Improved Bounds on Lossless Source Coding and Guessing Moments via Rényi Measures

Igal Sason Sergio Verdú

Abstract

This paper provides upper and lower bounds on the optimal guessing moments of a random variable taking values on a finite set when side information may be available. These moments quantify the number of guesses required for correctly identifying the unknown object and, similarly to Arikan's bounds, they are expressed in terms of the Arimoto-Rényi conditional entropy. Although Arikan's bounds are asymptotically tight, the improvement of the bounds in this paper is significant in the non-asymptotic regime. Relationships between moments of the optimal guessing function and the MAP error probability are also established, characterizing the exact locus of their attainable values. The bounds on optimal guessing moments serve to improve non-asymptotic bounds on the cumulant generating function of the codeword lengths for fixed-to-variable optimal lossless source coding without prefix constraints. Non-asymptotic bounds on the reliability function of discrete memoryless sources are derived as well. Relying on these techniques, lower bounds on the cumulant generating function of the codeword lengths are derived, by means of the smooth Rényi entropy, for source codes that allow decoding errors.

Keywords

Cumulant generating function, guessing moments, lossless source coding, M-ary hypothesis testing, Rényi entropy, Rényi divergence, Arimoto-Rényi conditional entropy, smooth Rényi entropy.

I. Introduction

A. Prior work

The problem of guessing discrete random variables has found a variety of applications in information theory, coding theory, cryptography, and searching and sorting algorithms. The central object of interest is the distribution of the number of guesses required to identify a realization of a random variable X, taking values on a finite or countably infinite set $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$, by asking questions of the form "Is X equal to x?". A *guessing function* is a one-to-one function $g \colon \mathcal{X} \to \mathcal{X}$, which can be viewed as a permutation of the elements of $\mathcal{X}$ in the order in which they are guessed. We can envision a generic algorithm that outputs $g^{-1}(1)$; a supervisor checks whether $X = g^{-1}(1)$, if so then the algorithm halts; otherwise, the algorithm outputs $g^{-1}(2)$ and the process repeats until the value of X is guessed correctly. Therefore, the number of guesses is $g(x)$ when the true outcome is $x \in \mathcal{X}$.

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 is with the Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA (e-mail: verdu@princeton.edu).

This manuscript has been submitted to the *IEEE Transactions on Information Theory* in November 1, 2017, and accepted for publication in January 28, 2018. It is published in the *IEEE Trans. on Information Theory*, vol. 64, no. 6, pp. 4323–4346, June 2018. This is a post-print which includes few corrections of printing typos.

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

Parts of this work have been presented at the 2018 International Zurich Seminar on Information and Communication, Zurich, Switzerland, February 21-23, 2018, and in the 2018 IEEE International Symposium on Information Theory, Vail, Colorado, USA, June 17–22, 2018.

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.

Lower and upper bounds on the minimal expected number of required guesses for correctly identifying the realization of X, expressed as a function of the Shannon entropy $H(X)$, have been respectively derived by Massey [31] and by McEliece and Yu [32]. More generally, given a probability mass function P_X on $\mathcal{X}$, it is of interest to minimize the generalized guessing moment

$$\mathbb{E}[g^\rho(X)] = \sum_{x \in \mathcal{X}} P_X(x) g^\rho(x), \quad \rho > 0. \tag{1}$$

For an arbitrary positive ρ, the ρ-th moment of the number of guesses is minimized by selecting the guessing function to be a *ranking function* g_X, for which $g_X(x) = k$ if $P_X(x)$ is the k-th largest mass. Upper and lower bounds on the ρ-th moment of ranking functions, expressed in terms of the Rényi entropy $H_\alpha(X)$ of order $\alpha = \frac{1}{1+\rho}$, were derived by Arikan [1], followed by a refined upper bound by Boztaş [6]. Although if $|\mathcal{X}|$ is small, it is straightforward to evaluate numerically the guessing moments, the benefit of bounds expressed in terms of Rényi entropies is particularly relevant when dealing with a random vector $X^n = (X_1, \ldots, X_n)$ whose letters belong to a finite alphabet $\mathcal{A}$; computing all the probabilities of the mass function P_{X^n} over the set $\mathcal{A}^n$, and then sorting them in decreasing order for the calculation of the ρ-th moment of the optimal guessing function for the elements of $\mathcal{A}^n$ has exponential complexity in n. Therefore, it becomes infeasible even for moderate values of n. In contrast, regardless of the value of n, bounds on guessing moments which depend on the Rényi entropy are readily computable if for example $\{X_i\}_{i=1}^n$ are independent; in which case, the Rényi entropy of the vector is equal to the sum of the Rényi entropies of its components (hence, the exponential complexity is reduced to linear complexity in n; furthermore, in the i.i.d. case, the complexity in calculating the Rényi entropy of X^n is independent of n). Arikan's bounds are asymptotically tight for random vectors of length n as $n \to \infty$, so another benefit of these bounds is that they provide the correct exponential growth rate of the guessing moments for sufficiently large n. In [1], Arikan generalized his bounds to allow side information, leading to asymptotically tight bounds which are expressed in terms of the Arimoto-Rényi conditional entropy [4].

The guessing problem has been studied in the information-theoretic literature in various contexts, which include: guessing subject to distortion [2], joint source-channel coding and guessing with application to sequential decoding [3], guessing with a prior access to a malicious oracle [11], a large deviations approach to guessing and source compression ([15], [22], [47]), guessing with limited memory [23], guesswork exponents for Markov sources [30], guessing in secrecy problems ([40], [51]), and guessing under source uncertainty [46].

For uniquely-decodable lossless source coding, Campbell ([13], [14]) proposed the normalized cumulant generating function of the codeword lengths as a generalization to the frequently used design criterion of normalized average code length. Campbell's motivation in [13] was to control the contribution of the longer codewords via a free parameter in the cumulant generating function: if the value of this parameter tends to zero, then the resulting design criterion becomes the normalized average code length while by increasing the value of the free parameter, the penalty for longer codewords is more severe, and the resulting code optimization yields a reduction in the fluctuations of the codeword lengths. In [13], Campbell obtained asymptotically tight upper and lower bounds on the minimum normalized cumulant generating function for discrete memoryless stationary sources with finite alphabet. These bounds, expressed in terms of the Rényi entropy, imply that for sufficiently long source sequences, it is possible to make the normalized cumulant generating function of the codeword lengths approach the Rényi entropy as closely as desired by a proper fixed-to-variable uniquely-decodable source code; moreover, a converse result in [13] shows that there is no uniquely-decodable source code for which the normalized cumulant generating function of its codeword lengths lies below the Rényi entropy. In addition, this type of bounds was studied in the context of various coding problems, including guessing (see, e.g., [1], [2], [3], [7], [8], [9], [16], [17], [22], [28], [33], [34], [35], [46], [50]).

Kontoyiannis and Verdú [26] studied the behavior of the best achievable rate and other fundamental limits in variable-rate lossless source compression without prefix constraints. In the non-asymptotic regime, the fundamental limits of fixed-to-variable lossless compression with and without prefix constraints were shown to be tightly coupled. Reference [26] obtains non-asymptotic upper and lower bounds on the distribution of codeword lengths, along with a rigorous proof of the Gaussian approximation put forward in 1962 by Strassen [45] for memoryless sources. An alternative approach was followed by Courtade and Verdú in [16], where they derived non-asymptotic bounds for the normalized cumulant generating function of the codeword lengths for optimal variable-length lossless codes without prefix constraints; these bounds are used in [16] to obtain simple proofs of the asymptotic normality and the reliability function of memoryless sources allowing countably infinite alphabets.

In [27], Kostina *et al.* studied the fundamental limits of the minimum average length of lossless and lossy variable-length compression, allowing a nonzero error probability $\varepsilon \in [0, 1)$ for almost lossless compression. The bounds in [27] were used to obtain a Gaussian approximation on the speed of convergence of the minimum average length, which was shown to be quite accurate for all but small blocklengths. In [24], Koga and Yamamoto followed an information-spectrum approach to obtain asymptotic properties of the codeword lengths for prefix fixed-to-variable source codes, allowing decoding errors. This work was refined in the non-asymptotic setting by Kuzuoka [28], which bounds the cumulant generating function of the codeword lengths via the smooth Rényi entropy.

B. Organization of the paper

Section II defines the Rényi information measures, and summarizes those properties used in this paper. Section III derives upper and lower bounds on the minimal guessing moments of a random variable taking a finite number of values where side information on its value may be available. In the non-asymptotic regime, these bounds improve earlier results by Arikan [1] and Boztaş [6]. Section III also derives tight lower and upper bounds which establish relationships between the MAP error probability in M-ary hypothesis testing, and the moments of the optimal guessing function for correctly identifying X when side information Y is available. The bounds on the guessing moments are applied in Section IV to optimal variable-length lossless data compression. We derive improved bounds on the normalized cumulant generating function of the codeword lengths for fixed-to-variable optimal codes, and on the non-asymptotic reliability function of discrete memoryless sources, tightening earlier results by Courtade and Verdú [16]. Following up the aforementioned work by Kuzuoka [28], Section V relies on the techniques in Sections III and IV in order to derive improved lower bounds on the cumulant generating function of the codeword lengths for fixed-to-variable source coding allowing errors via the use of the smooth Rényi entropy ([12], [25], [36], [37]). The bounds in Section V are derived for source codes allowing a given maximal or average error probability.

II. Preliminaries

The information measures used in this paper apply to discrete random variables.

Definition 1: [38] Let X be a discrete random variable taking values on a finite or countably infinite set $\mathcal{X}$, and let P_X be its probability mass function. The Rényi entropy of order $\alpha \in (0,1) \cup (1,\infty)$ is given by[1]

$$H_\alpha(X) = H_\alpha(P_X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X^\alpha(x). \tag{2}$$

By its continuous extension,

$$H_0(X) = \log \bigl|\{x \in \mathcal{X} \colon P_X(x) > 0\}\bigr|, \tag{3}$$
$$H_1(X) = H(X), \tag{4}$$
$$H_\infty(X) = \log \frac{1}{p_{\max}} \tag{5}$$

where $H(X)$ is the entropy of X, and

$$p_{\max} = \max_{x \in \mathcal{X}} P_X(x). \tag{6}$$

All definitions in this section extend in a natural way to random vectors.

Lemma 1: [38] Let $X^n = (X_1, \ldots, X_n)$ be an n-dimensional random vector with independent components. Then, for all $\alpha \in [0,\infty]$,

$$H_\alpha(X^n) = \sum_{i=1}^n H_\alpha(X_i). \tag{7}$$

Note however that in contrast to the Shannon entropy, $H_\alpha(X^n)$ may exceed the right side of (7) when $\alpha \neq 1$ and $\{X_i\}$ are dependent.

