{sp}}(r) = \left(\frac{1}{\alpha} - 1\right) r; \tag{271}$$

R_α is continuous and monotonically increasing in $\alpha \in [\alpha_{\mathrm{c}}, 1)$ from R_{c} to C.

Proof: For every memoryless binary-input output-symmetric channel, the symmetric input distribution P_X^* achieves the maximum of the error exponent $E_{\mathrm{r}}(R)$ in the right side of (257). Consequently, (270) and (271) readily follow from (244), (259), (267) and (269). Due to Theorem 17-3), R_α is monotonically increasing and continuous in $\alpha \in (0, 1)$. From (269) and (270)

$$R_{\alpha_{\mathrm{c}}} = R_{\mathrm{c}}, \tag{272}$$

and (246) and (266) yield

$$\lim_{\alpha \uparrow 1} R_\alpha = I(P_X^*, P_{Y|X}) = C. \tag{273}$$

Consequently, R_α is monotonically increasing in $\alpha \in [\alpha_{\mathrm{c}}, 1)$ from R_{c} to C. ■

[5]In general $\alpha_{\mathrm{c}} \in [0, 1]$ (see (268)); the cases $\alpha_{\mathrm{c}} = 0$ and $\alpha_{\mathrm{c}} = 1$ imply that Theorem 18-a) or b) are not applicable, respectively.

Example 4: Let $P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1-\delta$, $P_{Y|X}(0|1) = P_{Y|X}(1|0) = \delta$, and $P_X(0) = P_X(1) = \frac{1}{2}$. In this case, it is convenient to express all rates in bits; in particular, it follows from (262)–(266) that the cutoff rate, critical rate and channel capacity are given, respectively, by (see, e.g., [30, p. 146])

$$R_0 = 1 - \log\bigl(1 + \sqrt{4\delta(1-\delta)}\bigr), \tag{274}$$

$$R_{\mathrm{c}} = 1 - h\left(\frac{\sqrt{\delta}}{\sqrt{\delta} + \sqrt{1-\delta}}\right), \tag{275}$$

$$C = I(P_X, P_{Y|X}) = 1 - h(\delta). \tag{276}$$

The sphere-packing error exponent for the binary symmetric channel is given by (see, e.g., [30, (5.8.26)–(5.8.27)])

$$E_{\mathrm{sp}}(R) = d\bigl(\delta_{\mathrm{GV}}(R) \,\|\, \delta\bigr) \tag{277}$$

where the normalized Gilbert-Varshamov distance (see, e.g., [63, Theorem 4.10]) is denoted by

$$\delta_{\mathrm{GV}}(R) = h^{-1}(1-R) \tag{278}$$

with $h^{-1}\colon [0, \log 2] \to \bigl[0, \frac{1}{2}\bigr]$ standing for the inverse of the binary entropy function.

In view of Theorem 18, Figure 3 shows R_α with $\alpha \in (0,1)$ for the case where the binary symmetric channel has capacity $\frac{1}{2}$ bit per channel use, namely, $\delta = h^{-1}\bigl(\frac{1}{2}\bigr) = 0.110$, in which case $\alpha_{\mathrm{c}} = 0.5791$ and $R_{\alpha_{\mathrm{c}}} = R_{\mathrm{c}} = 0.1731$ bits per channel use (see (275)).

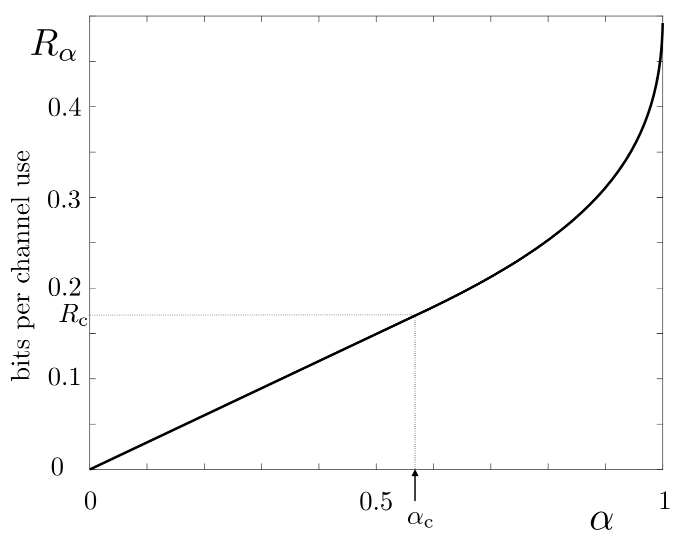

Fig. 3. The rate R_α for $\alpha \in (0,1)$ for a binary symmetric channel with crossover probability $\delta = 0.110$.