Definition 2: [38] Given probability mass functions P and Q on a finite or countably infinite set $\mathcal{X}$, the *Rényi divergence of order* $\alpha > 0$ is defined as follows:

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

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \sum_{x \in \mathcal{X}} P^\alpha(x)\, Q^{1-\alpha}(x). \tag{8}$$

- By the continuous extension of $D_\alpha(P\|Q)$, the Rényi divergences of orders 0, 1, and ∞ are defined as

$$D_0(P\|Q) = -\log Q\bigl(\{x \in \mathcal{X} \colon P(x) > 0\}\bigr), \tag{9}$$
$$D_1(P\|Q) = D(P\|Q), \tag{10}$$
$$D_\infty(P\|Q) = \log \sup_{x \in \mathcal{X}} \frac{P(x)}{Q(x)}, \tag{11}$$

where $D(P\|Q)$ denotes the relative entropy.

Properties of the Rényi divergence were studied in [5], [19], [41, Section 8] and [44]. The Rényi divergence of negative orders is defined by extending (8) to $\alpha \in (-\infty, 0)$ [19, Section 5].

Lemma 2: [19, Section 5] The following properties are satisfied by the Rényi divergence:

- For all $\alpha \neq 0$, $D_\alpha(P\|Q) = 0$ if and only if $P = Q$.
- $D_\alpha(P\|Q) \in [0,\infty]$ for $\alpha \in [0,\infty]$, and $D_\alpha(P\|Q) \in [-\infty, 0]$ for $\alpha \in [-\infty, 0)$ (with the continuous extension where $D_{-\infty}(P\|Q) \triangleq \lim_{\alpha \to -\infty} D_\alpha(P\|Q)$ [19, (81)]).

[1]Unless explicitly stated, the logarithm base can be chosen by the reader, with exp indicating the inverse function of log.

Definition 3: 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

$$d_\alpha(p\|q) = \frac{1}{\alpha-1}\ \log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Big). \tag{12}$$

Definition 4: [4] 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

$$H_\alpha(X|Y) = \frac{\alpha}{1-\alpha}\ \log\ \mathbb{E}\left[\left(\sum_{x\in\mathcal{X}} P^\alpha_{X|Y}(x|Y)\right)^{\frac{1}{\alpha}}\right] \tag{13}$$

$$= \frac{\alpha}{1-\alpha}\ \log \sum_{y\in\mathcal{Y}} P_Y(y)\ \exp\left(\frac{1-\alpha}{\alpha}\ H_\alpha(X|Y=y)\right), \tag{14}$$

  where (14) 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

$$H_0(X|Y) = \sup_{y\in\mathcal{Y}} H_0(X\,|\,Y=y), \tag{15}$$

$$H_1(X|Y) = H(X|Y), \tag{16}$$

$$H_\infty(X|Y) = \log \frac{1}{\mathbb{E}\Big[\max_{x\in\mathcal{X}} P_{X|Y}(x|Y)\Big]}. \tag{17}$$

Properties of the Arimoto-Rényi conditional entropy were studied in [20], [39] and [42].

As in [42, Section 4], we find several useful results satisfied by the Arimoto-Rényi conditional entropy of negative orders.

Another Rényi information measure used in this paper is the smooth Rényi entropy, introduced by Renner and Wolf [37] (after a different definition in [36]).

Definition 5: [37] Let X be a discrete random variable taking values on $\mathcal{X}$, and let P_X denote the probability mass function of X. Let $\alpha \in (0,1) \cup (1,\infty)$ and $\varepsilon \in [0,1)$. The *ε-smooth Rényi entropy of order* α is given by

$$H^{(\varepsilon)}_\alpha(X) = \frac{1}{1-\alpha}\ \min_{\mu \in B^{(\varepsilon)}(P_X)} \log \sum_{x\in\mathcal{X}} \mu^\alpha(x) \tag{18}$$

where $B^{(\varepsilon)}(P_X)$ is the following subset of sub-probability measures defined on $\mathcal{X}$:

$$B^{(\varepsilon)}(P_X) \triangleq \left\{\mu\colon \mathcal{X} \to [0,1] : \sum_{x\in\mathcal{X}} \mu(x) \geq 1-\varepsilon, \quad \mu(x) \leq P_X(x),\ \forall x \in \mathcal{X}\right\}. \tag{19}$$

The ε-smooth Rényi entropy becomes the Rényi entropy when $\varepsilon = 0$, i.e.,

$$H^{(0)}_\alpha(X) = H_\alpha(X) \tag{20}$$

for all $\alpha \in (0,1) \cup (1,\infty)$.
Properties of $H^{(\varepsilon)}_\alpha(X)$ were studied in [25] and [37].

Lemma 3: [25, Theorem 1] Let X be a random variable taking values on a finite set $\mathcal{X} = \{x_1, \ldots, x_M\}$, whose elements are ordered such that

$$P_X(x_1) \geq P_X(x_2) \geq \ldots \geq P_X(x_M), \tag{21}$$

and let $\varepsilon \in [0,1)$. Then,

a) For $\alpha \in (0,1)$, the minimum in the right side of (18) is achieved by $\mu_1 \in B^{(\varepsilon)}(P_X)$ given by

$$\mu_1(x_i) = \begin{cases} P_X(x_i), & i = 1, \ldots, J_\varepsilon - 1 \\ 1 - \varepsilon - \sum_{j=1}^{J_\varepsilon - 1} P_X(x_j), & i = J_\varepsilon \\ 0, & i = J_\varepsilon + 1, \ldots, M \end{cases} \tag{22}$$

with

$$J_\varepsilon = \min \left\{ 1 \leq j \leq M \colon \sum_{i=1}^{j} P_X(x_i) \geq 1 - \varepsilon \right\}. \tag{23}$$

b) For $\alpha > 1$, the minimum in the right side of (18) is achieved by $\mu_2 \in B^{(\varepsilon)}(P_X)$ given by

$$\mu_2(x_i) = \begin{cases} \beta, & i = 1, \ldots, K_\beta \\ P_X(x_i), & i = K_\beta + 1, \ldots, M \end{cases} \tag{24}$$

where $\beta \in (0,1)$ and $K_\beta \in \{1, \ldots, M\}$ are jointly selected such that

$$\sum_{i=1}^{M} \mu_2(x_i) = 1 - \varepsilon, \tag{25}$$

$$K_\beta = \max\big\{1 \leq j \leq M \colon P_X(x_j) \geq \beta\big\}. \tag{26}$$

Remark 1: The sub-probability measures $\mu_1, \mu_2 \in B^{(\varepsilon)}(P_X)$ in Lemma 3 are *independent* of α for $\alpha \in (0,1)$ or $\alpha > 1$, respectively.

Lemma 4: Under the assumption in Lemma 3, the following inequalities hold for $\varepsilon \in (0,1)$:

a) If $\alpha \in (0,1)$

$$\frac{1}{\alpha - 1} \log \frac{1}{1 - \varepsilon} \leq H_\alpha^{(\varepsilon)}(X) \leq \frac{1}{\alpha - 1} \log \left(\frac{1}{1 - \varepsilon} \right) + \log \left(\frac{1}{\mu_1(x_{J_\varepsilon})} \right) \tag{27}$$

with J_ε as defined in (23).

b) If $\alpha > 1$

$$\frac{1}{\alpha - 1} \log \frac{1}{1 - \varepsilon} \leq H_\alpha^{(\varepsilon)}(X) \leq \frac{1}{\alpha - 1} \log \left(\frac{1}{1 - \varepsilon} \right) + \log \left(\frac{1}{\min\{P_{\min}, \beta\}} \right) \tag{28}$$

where $P_{\min}$ denotes the minimal non-zero mass of P_X, and β is defined in (24)–(26).

Proof: From (22) and (24), it follows that for all $x \in \mathcal{X}$

$$\mu_1(x) \leq \mu_1^\alpha(x) \leq \mu_1(x)\, \mu_1^{\alpha-1}(x_{J_\varepsilon}), \qquad \alpha \in (0,1), \tag{29}$$

$$\mu_2(x) \min\{P_{\min}^{\alpha-1}, \beta^{\alpha-1}\} \leq \mu_2^\alpha(x) \leq \mu_2(x), \quad \alpha > 1. \tag{30}$$

The bounds in (27) and (28) can be verified from Definition 5, (22), (25), (29) and (30). ∎

Remark 2: The left inequality in (27) can be obtained from [37, Lemma 2].

The following result readily follows from Lemma 4.

Corollary 1: If X takes values on a finite set, then for all $\varepsilon \in (0,1)$,

$$\lim_{\alpha \uparrow 1} H_\alpha^{(\varepsilon)}(X) = -\infty, \tag{31}$$

$$\lim_{\alpha \downarrow 1} H_\alpha^{(\varepsilon)}(X) = +\infty. \tag{32}$$

Remark 3: In contrast to $H_\alpha(X)$ which is non-negative, continuous and monotonically non-increasing in α when X takes values on a finite set, Corollary 1 shows that these properties do not fully extend to $H_\alpha^{(\varepsilon)}(X)$ with $\varepsilon \in (0,1)$. Bearing in mind the discontinuity shown in Corollary 1, $H_\alpha^{(\varepsilon)}(X)$ is monotonically non-increasing on both $\alpha \in (0,1)$ and $\alpha \in (1,\infty)$ [37, Lemma 1].

III. Improved Bounds on Guessing Moments

This section provides improved upper and lower bounds on the guessing moments of a discrete random variable. The upper bounds in this section correspond to the case where the guessing function is a ranking function.

A. Key result

Theorem 1: Given a discrete random variable X taking values on a set $\mathcal{X}$, a function $g\colon \mathcal{X} \to (0,\infty)$, and a scalar $\rho \neq 0$, then

1)

$$\begin{aligned} &\sup_{\beta \in (-\rho,+\infty)\setminus\{0\}} \frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X) - \log \sum_{x\in\mathcal{X}} g^{-\beta}(x)\right] \\ &\leq \frac{1}{\rho}\log \mathbb{E}[g^\rho(X)] \qquad (33)\\ &\leq \inf_{\beta \in (-\infty,-\rho)\setminus\{0\}} \frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X) - \log \sum_{x\in\mathcal{X}} g^{-\beta}(x)\right]. \qquad (34) \end{aligned}$$

2) For $\tau \in \mathbb{R}$, define the probability mass function

$$Q_\tau(x) = \frac{g^{-\tau}(x)}{\sum\limits_{a\in\mathcal{X}} g^{-\tau}(a)}, \quad x \in \mathcal{X}, \qquad (35)$$

provided that the sum in the right side of (35) is finite. The following results hold:

a) If $P_X = Q_\rho$ and $\mathcal{X}$ is a finite set, then

$$\frac{1}{\rho}\log \mathbb{E}[g^\rho(X)] = -\frac{1}{\rho}\log\left(\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} g^{-\rho}(x)\right). \qquad (36)$$

b) If $P_X = Q_\nu$ with $\nu > 0$ and $\nu \neq \rho$, then the supremum in the left side of (33) is attained at $\beta = \nu - \rho$, and the inequality in (33) becomes an identity. Conversely, if (33) is an identity and the supremum is attained at $\beta^* \in (-\rho,+\infty)\setminus\{0\}$, then $P_X = Q_{\rho+\beta^*}$.

c) If $P_X = Q_\nu$ with $\nu < 0$ and $\nu \neq \rho$, then the infimum in the right side of (34) is attained at $\beta = \nu - \rho$, and the inequality in (34) becomes an identity. Conversely, if (34) is an identity and the infimum is attained at $\beta^* \in (-\infty,-\rho)\setminus\{0\}$, then $P_X = Q_{\rho+\beta^*}$.

Proof: It is instructive to prove a weaker result first where instead of optimizing with respect to β, we simply take $\beta = 1$ in the upper/lower bound.

Let $\alpha \neq 0$, and let R_α be the scaled probability mass function defined by

$$R_\alpha(x) = \frac{P_X^{\frac{1}{\alpha}}(x)}{\sum\limits_{a\in\mathcal{X}} P_X^{\frac{1}{\alpha}}(a)}, \quad x \in \mathcal{X}. \qquad (37)$$

Then,

$$D_{1+\rho}(R_{1+\rho}\|Q_1)=\frac{1}{\rho}\,\log\left(\sum_{x\in\mathcal{X}}P_X(x)g^{\rho}(x)\right)+\log\left(\sum_{x\in\mathcal{X}}\frac{1}{g(x)}\right)-\frac{1+\rho}{\rho}\,\log\left(\sum_{x\in\mathcal{X}}P_X^{\frac{1}{1+\rho}}(x)\right) \tag{38}$$

$$=\frac{1}{\rho}\log\mathbb{E}[g^{\rho}(X)]+\log\left(\sum_{x\in\mathcal{X}}\frac{1}{g(x)}\right)-H_{\frac{1}{1+\rho}}(X) \tag{39}$$

where (38) follows from (8), (35) and (37); (39) is due to (2). Since $D_{1+\rho}(R_{1+\rho}\|Q_1)\geq 0$ for $\rho>-1$ (see Lemma 2), the non-negativity of the right side of (39) implies that

$$\frac{1}{\rho}\,\log\mathbb{E}[g^{\rho}(X)]\geq\frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X)-\log\sum_{x\in\mathcal{X}}g^{-\beta}(x)\right] \tag{40}$$

holds for $\beta=1$ and $\rho\in(-1,0)\cup(0,\infty)$; furthermore, since $D_{1+\rho}(R_{1+\rho}\|Q_1)\leq 0$ for $\rho<-1$ (see Lemma 2), the non-positivity of the right side of (39) implies that

$$\frac{1}{\rho}\,\log\mathbb{E}[g^{\rho}(X)]\leq\frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X)-\log\sum_{x\in\mathcal{X}}g^{-\beta}(x)\right] \tag{41}$$

holds for $\beta=1$ and $\rho\in(-\infty,-1)$.

We proceed to show (40) and (41) for the range of β specified in (33) and (34) respectively. Consider next $\rho\neq 0$ and $\beta\neq 0$. Let $\widetilde{\rho}\triangleq\frac{\rho}{\beta}$.

To prove (33):

i) If $\beta\in(-\rho,0)$ with $\rho\in(0,\infty)$, then $\widetilde{\rho}\in(-\infty,-1)$. Since $1\in(-\infty,-\widetilde{\rho})$, (40) follows from the specialized version of (41) with $\beta\leftarrow 1$ and $(\rho,g)\leftarrow(\widetilde{\rho},g^{\beta})$.

ii) If $\beta\in(0,\infty)$ with $\rho\in(0,\infty)$, then $\widetilde{\rho}\in(0,\infty)$; and if $\beta\in(-\rho,\infty)$ with $\rho\in(-\infty,0)$, then $\widetilde{\rho}\in(-1,0)$. In both cases, $1\in(-\widetilde{\rho},\infty)$ and, consequently, (40) follows from its specialized version with $\beta\leftarrow 1$ and $(\rho,g)\leftarrow(\widetilde{\rho},g^{\beta})$.

To prove (34):

iii) If $\beta\in(-\infty,-\rho)$ with $\rho\in(0,\infty)$, then $\widetilde{\rho}\in(-1,0)$; and if $\beta\in(-\infty,0)$ with $\rho\in(-\infty,0)$, then $\widetilde{\rho}\in(0,\infty)$. In both cases $1\in(-\widetilde{\rho},\infty)$, and therefore (41) follows from the specialized version of (40) with $\beta\leftarrow 1$ and $(\rho,g)\leftarrow(\widetilde{\rho},g^{\beta})$.

iv) If $\beta\in(0,-\rho)$ with $\rho\in(-\infty,0)$, then $\widetilde{\rho}\in(-\infty,-1)$; this yields (41) from its specialized version with $\beta\leftarrow 1$ and $(\rho,g)\leftarrow(\widetilde{\rho},g^{\beta})$.

To prove Item 2):

- Suppose the set $\mathcal{X}$ is finite. By letting $\tau=\nu$ in (35), the identity in (36) follows easily.
- Suppose that $P_X=Q_\nu$ with $\nu>0$ and $\nu\neq\rho$. Let $\beta^*=\nu-\rho$ and let $(\rho,g)\leftarrow(\frac{\rho}{\beta^*},g^{\beta^*})$, which yields $Q_1\leftarrow Q_{\beta^*}$ (see (35)), $R_{1+\rho}\leftarrow R_{\nu/\beta^*}$, and $\frac{1}{1+\rho}\leftarrow\frac{\beta^*}{\beta^*+\rho}$. Then, from (38)–(39),

$$D_{\nu/\beta^*}(R_{\nu/\beta^*}\|Q_{\beta^*})=\frac{\beta^*}{\rho}\log\mathbb{E}[g^{\rho}(X)]+\log\left(\sum_{x\in\mathcal{X}}\frac{1}{g^{\beta^*}(x)}\right)-H_{\frac{\beta^*}{\beta^*+\rho}}(X). \tag{42}$$

Since by assumption $P_X=Q_\nu$, it is easy to verify from (35) and (37) that $R_{\nu/\beta^*}=Q_{\beta^*}$. Hence, both sides of (42) are equal to zero, which therefore implies that the supremum in the left side of (33) is attained at $\beta=\beta^*\in(-\rho,+\infty)\setminus\{0\}$, and the inequality in (33) becomes an identity. This proves the first part of Item 2b). To prove its second part, assume that (33) is an identity and the supremum is attained at β^* in the range of β specified in (33). This implies that the right side of (42) is equal to zero, and so is its left side. In view of Lemma 2, since $\nu = 0$, it follows that $R_{\nu/\beta^*} = Q_{\beta^*}$ which then gives that $P_X = Q_\nu$.

- The proof of Item 2c) is analogous to the proof of Item 2b).

■

Remark 4: Theorem 1 can be proved in the following alternative way: For $\beta \neq 0$, let

$$\alpha = \frac{\rho + \beta}{\beta}. \tag{43}$$

It can be verified from (8), (35) and (37) that

$$D_\alpha(R_\alpha \| Q_\beta) = \frac{\beta}{\rho} \log \mathbb{E}[g^\rho(X)] + \log \left(\sum_{x \in \mathcal{X}} g^{-\beta}(x) \right) - H_{\frac{\beta}{\beta+\rho}}(X), \tag{44}$$

and then, in view of Lemma 2, the following cases are considered for the free parameter β:

i) If $\beta \in (-\rho, \infty)$ and $\beta > 0$, then $\alpha > 0$ and $D_\alpha(R_\alpha \| Q_\beta) \geq 0$; hence, the right side of (44) is non-negative, and dividing it by β gives (40);
ii) If $\beta \in (-\rho, \infty)$ and $\beta < 0$, then $\alpha < 0$, and therefore $D_\alpha(R_\alpha \| Q_\beta) \leq 0$; since the right side of (44) is non-positive, dividing it by β gives (40);
iii) If $\beta \in (-\infty, -\rho)$ and $\beta < 0$, then $\alpha > 0$ and $D_\alpha(R_\alpha \| Q_\beta) \geq 0$; hence, the right side of (44) is non-negative, and dividing it by β gives (41);
iv) If $\beta \in (-\infty, -\rho)$ and $\beta > 0$, then $\alpha < 0$, and therefore $D_\alpha(R_\alpha \| Q_\beta) \leq 0$; since the right side of (44) is non-positive, dividing it by β gives (41).

This gives the lower and upper bounds in (33) and (34), respectively, after an optimization of the right sides of (40) and (41) over the free parameter $\beta \in (-\rho, \infty) \setminus \{0\}$ and $\beta \in (-\infty, -\rho) \setminus \{0\}$, respectively. Item 2) can be proved in a similar way to our earlier proof by relying on (43), (44), and Lemma 2; note that in view of (35), (37) and (43), if $\beta = \nu - \rho$, then $R_\alpha = Q_\beta$ if and only if $P_X = Q_\nu$.

Remark 5: For $\rho > 0$, the supremum over β in the right side of (33) involves negative orders of the Rényi entropy whenever $\beta \in (-\rho, 0)$. As shown in Example 1, the optimal value of $\beta \in (-\rho, \infty) \setminus \{0\}$ can be negative; furthermore, for every such β, Theorem 1 asserts the existence of a probability mass function for which (33) is achieved with equality. Allowing Rényi entropy of negative orders in Theorem 1 is therefore beneficial.

The particularization of Item 1) in Theorem 1 to $\beta = 1$ yields the following, generally looser, bound:

Corollary 2: [16, Lemma 2] Let X and g be as in Theorem 1, and $\rho \in (-1, 0) \cup (0, \infty)$. Then,

$$\frac{1}{\rho} \log \mathbb{E}\big[g^\rho(X)\big] \geq H_{\frac{1}{1+\rho}}(X) - \log \sum_{x \in \mathcal{X}} \frac{1}{g(x)}. \tag{45}$$

Remark 6: The proof of Corollary 2, which serves as a first step in proving Theorem 1, differs from its proof in [16, Lemma 2] (see also [1, Theorem 1] for a specialized version where $\rho > 0$). The proofs in [1] and [16] rely on the reverse Hölder inequality, whereas the proof here is based on the Rényi divergence.

B. Lower bounds

Theorem 1 is applied in this section to derive lower bounds on guessing moments with or without side information, improving the bounds in [1].