VII. Conclusions

The interplay between information-theoretic measures and the analysis of hypothesis testing has been a fruitful area of research. Along these lines, we have shown new bounds on the minimum Bayesian error probability $\varepsilon_{X|Y}$ of arbitrary M-ary hypothesis testing problems as a function of information measures. In particular, our major focus has been the Arimoto-Rényi conditional entropy of the hypothesis index given the observation. We have seen how changing the conventional form of Fano's inequality from

$$H(X|Y) \leq h(\varepsilon_{X|Y}) + \varepsilon_{X|Y} \log(M-1) \tag{279}$$

$$= \log M - d\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big) \tag{280}$$

to the right side of (280), where $d(\cdot\|\cdot)$ is the binary relative entropy, allows a natural generalization where the Arimoto-Rényi conditional entropy of an arbitrary positive order α is upper bounded by

$$H_\alpha(X|Y) \leq \log M - d_\alpha\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big) \tag{281}$$

with $d_\alpha(\cdot\|\cdot)$ denoting the binary Rényi divergence.

Likewise, thanks to the Schur-concavity of Rényi entropy, we obtain a lower bound on $H_\alpha(X|Y)$ in terms of $\varepsilon_{X|Y}$, which holds even if $M = \infty$. Again, we are able to recover existing bounds by letting $\alpha \to 1$.

In addition to the aforementioned bounds on $H_\alpha(X|Y)$ as a function of $\varepsilon_{X|Y}$, it is also of interest to give explicit lower and upper bounds on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$. As $\alpha \to \infty$, these bounds converge to $\varepsilon_{X|Y}$.

The list decoding setting, in which the hypothesis tester is allowed to output a subset of given cardinality and an error occurs if the true hypothesis is not in the list, has considerable interest in information theory. We have shown that our techniques readily generalize to that setting and we have found generalizations of all the $H_\alpha(X|Y)$–$\varepsilon_{X|Y}$ bounds to the list decoding setting.

We have also explored some facets of the role of binary hypothesis testing in analyzing M-ary hypothesis testing problems, and have shown new bounds in terms of Rényi divergence.

As an illustration of the application of the $H_\alpha(X|Y)$–$\varepsilon_{X|Y}$ bounds, we have analyzed the exponentially vanishing decay of the Arimoto-Rényi conditional entropy of the transmitted codeword given the channel output for discrete memoryless channels and random coding ensembles.

Appendix A
Proof of Proposition 1

We prove that $H_\alpha(X|Y) \leq H_\beta(X|Y)$ in three different cases:

- $1 < \beta < \alpha$
- $0 < \beta < \alpha < 1$
- $\beta < \alpha < 0$.

Proposition 1 then follows by transitivity, and the continuous extension of $H_\alpha(X|Y)$ at $\alpha = 0$ and $\alpha = 1$. The following notation is handy.

$$\theta = \frac{\beta - 1}{\alpha - 1} > 0, \tag{282}$$

$$\rho = \frac{\alpha\theta}{\beta} > 0. \tag{283}$$

Case 1: If $1 < \beta < \alpha$, then $(\rho, \theta) \in (0,1)^2$, and

$$\begin{aligned}
&\exp\bigl(-H_\alpha(X|Y)\bigr) \\
&= \mathbb{E}^{\frac{\beta\rho}{\beta-1}} \left[\left(\sum_{x \in \mathcal{X}} P^\alpha_{X|Y}(x|Y) \right)^{\frac{1}{\alpha}} \right] && (284)\\
&\geq \mathbb{E}^{\frac{\beta}{\beta-1}} \left[\left(\sum_{x \in \mathcal{X}} P^\alpha_{X|Y}(x|Y) \right)^{\frac{\rho}{\alpha}} \right] && (285)\\
&= \mathbb{E}^{\frac{\beta}{\beta-1}} \left[\left(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)\, P^{\alpha-1}_{X|Y}(x|Y) \right)^{\frac{\theta}{\beta}} \right] && (286)\\
&\geq \mathbb{E}^{\frac{\beta}{\beta-1}} \left[\left(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)\, P^{(\alpha-1)\theta}_{X|Y}(x|Y) \right)^{\frac{1}{\beta}} \right] && (287)\\
&= \mathbb{E}^{\frac{\beta}{\beta-1}} \left[\left(\sum_{x \in \mathcal{X}} P^\beta_{X|Y}(x|Y) \right)^{\frac{1}{\beta}} \right] && (288)\\
&= \exp\bigl(-H_\beta(X|Y)\bigr). && (289)
\end{aligned}$$

where the outer expectations are with respect to $Y \sim P_Y$; (284) follows from (14) and the equality $\frac{\beta\rho}{\beta-1} = \frac{\alpha}{\alpha-1}$ (see (282) and (283)); (285) follows from $\frac{\beta}{\beta-1} > 0$, $\frac{1}{\beta} > 0$, Jensen's inequality and the concavity of t^ρ on $[0, \infty)$; (286) follows from (283) and (282); (287) holds due to the concavity of t^θ on $[0, \infty)$ and Jensen's inequality; (288) follows from (282); (289) follows from (14).

Case 2: If $0 < \beta < \alpha < 1$, then $(\rho, \theta) \in (1, \infty)^2$, and both t^ρ and t^θ are convex. However, (285) and (287) continue to hold since now $\frac{\beta}{\beta-1} < 0$.

Case 3: If $\beta < \alpha < 0$, then $\rho \in (0,1)$ and $\theta \in (1, \infty)$. Inequality (285) holds since t^ρ is concave on $[0, \infty)$ and $\frac{\beta}{\beta-1} > 0$; furthermore, although t^θ is now convex on $[0, \infty)$, (287) also holds since $\frac{1}{\beta} < 0$.

APPENDIX B
PROOF OF PROPOSITION 3

Proof of Proposition 3a): We need to show that the lower bound on $H_\alpha(X|Y)$ in (166) coincides with its upper bound in (82) if and only if $\varepsilon_{X|Y} = 0$ or $\varepsilon_{X|Y} = 1 - \frac{1}{M}$. This corresponds, respectively, to the cases where X is a deterministic function of the observation Y or X is equiprobable on the set $\mathcal{X}$ and independent of Y. In view of (35), (82) and (166), and by the use of the parameter $t = \varepsilon_{X|Y} \in \bigl[0, 1 - \frac{1}{M}\bigr]$, this claim can be verified by proving that

- if $\alpha \in (0,1)$ and $k \in \{1, \ldots, M-1\}$, then

$$\Bigl[(1-t)^\alpha + (M-1)^{1-\alpha} t^\alpha\Bigr]^{\frac{1}{\alpha}} > \Bigl[k(k+1)^{\frac{1}{\alpha}} - k^{\frac{1}{\alpha}}(k+1)\Bigr]\, t + k^{\frac{1}{\alpha}+1} - (k-1)(k+1)^{\frac{1}{\alpha}} \tag{290}$$

  for all $t \in [1 - \frac{1}{k}, 1 - \frac{1}{k+1})$ with the point $t = 0$ excluded if $k = 1$;
- the opposite inequality in (290) holds if $\alpha \in (1, \infty)$.