Theorem 2: Let X be a random variable taking values on the finite set $\mathcal{X} = \{1, \dots, M\}$, and let $g \colon \mathcal{X} \to \mathcal{X}$ be an arbitrary guessing function. Then, for every $\rho \neq 0$,

$$\frac{1}{\rho} \log \mathbb{E}\big[g^\rho(X)\big] \geq \sup_{\beta \in (-\rho, \infty) \setminus \{0\}} \frac{1}{\beta} \Big[H_{\frac{\beta}{\beta+\rho}}(X) - \log u_M(\beta)\Big] \tag{46}$$

with

$$u_M(\beta) = \begin{cases} \log_{\mathrm{e}} M + \gamma + \dfrac{1}{2M} - \dfrac{5}{6(10M^2+1)} & \beta = 1, \\[2ex] \min\left\{\zeta(\beta) - \dfrac{(M+1)^{1-\beta}}{\beta - 1} - \dfrac{(M+1)^{-\beta}}{2},\, u_M(1)\right\} & \beta > 1, \\[2ex] 1 + \dfrac{1}{1-\beta} \Big[\big(M + \tfrac{1}{2}\big)^{1-\beta} - \big(\tfrac{3}{2}\big)^{1-\beta}\Big] & |\beta| < 1, \\[2ex] \dfrac{M^{1-\beta} - 1}{1 - \beta} + \tfrac{1}{2}\Big(1 + M^{-\beta}\Big) & \beta \leq -1 \end{cases} \tag{47}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant, and $\zeta(\beta) = \sum_{n=1}^{\infty} \frac{1}{n^\beta}$ is the Riemann zeta function for $\beta > 1$.

Proof: Since $g \colon \mathcal{X} \to \mathcal{X}$ is a one-to-one function with $\mathcal{X} = \{1, \dots, M\}$,

$$\sum_{x \in \mathcal{X}} g^{-\beta}(x) = \sum_{i=1}^{M} \frac{1}{i^\beta}. \tag{48}$$

In view of (33), we derive (46) by proving that

$$\sum_{i=1}^{M} \frac{1}{i^\beta} \leq u_M(\beta), \quad \beta \geq 0, \tag{49a}$$

$$\sum_{i=1}^{M} \frac{1}{i^\beta} \geq u_M(\beta), \quad \beta \leq 0 \tag{49b}$$

where $u_M(\beta)$ is given in (47). Note that the restriction $\beta > -\rho$ in (46) is due to (33). The proof of (49a) and (49b) is deferred to Appendix A. ■

Remark 7: Specializing Theorem 2 to $\beta = 1$ and using $u_M(1) \leq 1 + \log_{\mathrm{e}} M$ for $M \geq 2$, we obtain

$$\frac{1}{\rho} \log \mathbb{E}\big[g^\rho(X)\big] \geq H_{\frac{1}{1+\rho}}(X) - \log\big(1 + \log_{\mathrm{e}} M\big) \tag{50}$$

for $\rho \in (-1, \infty)$. This bound was obtained in the range $\rho > 0$ by Arikan [1, (1)].

The following remark justifies the utility of Theorem 2.

Remark 8: Since Theorem 1 applies in particular to guessing functions, it gives a lower bound on $\frac{1}{\rho} \log \mathbb{E}\big[g^\rho(X)\big]$ where $u_M(\beta)$ in (46)–(47) is replaced by the finite sum $\sum_{j=1}^{M} \frac{1}{j^\beta}$ for $\beta \in (-\rho, \infty) \setminus \{0\}$. While numerical evidence shows that it is slightly better than the bound in Theorem 2 for large M, the latter bound is much easier to compute if M is large.

The following simple example illustrates the improvement afforded by the lower bound in Theorem 2, as well as the sub-optimality of $\beta = 1$.

Example 1: Let X be geometrically distributed restricted to $\{1, \dots, M\}$ with the probability mass function

$$P_X(k) = \frac{(1-a)\, a^{k-1}}{1-a^M}, \quad k \in \{1, \dots, M\} \tag{51}$$

where $a = \frac{24}{25}$ and $M = 128$. Since P_X is monotonically decreasing on $\{1, \dots, M\}$, the optimal guessing function is $g_X(i) = i$ for $i \in \{1, \dots, M\}$. Figure 1 compares $\frac{1}{6} \log_e \mathbb{E}\big[g_X^6(X)\big]$ with its lower bounds in (46) and (50). The numerical results exemplify the sub-optimality of $\beta = 1$ in the right side of (46). Not only is the

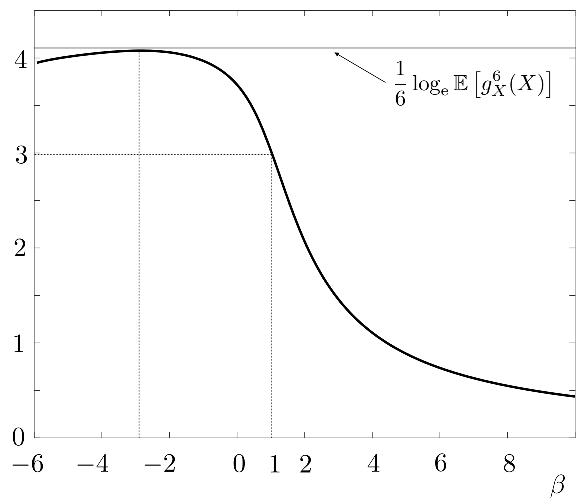

Fig. 1. Comparison of $\frac{1}{6} \log_e \mathbb{E}\big[g_X^6(X)\big] = 4.084$ for the random variable X in Example 1 with its lower bounds. The lower bound in (50) is equal to 2.953, while Theorem 2 improves this lower bound to 4.078 (attained at $\beta = -2.85$).

improvement over the lower bound in [1, (1)] significant, but the improved lower bound is very close to the actual value of the sixth guessing moment. Furthermore, the improved lower bound is attained at $\beta = -2.85$, thereby showing the benefit of allowing negative orders in the definition of Rényi entropy.

Remark 9: Massey [31] obtained a lower bound on the expected value of the minimal number of guesses for correctly guessing the value of a discrete random variable X which does not necessarily take a finite number of values:

$$H(X) \leq \mathbb{E}[g(X)]\, h\left(\frac{1}{\mathbb{E}[g(X)]}\right) \tag{52}$$

$$\leq \log \mathbb{E}[g(X)] + \log \mathrm{e}, \tag{53}$$

where h is the binary entropy function, and (52) is achieved with equality if and only if X is geometrically distributed. The idea behind (52) is that $H(X) = H(g(X))$ (since X and $g(X)$ are in one-to-one correspondence), and the entropy of the positive integer-valued random variable $g(X)$ with a given value of its expectation is maximized when X is geometrically distributed, which gives the right side of (52). Note that (52) provides an implicit lower bound on $\mathbb{E}[g(X)]$ as a function of the Shannon entropy $H(X)$, whereas (53) gives a looser though explicit lower bound.

Massey's bound (52) can be generalized to obtain a lower bound on the ρ-th moment of the guessing function as a function of $H_\alpha(X)$ for arbitrary $\alpha, \rho > 0$. To this end, we first rely on the basic equality

$$H_\alpha(X) = H_\alpha(g(X)) \tag{54}$$

since g is a one-to-one function, and then we seek the distribution of the random variable $Z = g(X)$, taking values on $\{1, \ldots, M\}$ with possibly $M = \infty$, which maximizes $H_\alpha(Z)$ under the equality constraint

$$\mathbb{E}[Z^\rho] = \theta \tag{55}$$

for fixed $\theta > 1$. The problem of maximizing the Rényi entropy under general equality constraints was studied by Bunte and Lapidoth (see [10, Theorem II.1]) in the setting of continuous random variables. In the discrete setting, a similar proof which relies on the non-negativity of Sundaresan's divergence [46] provides the maximizing distribution of the Rényi entropy $H_\alpha(Z)$ for a fixed value of $\mathbb{E}[Z^\rho]$:

$$P_Z(n) = \begin{cases} c\,(1+\lambda n^\rho)^{\frac{1}{\alpha-1}} \; 1\{1 \le n \le M\}, & \alpha \in (0,1), \; \lambda > -M^{-\rho} \\ c\,(1+\lambda n^\rho)^{\frac{1}{\alpha-1}} \; 1\{n \in \mathcal{T}\}, & \alpha > 1 \end{cases} \tag{56}$$

where $\mathcal{T} \triangleq \mathcal{T}_{\lambda,\rho} = \{1, \ldots, M\} \cap \{n : 1 + \lambda n^\rho \ge 0\}$, and $\lambda \in \mathbb{R}$ is set to satisfy (55) with the normalizing constant c in (56). Hence, we get from (54) and (56) that

$$H_\alpha(X) \le H_\alpha(Z) \tag{57}$$

with equality for the maximizing distribution P_Z in (56) which is selected to satisfy (55). An analogous way to view this problem is to minimize $\mathbb{E}[g^\rho(X)]$ for a given value of $H_\alpha(X)$. This leads to an extension of Massey's bound in (52) which, however, does not lend itself to a closed-form generalized bound and its computation is rather involved (especially, for $\alpha > 1$). The lower bound in Theorem 2, on the other hand, can be calculated more easily.

C. *Upper bounds*

The average number of guesses is minimized by taking the guessing function to be the ranking function g_X, for which $g_X(x) = k$ if $P_X(x)$ is the k-th largest mass [31]. Although the tie breaking affects the choice of g_X, the distribution of $g_X(X)$ does not depend on how ties are resolved. Not only does this strategy minimize the average number of guesses, but it also minimizes the ρ-th moment of the number of guesses for every $\rho > 0$.

Theorem 3: [1, Proposition 4] Let X be a discrete random variable taking values on a set $\mathcal{X}$, and let g_X be the ranking function according to P_X. Then, for all $\rho > 0$,

$$\mathbb{E}[g_X^\rho(X)] \le \exp\Bigl(\rho H_{\frac{1}{1+\rho}}(X)\Bigr). \tag{58}$$

The following result tightens Theorem 3.

Theorem 4: Under the assumptions in Theorem 3, for all $\rho \ge 0$,

$$\mathbb{E}[g_X^\rho(X)] \le \frac{1}{1+\rho}\left[\exp\Bigl(\rho H_{\frac{1}{1+\rho}}(X)\Bigr) - 1\right] + \exp\Bigl((\rho-1)^+ H_{\frac{1}{\rho}}(X)\Bigr) \tag{59}$$

where $(x)^+ \triangleq \max\{x, 0\}$ for $x \in \mathbb{R}$.

Proof: From [6, Lemma 2], if $\rho \ge 0$, then

$$\mathbb{E}\bigl[g_X^{1+\rho}(X)\bigr] - \mathbb{E}\bigl[\bigl(g_X(X)-1\bigr)^{1+\rho}\bigr] \le \left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x)\right)^{1+\rho} = \exp\Bigl(\rho H_{\frac{1}{1+\rho}}(X)\Bigr). \tag{60}$$

At this point we deviate from the analysis in [6]. For $\rho \ge 1$, let $r \colon [1,\infty) \to \mathbb{R}$ and $v \colon [0,\infty) \to \mathbb{R}$ be given by

$$r(u) = \frac{u^{1+\rho} - (u-1)^{1+\rho} - 1}{1+\rho} - (u-1)^\rho, \quad u \ge 1 \tag{61}$$

$$v(u) = u^\rho, \quad u \ge 0. \tag{62}$$

Expressing the derivative of (61) with the aid of (62) implies that, for all $u \geq 1$,

$$r'(u) = v(u) - v(u-1) - v'(u-1) \geq 0, \tag{63}$$

where the inequality is due to the convexity of $v(\cdot)$ in $(0, \infty)$. Since $r(1) = 0$, it follows that $r(\cdot)$ is non-negative in $[1, \infty)$. Invoking the result in (60), along with $\mathbb{E}[r(g_X(X))] \geq 0$, implies that for $\rho \geq 1$

$$\mathbb{E}\big[\big(g_X(X) - 1\big)^{\rho}\big] \leq \frac{1}{1+\rho} \left[\exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) - 1\right]. \tag{64}$$

A replacement of ρ by $\rho - 1$ in (60) gives that for $\rho \geq 1$

$$\mathbb{E}\big[g_X^{\rho}(X)\big] - \mathbb{E}\big[\big(g_X(X) - 1\big)^{\rho}\big] \leq \exp\Big((\rho - 1)\, H_{\frac{1}{\rho}}(X)\Big). \tag{65}$$

Hence, for $\rho \geq 1$, adding (64) and (65) yields

$$\mathbb{E}[g_X^{\rho}(X)] \leq \frac{1}{1+\rho} \left[\exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) - 1\right] + \exp\Big((\rho-1) H_{\frac{1}{\rho}}(X)\Big). \tag{66}$$

This proves (59) for $\rho \geq 1$. We next consider the case where $\rho \in (0,1)$. Lemma 8 in Appendix B yields

$$u^{\rho} \leq \frac{u^{1+\rho} - (u-1)^{1+\rho} + \rho}{1+\rho}, \quad u \geq 1, \tag{67}$$

and therefore, for $\rho \in (0,1)$,

$$\mathbb{E}\big[g_X^{\rho}(X)\big] \leq \frac{1}{1+\rho} \left(\mathbb{E}\left[g_X^{1+\rho}(X)\right] - \mathbb{E}\left[\big(g_X(X) - 1\big)^{1+\rho}\right]\right) + \frac{\rho}{1+\rho}. \tag{68}$$

Combining (60) and (68) gives (59) for $\rho \in (0,1)$. The case $\rho = 0$, for which (59) holds with equality, is trivial. ■

Remark 10: Bound (59) strictly (unless X is deterministic) improves the bound in (58). This holds for $\rho \in (0,1)$ since $H_{\frac{1}{1+\rho}}(X) \geq 0$; for $\rho \in [1, \infty)$, the difference between the bounds in (58) and (59) is equal to

$$\begin{aligned}
&\exp\Big(\rho\, H_{\frac{1}{1+\rho}}(X)\Big) - \left\{\frac{1}{1+\rho}\left[\exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) - 1\right] + \exp\Big((\rho-1) H_{\frac{1}{\rho}}(X)\Big)\right\} \\
&= \frac{\rho}{1+\rho} \left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x)\right)^{1+\rho} - \left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{\rho}}(x)\right)^{\rho} + \frac{1}{1+\rho} \\
&\geq \frac{\rho}{1+\rho} \left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x)\right)^{1+\rho} - \left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x)\right)^{\rho} + \frac{1}{1+\rho} &(69)\\
&= s\left(\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x)\right) &(70)
\end{aligned}$$

where (69) holds since $P_X^{\frac{1}{\rho}}(x) \leq P_X^{\frac{1}{1+\rho}}(x)$, and (70) holds with $s \colon [1, \infty) \to \mathbb{R}$ given by

$$s(u) = u^{\rho}\left(\frac{\rho u}{1+\rho} - 1\right) + \frac{1}{1+\rho}, \quad u \geq 1. \tag{71}$$

In (70), note that $\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x) \geq \sum_{x \in \mathcal{X}} P_X(x) = 1$. The right side of (70) is non-negative since $s(1) = 0$, and

$$s'(u) = \rho(u-1)u^{\rho-1} > 0 \tag{72}$$

for all $u > 1$. This shows that the bound in (59) is at least as tight as the bound in (58) with a strict improvement unless X is deterministic.

In the range $\rho \in [0, 2]$, we can tighten (59) according to the following result.

Theorem 5: Under the assumptions in Theorem 3,

a) For $\rho \in [0, 1]$

$$\mathbb{E}[g_X^\rho(X)] \leq \frac{1}{1+\rho} \exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) + \frac{\rho - (1-\rho)(2^\rho - 1)(1-p_{\max})}{1+\rho}. \tag{73}$$

b) For $\rho \in [1, 2]$

$$\mathbb{E}[g_X^\rho(X)] \leq \frac{1}{1+\rho} \exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) + \frac{1}{\rho} \exp\Big((\rho-1) H_{\frac{1}{\rho}}(X)\Big) + \frac{\rho^2 - \rho - 1}{\rho(1+\rho)}. \tag{74}$$

Furthermore, both (73) and (74) hold with equality if X is deterministic.

Proof: See Appendix B. ■

Remark 11: Particularizing (74) to $\rho = 1$ and $\rho = 2$, we recover the bounds on the first and second moments in [6, Theorem 3]. Furthermore, the bounds in (73) and (74) provide a continuous transition at $\rho = 1$.

Theorem 6: Under the assumptions in Theorem 3, for $\rho \geq 2$,

$$\mathbb{E}[g_X^\rho(X)] \leq 1 + \sum_{j=0}^{\lfloor \rho \rfloor} c_j(\rho) \left[\exp\Big((\rho - j)\, H_{\frac{1}{1+\rho-j}}(X) \Big) - 1 \right], \tag{75}$$

where $\{c_j(\rho)\}$ is given by

$$c_j(\rho) = \begin{cases} \dfrac{1}{1+\rho} & j = 0 \\ \frac{1}{2} & j = 1 \\ \dfrac{\rho \ldots (\rho - j + 2)}{2^j} & j \in \{2, \ldots, \lfloor \rho \rfloor - 1\} \\ \dfrac{\rho \ldots (\rho - j + 2)}{2^{j-1}\, (\rho - j + 1)} & j = \lfloor \rho \rfloor \end{cases} \tag{76}$$

and $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to x.

Proof: See Appendix C. ■

Remark 12: In contrast to [6, Theorem 3], the results in Theorems 5 and 6 provide an explicit upper bound on $\mathbb{E}[g_X^\rho(X)]$ for $\rho \in (0, \infty)$ as a function of Rényi entropies of X. Note also that the upper bounds in (74) and (75) coincide at $\rho = 2$.

Remark 13: Numerical evidence shows that none of the bounds in (59) and (75) supersedes the other for $\rho > 2$ (as it is next illustrated in Examples 2 and 3). Since (74) and (75) coincide at $\rho = 2$, Theorem 5-b) implies that the bound in (75) is tighter than (59) for this value of ρ.

Example 2: Let $X \in \{1, \ldots, 32\}$ have the probability distribution in (51) with $a = 0.9$ and $M = 32$. Table I compares $\frac{1}{3} \log_e \mathbb{E}[g_X^3(X)]$ to its various lower and upper bounds. Notice that in this example, the upper bound in (75) improves the bound in (59).

Example 3: Let $X \in \{1, \ldots, 16\}$ have the probability distribution in (51) with $a = 0.9$ and $M = 16$. Table II compares $\frac{1}{20} \log_e \mathbb{E}[g_X^{20}(X)]$ to its various lower and upper bounds. Note that in Table II, the lower bound in (50) is quite weaker than the lower bound in Theorem 2, which shows an excellent match with its exact value. In contrast to Example 2, Example 3 shows that the upper bound (75) may be weaker than (59).

TABLE I. COMPARISON OF $\frac{1}{3}\log_e \mathbb{E}[g_X^3(X)]$ AND BOUNDS IN EXAMPLE 2.

(50) lower bound	Theorem 2 lower bound	$\frac{1}{3}\log_e \mathbb{E}[g_X^3(X)]$ exact value	(75) upper bound	(59) upper bound	(58) upper bound
1.864	2.593	2.609	2.920	2.939	3.360

TABLE II. COMPARISON OF $\frac{1}{20}\log_e \mathbb{E}[g_X^{20}(X)]$ AND BOUNDS IN EXAMPLE 3.

(50) lower bound	Theorem 2 lower bound	$\frac{1}{20}\log_e \mathbb{E}[g_X^{20}(X)]$ exact value	(75) upper bound	(59) upper bound	(58) upper bound
1.439	2.602	2.606	2.662	2.657	2.767

D. Improved bounds on guessing moments with side information

This subsection extends the lower and upper bounds in Sections III-B and III-C to allow side information Y for guessing the value of X. These bounds tighten the results in [1, Theorem 1] and [1, Proposition 4] for all $\rho > 0$.

Theorem 7: Let X and Y be discrete random variables taking values on the sets $\mathcal{X} = \{1, \ldots, M\}$ and $\mathcal{Y}$, respectively. For all $y \in \mathcal{Y}$, let $g(\cdot|y)$ be a guessing function of X given that $Y = y$. Then, for $\rho \in (0, \infty)$,

$$\frac{1}{\rho}\,\log \mathbb{E}\big[g^\rho(X|Y)\big] \geq \sup_{\beta \in (-\rho,0)\cup(0,\infty)} \frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X|Y) - \log u_M(\beta)\right] \tag{77}$$

with $u_M(\cdot)$ as defined in (47).

Proof:

$$\begin{aligned}
\mathbb{E}[g^\rho(X|Y)] &= \sum_{y\in\mathcal{Y}} P_Y(y)\,\mathbb{E}[g^\rho(X|y)] && (78)\\
&\geq \sum_{y\in\mathcal{Y}} P_Y(y)\,\exp\left(\sup_{\beta\in(-\rho,0)\cup(0,\infty)} \frac{\rho}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X|Y=y) - \log u_M(\beta)\right]\right) && (79)\\
&\geq \sup_{\beta\in(-\rho,0)\cup(0,\infty)} \sum_{y\in\mathcal{Y}} P_Y(y)\,\exp\left(\frac{\rho}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X|Y=y) - \log u_M(\beta)\right]\right) && (80)\\
&= \sup_{\beta\in(-\rho,0)\cup(0,\infty)} \exp\left(\frac{\rho}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X|Y) - \log u_M(\beta)\right]\right) && (81)
\end{aligned}$$

where (79) follows from (46) with the conditional guessing function $g(\cdot|Y = y)\colon \mathcal{X} \to \mathcal{X}$ where $|\mathcal{X}| = M$; (81) follows from (14) with $\alpha = \frac{\beta}{\rho+\beta}$. ■

Remark 14: Similarly to Remark 7, the loosening of the result in Theorem 7 by replacing the supremum over $\beta \in (-\rho, 0) \cup (0, \infty)$ in the right side of (77) with the value $\beta = 1$ and further using the inequality $u_M(1) \leq 1 + \log_e M$ for $M \geq 2$ (see (47)) yields Arikan's result in [1, (2)]. As explained in [1], Theorem 7 can be used to obtain an improved non-asymptotic lower bound on the moments of the number of computational steps used to decode tree codes by an arbitrary sequential decoder.