Let the function $v_\alpha \colon \bigl[0, 1 - \frac{1}{M}\bigr] \to \mathbb{R}$ be defined such that for all $k \in \{1, \ldots, M-1\}$

$$v_\alpha(t) = v_{\alpha,k}(t), \quad t \in \bigl[1 - \tfrac{1}{k}, 1 - \tfrac{1}{k+1}\bigr) \tag{291}$$

where $v_{\alpha,k}\colon \big[1-\frac{1}{k}, 1-\frac{1}{k+1}\big) \to \mathbb{R}$ is given by

$$v_{\alpha,k}(t) = \Big[(1-t)^{\alpha} + (M-1)^{1-\alpha} t^{\alpha}\Big]^{\frac{1}{\alpha}} - \Big[k(k+1)^{\frac{1}{\alpha}} - k^{\frac{1}{\alpha}}(k+1)\Big]\, t - k^{\frac{1}{\alpha}+1} + (k-1)(k+1)^{\frac{1}{\alpha}}. \tag{292}$$

Note that $v_{\alpha,k}(t)$ is the difference between the left and right sides of (290). Moreover, let $v_{\alpha}(\cdot)$ be continuously extended at $t = 1-\frac{1}{M}$; it can be verified that $v_{\alpha}(0) = v_{\alpha}\left(1-\frac{1}{M}\right) = 0$. We need to prove that for all $\alpha \in (0,1)$

$$v_{\alpha}(t) > 0, \quad \forall\, t \in \left(0, 1-\tfrac{1}{M}\right), \tag{293}$$

together with the opposite inequality in (293) for $\alpha \in (1,\infty)$. To that end, we prove that

a) for $\alpha \in (0,1) \cup (1,\infty)$, $v_{\alpha}(\cdot)$ is continuous at all points

$$t_k = \tfrac{k}{k+1}, \quad k \in \{0, \ldots, M-1\}. \tag{294}$$

To show continuity, note that the right continuity at $t_0 = 0$ and the left continuity at $t_{M-1} = 1 - \frac{1}{M}$ follow from (291); for $k \in \{1, \ldots, M-2\}$, the left continuity of $v_{\alpha}(\cdot)$ at t_k is demonstrated by showing that

$$\lim_{t \uparrow t_k} v_{\alpha,k}(t) = v_{\alpha,k+1}(t_k), \tag{295}$$

and its right continuity at t_k is trivial from (291), (292) and (294);

b) for all $k \in \{0, 1, \ldots, M-1\}$

$$v_{\alpha}(t_k) = \frac{\Big[1 + (M-1)^{1-\alpha} k^{\alpha}\Big]^{\frac{1}{\alpha}} - (k+1)^{\frac{1}{\alpha}}}{k+1}, \tag{296}$$

which implies that $v_{\alpha}(\cdot)$ is zero at the endpoints of the interval $\left[0, \frac{M-1}{M}\right]$, and if $M \geq 3$

$$v_{\alpha}(t_k) > 0, \quad \alpha \in (0,1), \tag{297}$$

$$v_{\alpha}(t_k) < 0, \quad \alpha \in (1,\infty) \tag{298}$$

for all $k \in \{1, \ldots, M-2\}$;

c) the following convexity results hold:
   - for $\alpha \in (0,1)$, the function $v_{\alpha}(\cdot)$ is strictly concave on $[t_k, t_{k+1}]$ for all $k \in \{0, \ldots, M-1\}$;
   - for $\alpha \in (1,\infty)$, the function $v_{\alpha}(\cdot)$ is strictly convex on $[t_k, t_{k+1}]$ for all $k \in \{0, \ldots, M-1\}$.

   These properties hold since, due to the linearity in t of the right side of (290), $v_{\alpha}(\cdot)$ is convex or concave on $\big[t_k, t_{k+1}\big]$ for all $k \in \{0, 1, \ldots, M-1\}$ if and only if the left side of (290) is, respectively, a convex or concave function in t; due to Lemma 1, the left side of (290) is strictly concave as a function of t if $\alpha \in (0,1)$ and it is strictly convex if $\alpha \in (1,\infty)$.

Due to Items a)–c), (293) holds for $\alpha \in (0,1)$ and the opposite inequality holds for $\alpha \in (1,\infty)$. In order to prove (293) for $\alpha \in (0,1)$, note that $\left[0, 1-\frac{1}{M}\right] = \bigcup_{i=1}^{M-1} [t_{i-1}, t_i]$ where $v_{\alpha}(t_0) = v_{\alpha}(t_{M-1}) = 0$ and $v_{\alpha}(t_k) > 0$ for every $k \in \{1, \ldots, M-2\}$, $v_{\alpha}(\cdot)$ is continuous at all points t_k for $k \in \{0, \ldots, M-1\}$, and v_{α} is strictly concave on $[t_k, t_{k+1}]$ for all $k \in \{0, \ldots, M-2\}$, yielding the positivity of $v_{\alpha}(\cdot)$ on the interval $(0, 1-\frac{1}{M})$. The justification for the opposite inequality of (293) if $\alpha \in (1,\infty)$ is similar, yielding the negativity of $v_{\alpha}(\cdot)$ on $(0, 1-\frac{1}{M})$.

Proof of Proposition 3b): In view of (35), (82) and (174), for all $\varepsilon_{X|Y} \in [0, \frac{1}{2})$ and $\alpha \in (0,1) \cup (1,\infty)$:

$$u_{\alpha,M}(\varepsilon_{X|Y}) = \frac{1}{1-\alpha} \log\Big((1-\varepsilon_{X|Y})^{\alpha} + (M-1)^{1-\alpha} \varepsilon_{X|Y}^{\alpha}\Big), \tag{299}$$

$$l_{\alpha}(\varepsilon_{X|Y}) = \frac{\alpha}{1-\alpha} \log\Big(1 + (2^{\frac{1}{\alpha}} - 2)\varepsilon_{X|Y}\Big). \tag{300}$$

Equality (179) follows by using L'Hôpital's rule for the calculation of the limit in the left side of (179).

ACKNOWLEDGMENT

The authors gratefully acknowledge the detailed reading of the manuscript by the Associate Editor and anonymous referees.

REFERENCES

[1] R. Ahlswede, P. Gács and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, no. 2, pp. 157–177, 1976 (correction in vol. 39, no. 4, pp. 353–354, 1977).

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. on Information Theory*, vol. 42, no. 1, pp. 99–105, January 1996.

[3] S. Arimoto, "Information-theoretical considerations on estimation problems," *Information and Control*, vol. 19, no. 3, pp. 181–194, October 1971.

[4] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Information Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[5] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," in *Topics in Information Theory - 2nd Colloquium*, Keszthely, Hungary, 1975, Colloquia Mathematica Societatis Janós Bolyai (I. Csiszár and P. Elias editors), Amsterdam, Netherlands: North Holland, vol. 16, pp. 41–52, 1977.

[6] L. Baladová, "Minimum of average conditional entropy for given mimimum probability of error," *Kybernetika*, vol. 2, no. 5, pp. 416–422, 1966.

[7] M. Ben-Bassat and J. Raviv, "Rényi's entropy and probability of error," *IEEE Trans. on Information Theory*, vol. 24, no. 3, pp. 324–331, May 1978.

[8] S. Berens, *Conditional Rényi Entropy*, Master's thesis, Department of Mathematics, Leiden university, Leiden, Netherlands, August 2013. [Online]. Available at https://www.math.leidenuniv.nl/scripties/MasterBerens.pdf.

[9] L. Birgé, "A new lower bound for multiple hypothesis testing," *IEEE Trans. Information Theory*, vol. 51, no. 4, pp. 1611–1615, April 2005.

[10] A. Bracher, E. Hof and A. Lapidoth, "Guessing attacks on distributed-storage systems," *Proceedings of the 2015 IEEE International Symposium on Information Theory*, pp. 1585–1589, Hong-Kong, China, June 2015.

[11] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Trans. on Information Theory*, vol. 60, no. 9, pp. 5065–5076, September 2014.

[12] C. Bunte and A. Lapidoth, "On the listsize capacity with feedback," *IEEE Trans. on Information Theory*, vol. 60, no. 11, pp. 6733–6748, November 2014.

[13] L. L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, pp. 423–429, 1965.

[14] P. N. Chen and F. Alajaji, "A generalized Poor-Verdú error bound for multihypothesis testing," *IEEE Trans. on Information Theory*, vol. 58, no. 1, pp. 311–316, January 2012.

[15] J. Chu and J. Chueh, "Inequalities between information measures and error probability," *Journal of the Franklin Institute*, vol. 282, no. 2, pp. 121–125, August 1966.