Theorem 8: Let X and Y be discrete random variables taking values on sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. For all $y \in \mathcal{Y}$, let $g_{X|Y}(\cdot|y)$ be a ranking function of X given that $Y = y$. Then, for $\rho \in (0, \infty)$,

$$\mathbb{E}[g^\rho_{X|Y}(X|Y)] \leq \frac{1}{1+\rho}\left[\exp\Big(\rho H_{\frac{1}{1+\rho}}(X|Y)\Big) - 1\right] + \exp\Big((\rho-1)^+\, H_{\frac{1}{\rho}}(X|Y)\Big). \tag{82}$$

For $\rho \in (0,1)$, (82) can be tightened to

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] \leq \frac{1}{1+\rho}\, \exp\Big(\rho H_{\frac{1}{1+\rho}}(X|Y)\Big) + \frac{\rho - (1-\rho)(2^{\rho}-1)(1-p^*)}{1+\rho} \tag{83}$$

with

$$p^* \triangleq \sup_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}} p_{X|Y}(x|y). \tag{84}$$

For $\rho \in [1,2]$, (82) can be tightened to

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] \leq \frac{1}{1+\rho}\, \exp\Big(\rho H_{\frac{1}{1+\rho}}(X|Y)\Big) + \frac{1}{\rho}\, \exp\Big((\rho-1) H_{\frac{1}{\rho}}(X|Y)\Big) + \frac{\rho^2-\rho-1}{\rho(1+\rho)}. \tag{85}$$

Moreover, for $\rho \geq 2$,

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] \leq 1 + \sum_{j=0}^{\lfloor \rho \rfloor} c_j(\rho) \left[\exp\Big((\rho-j)\, H_{\frac{1}{1+\rho-j}}(X|Y)\Big) - 1\right], \tag{86}$$

with $\{c_j(\rho)\}$ defined in (76).

Proof: From Theorem 4 and (78), for all $\rho \in [1,\infty)$,

$$\begin{aligned}
&\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] \\
&\leq \frac{1}{1+\rho}\left[\sum_{y\in\mathcal{Y}} P_Y(y)\, \exp\Big(\rho H_{\frac{1}{1+\rho}}(X \,|\, Y=y)\Big) - 1\right] + \sum_{y\in\mathcal{Y}} P_Y(y)\, \exp\Big((\rho-1)\, H_{\frac{1}{\rho}}(X \,|\, Y=y)\Big) && (87)\\
&= \frac{1}{1+\rho}\left[\exp\Big(\rho H_{\frac{1}{1+\rho}}(X|Y)\Big) - 1\right] + \exp\Big((\rho-1)\, H_{\frac{1}{\rho}}(X|Y)\Big) && (88)
\end{aligned}$$

where (87) follows from (59), and (88) follows from (14) with $\alpha = \frac{1}{1+\rho}$ and $\alpha = \frac{1}{\rho}$. The proof of (82) for $\rho \in (0,1)$ is similar.

Extending (73), (74) and (75) to (83), (85) and (86), respectively, relies on (14) and it is similar to the extension of the result in Theorem 4 to (82). ■

Remark 15: The bound in (82) improves the result in [1, Proposition 4] since the superiority of the bound in Theorem 4 over the result in Theorem 3 is preserved after the averaging in (78).

Remark 16: Following Remark 13, neither of the bounds in (82) and (86) supersedes the other for $\rho \geq 2$. Since (85) and (86) coincide for $\rho = 2$, it follows from Theorem 8 that the upper bound in (86) is tighter than (82) for this value of ρ. Note that both bounds are tight if X, conditioned on Y, is deterministic, for which they are equal to $\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] = 1$. Finally, note that the transition from (83) to (85) is continuous at $\rho = 1$.

E. Relationship between guessing moments and minimum probability of error

Let X and Y be discrete random variables,[2] taking values on the sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. The minimum probability of error of X given Y, denoted by $\varepsilon_{X|Y}$, is achieved by the maximum-a-posteriori (MAP) decision rule. Hence,

$$\varepsilon_{X|Y} = \sum_{y\in\mathcal{Y}} P_Y(y) \left[1 - \max_{x\in\mathcal{X}} P_{X|Y}(x|y)\right]. \tag{89}$$

---

[2]The assumption that Y is a discrete random variable can be easily dispensed with.

In contrast, the moments of the ranking function $\mathbb{E}[g^{\rho}_{X|Y}(X|Y)]$ quantify the number of guesses required for correctly identifying the unknown object X on the basis of Y. It is therefore natural to establish relationships between both quantities. First note that by definition,

$$1 - \varepsilon_{X|Y} = \mathbb{P}[g_{X|Y}(X|Y) = 1]. \tag{90}$$

In [42], we derived lower and upper bounds on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$ for an arbitrary order α. In this section, Theorems 7–8 provide lower and upper bounds on guessing moments of a ranking function $g_{X|Y}(X|Y)$ as a function of Arimoto-Rényi conditional entropies. As a natural continuation to these studies, we derive tight lower and upper bounds on $\mathbb{E}[g^{\rho}_{X|Y}(X|Y)]$ as a function of $\varepsilon_{X|Y}$.

Theorem 9: Let X and Y be discrete random variables taking values on sets $\mathcal{X} = \{1, \ldots, M\}$ and $\mathcal{Y}$, respectively. Then, for $\rho > 0$,

$$\begin{aligned} f_\rho(\varepsilon_{X|Y}) &\le \mathbb{E}[g^{\rho}_{X|Y}(X|Y)] &(91)\\ &\le 1 + \left(\tfrac{1}{M-1} \sum_{j=2}^{M} j^\rho - 1 \right) \varepsilon_{X|Y} &(92) \end{aligned}$$

where the function $f_\rho \colon [0,1) \to [0, \infty)$ is given by

$$\begin{aligned} f_\rho(u) &= (1-u) \sum_{j=1}^{k_u} j^\rho + [1 - (1-u)k_u](k_u+1)^\rho, \quad u \in [0,1) &(93)\\ k_u &= \left\lfloor \frac{1}{1-u} \right\rfloor. &(94) \end{aligned}$$

Furthermore, the lower and upper bounds in (91) and (92) are tight:

- Let $p_{\max}(y) = \max\limits_{x \in \mathcal{X}} P_{X|Y}(x|y)$ for $y \in \mathcal{Y}$. The lower bound is attained if and only if $p_{\max}(y) = p_{\max}$ is fixed for all $y \in \mathcal{Y}$, and conditioned on $Y = y$, 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.
- The upper bound is attained if and only if regardless of $y \in \mathcal{Y}$, conditioned on $Y = y$, X is equiprobable among its $M-1$ conditionally least likely values on $\mathcal{X}$.

Proof: For $y \in \mathcal{Y}$ and $i \in \{1, \ldots, M\}$, let $x_i(y) \in \mathcal{X}$ satisfy $g_{X|Y}\big(x_i(y)|y\big) = i$. By the definition of $g_{X|Y}(\cdot|\cdot)$ on $\mathcal{X} \times \mathcal{Y}$, it follows that

$$P_{X|Y}\big(x_1(y)|y\big) \ge P_{X|Y}\big(x_2(y)|y\big) \ge \ldots \ge P_{X|Y}\big(x_M(y)|y\big). \tag{95}$$

For $\rho \in (0, \infty)$,

$$\begin{aligned} \mathbb{E}[g^{\rho}_{X|Y}(X|Y)] &= \sum_{y \in \mathcal{Y}} P_Y(y)\, \mathbb{E}[g^{\rho}_{X|Y}(X|y)] &(96)\\ &= \sum_{y \in \mathcal{Y}} \left\{ P_Y(y) \left(\max_{x \in \mathcal{X}} P_{X|Y}(x|y) + \sum_{i=2}^{M} i^\rho P_{X|Y}\big(x_i(y)|y\big) \right) \right\} &(97)\\ &= 1 - \varepsilon_{X|Y} + \sum_{y \in \mathcal{Y}} \left\{ P_Y(y) \sum_{i=2}^{M} i^\rho P_{X|Y}\big(x_i(y)|y\big) \right\} &(98)\\ &\le 1 - \varepsilon_{X|Y} + \frac{1}{M-1} \sum_{y \in \mathcal{Y}} \left\{ P_Y(y) \left(\sum_{i=2}^{M} i^\rho \right) \left(\sum_{i=2}^{M} P_{X|Y}\big(x_i(y)|y\big) \right) \right\} &(99) \end{aligned}$$

$$= 1 - \varepsilon_{X|Y} + \frac{1}{M-1} \sum_{y \in \mathcal{Y}} \left\{ P_Y(y) \left(1 - \max_{x \in \mathcal{X}} P_{X|Y}(x|y) \right) \right\} \sum_{i=2}^{M} i^\rho \tag{100}$$

$$= 1 + \left(\frac{1}{M-1} \sum_{i=2}^{M} i^\rho - 1 \right) \varepsilon_{X|Y} \tag{101}$$

where (97) and (100) follow from (95); (98) and (101) follow from (89); (99) follows from Lemma 12 (see Appendix D), for which we take the strictly monotonically increasing function f to be given by

$$f(i) = (i+1)^\rho, \quad i = 1, \ldots, M-1, \tag{102}$$

and $q_i \leftarrow P_{X|Y}\big(x_{i+1}(y)|y\big)$. This proves the upper bound in (92). A necessary and sufficient condition for attaining the upper bound in (92) follows from Lemma 12 and due to the strict monotonicity of the function in (102) for $\rho > 0$. Hence, it follows that (99) is satisfied with equality if and only if, for every $y \in \mathcal{Y}$, $q_i(y) \triangleq P_{X|Y}\big(x_{i+1}(y)|y\big)$ is fixed for all $i \in \{1, \ldots, M-1\}$; due to (95), this is equivalent to the requirement that regardless of $y \in \mathcal{Y}$, conditioned on $Y = y$, the $M-1$ least probable values of X are equiprobable.

To show (91), we write the conditional moment of the ranking function as

$$\mathbb{E}[g^\rho_{X|Y}(X|Y)] = \sum_{y \in \mathcal{Y}} \left\{ P_Y(y) \sum_{i=1}^{M} i^\rho P_{X|Y}\big(x_i(y)|y\big) \right\}, \tag{103}$$

and we denote the conditional error probability given the observation $Y = y$ by $\varepsilon_{X|Y}(y) = 1 - p_{\max}(y)$. Note from (95) that $x_1(y)$ is a mode of $P_{X|Y}(\cdot|y)$, and we have

$$P_{X|Y}\big(x_1(y)|y\big) = p_{\max}(y) = 1 - \varepsilon_{X|Y}(y). \tag{104}$$

The inner sum in the right side of (103) is minimized, for a given value of $\varepsilon_{X|Y}(y)$, by

$$P_{X|Y}(x_i(y)|y) = \begin{cases} 1 - \varepsilon_{X|Y}(y) & i = 1, \ldots, \left\lfloor \frac{1}{1-\varepsilon_{X|Y}(y)} \right\rfloor \\ 1 - (1 - \varepsilon_{X|Y}(y)) \left\lfloor \frac{1}{1-\varepsilon_{X|Y}(y)} \right\rfloor & i = \left\lfloor \frac{1}{1-\varepsilon_{X|Y}(y)} \right\rfloor + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{105}$$

In order to show it, note that according to (95), any perturbation of $P_{X|Y}(\cdot|y)$ in (105) necessarily shifts mass from $x_i(y)$ to $x_j(y)$ with $j > i$; since $\{k^\rho\}_{k \geq 1}$ is positive and monotonically increasing in k for $\rho > 0$, this can only increase the inner sum in the right side of (103).