[16] T. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," *Proceedings of the 2014 IEEE International Symposium on Information Theory*, pp. 2494–2498, Honolulu, Hawaii, USA, July 2014.

[17] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 21–27, January 1967.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, Second Edition, 2006.

[19] I. Csiszár, "Generalized cutoff rates and Rényi information measures," *IEEE Trans. on Information Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

[20] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Second Edition, Cambridge University Press, 2011.

[21] P. A. Devijver, "On a new class of bounds on Bayes risk in multihypothesis pattern recognition," *IEEE Trans. on Computers*, vol. C-23, no. 1, pp. 70–80, January 1974.

[22] D. Erdogmus and J. Principe, "Lower and upper bounds for misclasification probability based on Rényi's information," *Journal of VLSI Signal Processing*, vol. 37, no. 2, pp. 305–317, June 2004.

[23] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[24] R. M. Fano, *Class Notes for Course 6.574: Transmission of Information*, MIT, Cambridge, MA, USA, 1952.

[25] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Trans. on Information Theory*, vol. 40, no. 1, pp. 259–266, January 1994.

[26] A. Feinstein, “A new basic theorem of information theory,” *IEEE Trans. on Information Theory*, vol. 3, no. 4, pp. 2–22, September 1954.

[27] S. Fehr and S. Berens, “On the conditional Rényi entropy,” *IEEE Trans. on Information Theory*, vol. 60, no. 11, pp. 6801–6810, November 2014.

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

[29] R. G. Gallager, “A simple derivation of the coding theorem and some applications,” *IEEE Trans. on Information Theory*, vol. 11, no. 1, pp. 3–18, January 1965.

[30] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley, 1968.

[31] R. G. Gallager, “The random coding bound is tight for the average code,” *IEEE Trans. on Information Theory*, vol. 19, no. 2, pp. 244–246, March 1973.

[32] A. Guntuboyina, “Lower bounds for the minimax risk using f-divergences, and applications,” *IEEE Trans. on Information Theory*, vol. 57, no. 4, pp. 2386–2399, April 2011.

[33] N. Grigoryan, A. Harutyunyan, S. Voloshynovskiy and O. Koval, “On multiple hypothesis testing with rejection option,” *Proceedings of the 2011 IEEE Information Theory Workshop*, pp. 75–79, Paraty, Brazil, October 2011.

[34] T. S. Han and S. Verdú, “Generalizing the Fano inequality,” *IEEE Trans. on Information Theory*, vol. 40, no. 4, pp. 1247–1251, July 1994.

[35] M. Hayashi and V. Y. F. Tan, “Equivocation and exponents under various Rényi information measures,” *IEEE Trans. on Information Theory*, vol. 63, no. 2, pp. 975–1005, February 2017.

[36] M. Hayashi and V. Y. F. Tan, “Remaining uncertainties and exponents under Rényi information measures,” *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pp. 1536–1540, Barcelona, Spain, July 2016.

[37] M. Hayashi and M. Tomamichel, “Correlation detection and an operational interpretation of the Rényi mutual information,” *Proceedings of the 2015 IEEE International Symposium on Information Theory*, pp. 1447–1451, Hong Kong, June 14–19, 2015.

[38] M. E. Hellman and J. Raviv, “Probability of error, equivocation, and the Chernoff bound,” *IEEE Trans. on Information Theory*, vol. 16, no. 4, pp. 368–372, July 1970.

[39] S. W. Ho and S. Verdú, “On the interplay between conditional entropy and error probability,” *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 5930–5942, December 2010.

[40] S. W. Ho and S. Verdú, “Convexity/concavity of the Rényi entropy and α-mutual information,” *Proceedings of the 2015 IEEE International Symposium on Information Theory*, pp. 745–749, Hong Kong, China, June 2015.

[41] S. W. Ho and R. W. Yeung, “The interplay between entropy and variational distance,” *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 5906–5929, December 2010.

[42] E. Hof, I. Sason and S. Shamai (Shitz), “Performance bounds for erasure, list and feedback schemes with linear block codes,” *IEEE Trans. on Information Theory*, vol. 56, no. 8, pp. 3754–3778, August 2010.

[43] M. Iwamoto and J. Shikata, “Information theoretic security for encryption based on conditional Rényi entropies,” *Proceedings of the 7th International Conference on Information Theoretic Security*, Singapore, Nov. 28-30, 2013. See *Lecture Notes in Computer Science*, vol. 8317, pp. 103–121, Springer 2014.

[44] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *IEEE Trans. on Communication Technology*, vol. 15, no. 1, pp. 52–60, February 1967.

[45] F. Kanaya and T. S. Han, “The asymptotics of posterior entropy and error probability for Bayesian estimation,” *IEEE Trans. on Information Theory*, vol. 41, no. 6, pp. 1988–1992, November 1995.

[46] Y. H. Kim, A. Sutivong and T. M. Cover, “State amplification,” *IEEE Trans. on Information Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.

[47] R. König, R. Renner, and C. Schaffner, “The operational meaning of min- and max-entropy,” *IEEE Trans. on Information Theory*, vol. 55, no. 9, pp. 4337–4347, September 2009.

[48] V. A. Kovalevsky, “The problem of character recognition from the point of view of mathematical statistics,” in *Reading Automata and Pattern Recognition* (in Russian), Naukova Dumka, Kiev, ed. 1965. English translation in *Character Readers and Pattern Recognition*, Spartan Books, pp. 3–30, New York, 1968.

[49] A. Lalitha and T. Javidi, “Reliability of sequential hypothesis testing can be achieved by an almost-fixed-length-test,” *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pp. 1710–1714, Barcelona, Spain, July 2016.

[50] A. Lapidoth and C. Pfister, “Two measures of dependence,” *Proceedings of the 2016 IEEE International Conference on the Science of Electrical Engineering*, pp. 1–5, Eilat, Israel, November 2016.

[51] C. C. Leang and D. H. Johnson, “On the asymptotics of M-hypothesis Bayesian detection,” *IEEE Trans. on Information Theory*, vol. 43, no. 1, pp. 280–282, January 1997.

[52] F. Leditzky, M. M. Wilde and N. Datta, “Strong converse theorems using Rényi entropies,” *Journal of Mathematical Physics*, vol. 57, paper no. 082202, pp. 1–33, August 2016.

[53] N. Merhav, "List decoding – random coding exponents and expurgated exponents," *IEEE Trans. on Information Theory*, vol. 60, no. 11, pp. 6749-6759, November 2014.

[54] M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr and M. Tomamichel, "On quantum Rényi entropies: a new definition and some properties," *Journal of Mathematical Physics*, vol. 54, paper no. 122203, pp. 1–20, December 2013.

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

[56] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," *Proceedings of the Forty-Eighth Annual Allerton Conference on Communication, Control and Computing*, pp. 1327–1333, Monticello, Illinois, USA, October 2010.

[57] H. V. Poor and S. Verdú, "A lower bound on the probability of error in multihypothesis testing," *IEEE Trans. on Information Theory*, vol. 41, no. 6, pp. 1992–1994, November 1995.

[58] S. Prasad, "Bayesian error-based sequences of statistical information bounds," *IEEE Trans. on Information Theory*, vol. 61, no. 9, pp. 5052–5062, September 2015.

[59] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications and coding," Second Edition, *Foundations and Trends in Communications and Information Theory*, NOW Publishers, vol. 10, no. 1–2, pp. 1–250, October 2014.

[60] J. M. Renes, "Better bounds on optimal measurement and entanglement recovery, with applications to uncertainty and monogamy relations," July 2017. [Online]. Available at https://arxiv.org/abs/1707.01114.

[61] R. Renner, *Security of Quantum Key Distribution*, PhD dissertation, ETH Zurich, Switzerland, December 2005.

[62] A. Rényi, "On measures of entropy and information," *Proceedings of the Fourth Berkeley Symposium on Probability Theory and Mathematical Statistics*, pp. 547–561, Berkeley, California, USA, 1961.

[63] R. M. Roth, *Introduction to Coding Theory*, Cambridge University Press, Cambridge, UK, 2006.