From the minimizing conditional distribution in (105), for a given value of $\varepsilon_{X|Y}(y)$, and (93)–(94)

$$\sum_{i=1}^{M} i^\rho P_{X|Y}\big(x_i(y)|y\big) \geq f_\rho\big(\varepsilon_{X|Y}(y)\big). \tag{106}$$

Due to the convexity of $f_\rho \colon [0,1) \to [0, \infty)$ in (93)–(94) for $\rho > 0$ (see Lemma 13 in Appendix D), and since

$$\varepsilon_{X|Y} = \sum_{y \in \mathcal{Y}} P_Y(y) \varepsilon_{X|Y}(y), \tag{107}$$

the lower bound in (91) follows from (103), (106) and Jensen's inequality. This also yields the necessary and sufficient condition for the attainability of the bound in (91), as it specified in the statement of the theorem (note that, from (104), $p_{\max}(y)$ is fixed for all $y \in \mathcal{Y}$ if and only if $\varepsilon_{X|Y}(y)$ is so). ∎

Remark 17: The lower and upper bounds in (91) and (92) coincide in each of the extreme cases $\varepsilon_{X|Y} = 0$ and $\varepsilon_{X|Y} = 1 - \frac{1}{M}$. The former and latter values refer, respectively, to the following cases:

- X is a deterministic function of Y,
- X and Y are independent, and X is equiprobable.

Example 4: Let X and Y be random variables which take values on $\mathcal{X} = \{1, 2, 3, 4\}$, and let

$$\big[P_{XY}(x, y)\big]_{(x,y)\in\mathcal{X}^2} = \frac{1}{52} \begin{pmatrix} 10 & 1 & 1 & 1 \\ 1 & 10 & 1 & 1 \\ 1 & 1 & 10 & 1 \\ 1 & 1 & 1 & 10 \end{pmatrix}. \tag{108}$$

It follows from (108) that $P_Y(y) = \frac{1}{4}$ for all $y \in \mathcal{X}$, and we can select the conditional ranking function of X given Y to satisfy $g_{X|Y}(x|1) = x$ for all $x \in \mathcal{X}$ (since 1 is the most likely value of X given $Y = 1$, and 2,3,4 are conditionally equiprobable given $Y = 1$); moreover, symmetry in (108) yields

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] = \tfrac{1}{13}\left(10 + 2^{\rho} + 3^{\rho} + 4^{\rho}\right). \tag{109}$$

The result in (109) coincides with the upper bound in (92) since, regardless of y, $P_{X|Y}(x|y) = \frac{1}{13}$ for the $|\mathcal{X}| - 1 = 3$ least probable values of X given $Y = y$. This can be verified directly since it follows from (89) and (108) that $\varepsilon_{X|Y} = \frac{3}{13}$, and then the upper bound in (92) (with $M = 4$) is equal to the right side of (109).

Next example illustrates the locus of attainable values of $(\varepsilon_{X|Y}, \log_{\mathrm{e}} \mathbb{E}[g^{\rho}_{X|Y}(X|Y)])$ for a fixed M. The extreme cases identified in Remark 17 correspond to $(0, 0)$ and $\left(1 - \frac{1}{M}, \log_{\mathrm{e}}\left(\frac{1}{M}\sum_{j=1}^{M} j^{\rho}\right)\right)$; in these extreme cases, there is a one-to-one correspondence between the two quantities.

Example 5: Let X be a random variable taking values on $\mathcal{X}$ with $|\mathcal{X}| = M \geq 2$. Letting $\rho = 1$ in (91)–(94) yields

$$1 + \tfrac{1}{2}(1 + \varepsilon_{X|Y}) \left\lfloor \frac{1}{1 - \varepsilon_{X|Y}} \right\rfloor - \tfrac{1}{2}(1 - \varepsilon_{X|Y}) \left\lfloor \frac{1}{1 - \varepsilon_{X|Y}} \right\rfloor^2 \leq \mathbb{E}[g_{X|Y}(X|Y)] \leq 1 + \tfrac{1}{2} M \varepsilon_{X|Y} \tag{110}$$

where both upper and lower bounds on the expected number of guesses in (110) are attainable for any value of $\varepsilon_{X|Y} \in [0, 1 - \frac{1}{M}]$. The plots in Figure 2 illustrate the tight bounds in (110) for a fixed M.

In view of Theorems 2 and 9, the next result provides an explicit lower bound on $\varepsilon_{X|Y}$ as a function of $H_{\alpha}(X|Y)$ for any non-zero $\alpha < 1$.

Theorem 10: Let X and Y be discrete random variables taking values on sets $\mathcal{X} = \{1, \ldots, M\}$ and $\mathcal{Y}$, respectively. Then, for all $\alpha \in (-\infty, 0) \cup (0, 1)$,

$$\varepsilon_{X|Y} \geq \sup_{\rho > 0} \left\{ \frac{\exp\left(\left(\frac{1}{\alpha} - 1\right)\left[H_{\alpha}(X|Y) - \log u_M\left(\frac{\alpha\rho}{1-\alpha}\right)\right]\right) - 1}{\frac{1}{M-1}\sum\limits_{j=2}^{M} j^{\rho} - 1} \right\} \tag{111}$$

with $u_M(\cdot)$ as defined in (47).

Proof: Combining (77) and (92) yields, for every $\rho > 0$ and $\beta \in (-\rho, 0) \cup (0, \infty)$,

$$\varepsilon_{X|Y} \geq \frac{\exp\left(\frac{\rho}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X|Y) - \log u_M(\beta)\right]\right) - 1}{\frac{1}{M-1}\sum\limits_{j=2}^{M} j^{\rho} - 1}. \tag{112}$$

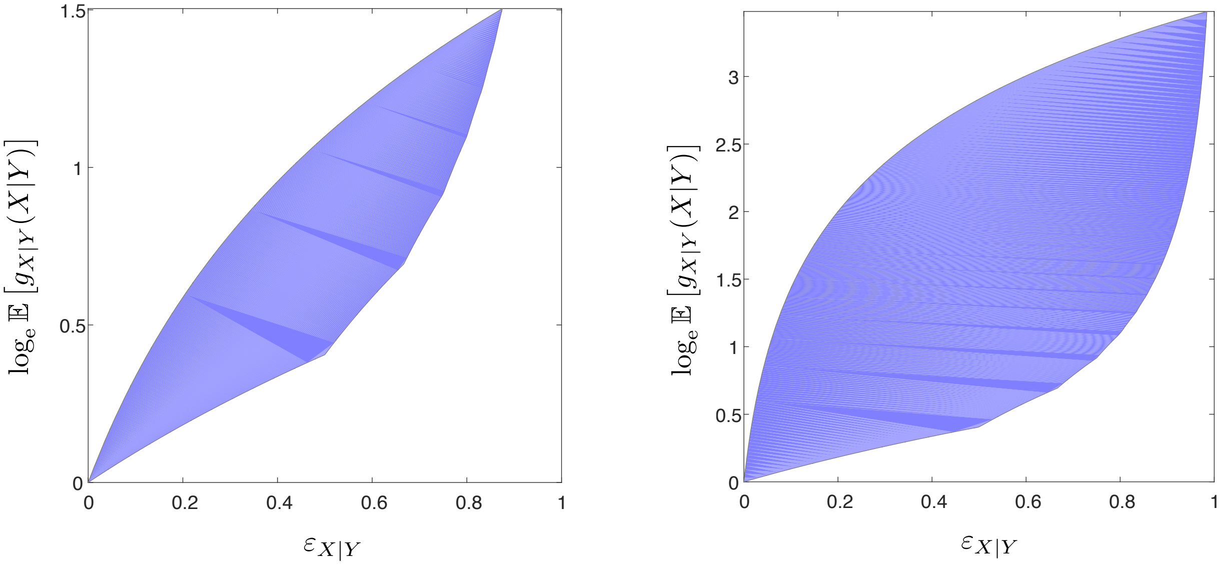

Fig. 2. Example 5: locus of attainable values of $(\varepsilon_{X|Y}, \log_e \mathbb{E}[g_{X|Y}(X|Y)])$. The random variable X takes $M = 8$ (left plot) or $M = 64$ (right plot) possible values.

Fix $\alpha \in (-\infty, 0) \cup (0, 1)$, and let the above free parameters β and ρ satisfy $\rho = \left(\frac{1}{\alpha} - 1\right)\beta$ (note that α cannot be zero or larger than 1, as otherwise $\beta \notin (-\rho, 0) \cup (0, \infty)$). Supremizing the right side of (112) over $\rho > 0$ yields (111). ■

In [42], we derived the following lower bounds on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$:

1) A generalization of Fano's inequality in [42, Theorem 3] holds for $\alpha > 0$, and it is given by

$$H_\alpha(X|Y) \leq \log M - d_\alpha\left(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\right) \tag{113}$$

with $d_\alpha(\cdot\|\cdot)$ as defined in (12).

2) An explicit lower bound in [42, Theorem 6] holds for $\alpha < 0$, and it is given by

$$\varepsilon_{X|Y} \geq \exp\left(\frac{1-\alpha}{\alpha}\left[H_\alpha(X|Y) - \log(M-1)\right]\right). \tag{114}$$

Example 6 includes numerical comparisons of the lower bounds on $\varepsilon_{X|Y}$ in (111), (113) and (114).

Remark 18: Shannon's inequality [43] (see also [49]) gives an explicit lower bound on $\varepsilon_{X|Y}$ as a function of $H(X|Y)$ when M is finite:

$$\varepsilon_{X|Y} \geq \frac{1}{6} \frac{H(X|Y)}{\log M + \log\log M - \log H(X|Y)}, \tag{115}$$

and the right side of (115) does not depend on the base of the logarithm. The bound in (111) becomes trivial in the limit where $\alpha \uparrow 1$ since, for any fixed $\rho > 0$, (47) implies that $u_M\left(\frac{\alpha\rho}{1-\alpha}\right) \to 1$, and therefore the lower bound on $\varepsilon_{X|Y}$ tends to zero in this case. Nevertheless, numerical experimentation shows that the convergence of this bound in (111) to zero is only affected by values of α very close to 1, as it is illustrated in Example 6 with a comparison to Shannon's lower bound in (115).