[64] T. Routtenberg and J. Tabrikian, "General class of lower bounds on the probability of error in multiple hypothesis testing," *Proceedings of the 25th IEEE Convention of Electrical and Electronics Engineers in Israel*, pp. 750–754, Eilat, Israel, December 2008.

[65] Y. Sakai and K. Iwata, "Relations between conditional Shannon entropy and expectation of l_α-norm," *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pp. 1641–1645, Barcelona, Spain, July 2016.

[66] I. Sason and S. Shamai (Shitz), "Performance analysis of linear codes under maximum-likelihood decoding: A tutorial," *Foundations and Trends in Communications and Information Theory*, NOW Publishers, vol. 3, no. 1–2, pp. 1–222, July 2006.

[67] I. Sason, "Moderate deviations analysis of binary hypothesis testing," *Proceedings of the 2012 IEEE International Symposium on Information Theory*, pp. 826–830, MIT, Cambridge, MA, USA, July 2012.

[68] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. on Information Theory*, vol. 62, no. 11, pp. 5973–6006, November 2016.

[69] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July-October 1948.

[70] C. E. Shannon, "Certain results in coding theory for noisy channels," *Information and Control*, vol. 1, pp. 6–25, September 1957.

[71] C. E. Shannon, "Channels with side information at the transmitter," *IBM Journal on Research and Development*, vol. 2, no. 4, pp. 289–293, October 1958.

[72] C. Shannon, R. Gallager and E. Berlekamp, "Lower bounds to error probability for decoding on discrete memoryless channels - Part I," *Information and Control*, vol. 10, pp. 65–103, February 1967.

[73] O. Shayevitz, "On Rényi measures and hypothesis testing," *Proceedings of the 2011 IEEE International Symposium on Information Theory*, pp. 800–804, Saint Petersburg, Russia, August 2011.

[74] R. Sibson, "Information radius," *Probability Theory and Related Fields*, vol. 14, no. 2, pp. 149–161, June 1969.

[75] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. on Information Theory*, vol. 53, no. 1, pp. 269–287, January 2007.

[76] D. Tebbe and S. Dwyer, "Uncertainty and the probability of error," *IEEE Trans. on Information Theory*, vol. 14, no. 3, pp. 516–518, May 1968.

[77] M. Tomamichel, M. Berta and M. Hayashi, "Relating different quantum generalizations of the conditional Rényi entropy," *Journal of Mathematical Physics*, vol. 55, no. 8, paper no. 082206, pp. 1–10, 2014.

[78] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi conditional mutual information via composite hypothesis testing against Markov distributions," *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pp. 585–589, Barcelona, Spain, July 2016.

[79] G. T. Toussiant, "Comments on the divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. on Communications*, vol. 20, no. 3, p. 485, July 1972.

[80] G. T. Toussiant, "A generalization of Shannon's equivocation and the Fano bound," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 7, no. 4, pp. 300–302, April 1977.

[81] I. Vajda, “Bounds on the minimal error probability on checking a finite or countable number of hypotheses,” *Problems of Information Transmission*, vol. 4, no. 1, pp. 6–14, 1968.

[82] G. Vázquez-Vilar, A. Tauste-Campo, A. Guillén-Fábregas, and A. Martinez, “Bayesian M-ary hypothesis testing: The meta-converse and Verdú–Han bounds are tight,” *IEEE Trans. on Information Theory*, vol. 62, no. 5, pp. 2324–2333, May 2016.

[83] S. Verdú, “Shannon’s inequality,” *2011 Workshop on Information Theory and Applications*, San Diego, California, USA, February 2011.

[84] S. Verdú, *Information Theory*, in preparation.

[85] S. Verdú, “α-mutual information,” *Workshop on Information Theory and Applications*, San-Diego, California, USA, February 2015. [Online]. Available at http://ita.ucsd.edu/workshop/15/files/paper/paper_374.pdf.

[86] A. M. Wilde, A. Winter, and D. Yang, “Strong converse for the classical capacity of entanglement-breaking and Hadamard channels via a sandwiched Rényi relative entropy,” *Communications in Mathematical Physics*, vol. 331, no. 2, pp. 559–622, July 2014.