Example 6: Let X and Y be random variables taking values on $\mathcal{X} = \{1, 2, 3, 4\}$, and let

$$\left[P_{XY}(x,y)\right]_{(x,y)\in\mathcal{X}^2} = \frac{1}{100}\begin{pmatrix} 9 & 3 & 4 & 9 \\ 9 & 9 & 3 & 4 \\ 4 & 9 & 9 & 3 \\ 3 & 4 & 9 & 9 \end{pmatrix}. \tag{116}$$

It can be verified from (89) that $\varepsilon_{X|Y} = \frac{16}{25} = 0.640$. Table III shows a slight advantage of the lower bound

TABLE III. EXAMPLE 6: LOWER BOUNDS ON $\varepsilon_{X|Y}$.

α	(111)	(113)	(114)
-1	0.463	–	0.447
$-\frac{1}{2}$	0.475	–	0.355
$-\frac{1}{4}$	0.482	–	0.206
$\frac{1}{5}$	0.494	0.523	–
$\frac{1}{2}$	0.502	0.530	–
$\frac{4}{5}$	0.510	0.536	–

in (113) over (111) for $\alpha \in (0,1)$, and a superiority of the lower bound in (111) over (114) for some negative values of α.

In view of Remark 18, the lower bound in (111) for α close to 1 is compared with Shannon's lower bound in (115). For $\alpha = 0.99$, the lower bound in (111) is equal to 0.515 (note that it is slightly looser than (113), which is equal to 0.540); on the other hand, the lower bound in (115) is equal to 0.146.

Theorem 9 establishes relationships between the ρ-th moment of the optimal guessing function, for fixed $\rho > 0$, and the MAP error probability. This characterizes the exact locus of their attainable values, as it is shown in Figure 2 for $\rho = 1$. A natural question is whether the MAP error probability can be uniquely determined on the basis of the knowledge of these ρ-th moments for all $\rho > 0$. The following result answers this question in the affirmative, also suggesting an easy way to determine the MAP error probability on the basis of the knowledge of these ρ-th moments at an arbitrarily small right neighborhood of $\rho = 0$.

Theorem 11: Let X and Y be discrete random variables taking values on sets $\mathcal{X} = \{1, \dots, M\}$ and $\mathcal{Y}$, respectively. For an integer $k \geq 0$, denote

$$z_k = \frac{\mathrm{d}^k}{\mathrm{d}\rho^k} \, \mathbb{E}[g^{\rho}_{X|Y}(X|Y)]\Big|_{\rho=0}. \tag{117}$$

Then,

$$\varepsilon_{X|Y} = 1 - \frac{1}{c_M} \begin{vmatrix} z_0 & 1 & \cdots & 1 \\ z_1 & \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ \vdots & \vdots & \vdots & \vdots \\ z_{M-1} & \log_{\mathrm{e}}^{M-1} 2 & \cdots & \log_{\mathrm{e}}^{M-1} M \end{vmatrix} \tag{118}$$

with

$$c_M = \begin{cases} \log_{\mathrm{e}} 2, & M = 2, \\ \displaystyle\prod_{k=2}^{M} \log_{\mathrm{e}} k \prod_{2 \leq i < j \leq M} \log_{\mathrm{e}}\Big(\frac{j}{i}\Big), & M \geq 3. \end{cases} \tag{119}$$

Proof: Let $x_i(y) \in \mathcal{X}$ be the element that satisfies (95) for $i \in \{1, \dots, M\}$ and $y \in \mathcal{Y}$, and consider the ranking function $g_{X|Y}$ that satisfies

$$g_{X|Y}(x_i(y)|y) = i \tag{120}$$

for all i and y as above. Then,

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] = \sum_{y\in\mathcal{Y}} \left\{ P_Y(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\, g^{\rho}_{X|Y}(x|y) \right\} \tag{121}$$

$$= \sum_{y\in\mathcal{Y}} \left\{ P_Y(y) \sum_{i=1}^{M} P_{X|Y}(x_i(y)|y)\, i^{\rho} \right\} \tag{122}$$

where (121) and (122) follow from (96) and (120), respectively. Swapping the order of summation in (122) yields

$$\mathbb{E}[g^{\rho}_{X|Y}(X|Y)] = \sum_{i=1}^{M} \left\{ i^{\rho} \sum_{y\in\mathcal{Y}} P_{X,Y}(x_i(y), y) \right\} \tag{123}$$

$$= \sum_{i=0}^{M-1} (i+1)^{\rho}\, u_i \tag{124}$$

where u_i is the inner sum in the right side of (123) with i replaced by $i+1$. Taking the k-th derivative of (124) at $\rho = 0$, and recalling (117) yields

$$z_k = \begin{cases} \displaystyle\sum_{i=0}^{M-1} u_i, & k = 0, \\ \displaystyle\sum_{i=1}^{M-1} u_i \log_{\mathrm{e}}^k (i+1), & k \in \{1, \ldots, M-1\}, \end{cases} \tag{125}$$

which gives the set of M linear equations

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ 0 & \vdots & \vdots & \vdots \\ 0 & \log_{\mathrm{e}}^{M-1} 2 & \cdots & \log_{\mathrm{e}}^{M-1} M \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{M-1} \end{pmatrix} = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{M-1} \end{pmatrix} \tag{126}$$

with the M unknown variables $\mathbf{u}^{\top} = (u_0, \ldots, u_{M-1})$. The equations in (126) are linearly independent since by Lemma 14 in Appendix D

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ 0 & \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ 0 & \vdots & \vdots & \vdots \\ 0 & \log_{\mathrm{e}}^{M-1} 2 & \cdots & \log_{\mathrm{e}}^{M-1} M \end{vmatrix} = c_M \neq 0 \tag{127}$$

with c_M as defined in (119), so u_0 in (126) is uniquely determined. We have

$$\varepsilon_{X|Y} = 1 - \sum_{y\in\mathcal{Y}} P_{X,Y}(x_1(y), y) \tag{128}$$

$$= 1 - u_0 \tag{129}$$

where (128) follows from (89) and since, by definition, $x_1(y)$ is a mode of $P_{X|Y}(\cdot|y)$ for all $y \in \mathcal{Y}$; (129) holds by the definition of u_0 as the inner sum in the right side of (123) with $i = 1$. Finally, (118) follows from (126)–(129) and Cramer's rule. ∎

Remark 19: Theorem 11 can be, alternatively, first proved without side information. Then, by replacing P_X by $P_{X|Y}(\cdot|y)$, (128)–(129) hold for $\varepsilon_{X|Y=y}$; finally, it holds by averaging in view of the linearity of $\mathbf{u}$ in $\mathbf{z}$.

IV. Non-Asymptotic Bounds for Optimal Fixed-to-Variable Lossless Compression

This section applies the improved bounds on guessing moments in Section III to analyze non-prefix one-to-one binary optimal codes, which do not satisfy Kraft's inequality. These are one-shot codes that assign distinct codewords to source strings; their average length per source symbol, which is smaller than the Shannon entropy of the source, is analyzed in [18, Section 7] and [48]. Preliminary material is introduced in Section IV-A, improved bounds on the distribution of optimal codeword lengths are derived in Section IV-B, and improved non-asymptotic bounds for fixed-to-variable codes are derived in Section IV-C.

A. Basic setup, notation and preliminaries

Definition 6: A variable-length lossless compression binary code for a discrete set $\mathcal{X}$ is an injective mapping:

$$f\colon \mathcal{X} \to \{0,1\}^* = \{\varnothing, 0, 1, 00, 01, 10, 11, 000, \ldots\} \tag{130}$$

where $f(x)$ is the codeword which is assigned to $x \in \mathcal{X}$; the length of this codeword is denoted by $\ell(f(x))$ where $\ell\colon \{0,1\}^* \to \{0,1,2,\ldots\}$ with the convention that $\ell(\varnothing) = 0$.

Definition 7: [50] A variable-length lossless source code is *compact* whenever it contains a codeword only if all shorter codewords also belong to the code.

Definition 8: [50] Given a probability mass function P_X on $\mathcal{X}$, a variable-length lossless source code is *P_X-efficient* if for all $(a,b) \in \mathcal{X}^2$,

$$\ell(f(a)) < \ell(f(b)) \Longrightarrow P_X(a) \geq P_X(b). \tag{131}$$

Definition 9: [50] Given a probability mass function P_X on $\mathcal{X}$, a variable-length lossless source code is *P_X-optimal* if it is both compact and P_X-efficient.

The optimality in Definition 9 is justified in Proposition 1. Let $f_X^*\colon \mathcal{X} \to \{0,1\}^*$ be a P_X-optimal variable-length lossless source code. If $|\mathcal{X}| < \infty$, then

a) $\varnothing$ is assigned to the most likely element in $\mathcal{X}$.
b) All the 2^ℓ binary strings of length ℓ are assigned to the 2^ℓ-th through $(2^{\ell+1}-1)$-th most likely elements with $\ell \in \{1, \ldots, \lfloor \log_2(1+|\mathcal{X}|) \rfloor - 1\}$. For example, 0 and 1 (or 1 and 0) are assigned, respectively, to the second and third most likely elements in $\mathcal{X}$.
c) If $\log_2(1+|\mathcal{X}|)$ is not an integer, then codewords of length $\lfloor \log_2(1+|\mathcal{X}|) \rfloor$ are assigned to each of the remaining $1 + |\mathcal{X}| - 2^{\lfloor \log_2(1+|\mathcal{X}|) \rfloor}$ elements in $\mathcal{X}$.

As long as $|\mathcal{X}| > 1$, there is more than one P_X-optimal code since compactness and P_X-efficiency are preserved by swapping codewords of the same length (and, if $|\mathcal{X}| = 2$, then the second most likely element can be either assigned 0 or 1). In the presence of ties among probabilities, the value of $\ell(f_X^*(x))$ for some $x \in \mathcal{X}$ may depend on the choice of f_X^*. The following result provides several relevant properties of optimal codes.

Proposition 1: ([26], [50]) Fix a probability mass function P_X on a finite set $\mathcal{X}$. The following results hold for P_X-optimal codes $f_X^*\colon \mathcal{X} \to \{0,1\}^*$:

a) The distribution of $\ell(f_X^*(X))$ is invariant to the actual choice of f_X^*, and it only depends on P_X.
b) For every lossless data compression code f, and for all $r \geq 0$,

$$\mathbb{P}\big[\ell\big(f(X)\big) \leq r\big] \leq \mathbb{P}\big[\ell\big(f_X^*(X)\big) \leq r\big]. \tag{132}$$

Furthermore, the inequality in (132) is strict for some $r \geq 0$ if f is not P_X-optimal.

c)

$$\sum_{x\in\mathcal{X}} 2^{-\ell(f_X^*(x))} \le \log_2(1+|\mathcal{X}|) \tag{133}$$

with equality if and only if $|\mathcal{X}|+1$ is a positive integral power of 2. Furthermore, all compact codes for $\mathcal{X}$ achieve the same value of $\sum\limits_{x\in\mathcal{X}} 2^{-\ell(f(x))}$, which is larger than that achieved by a non-compact code.

Definition 10: The *cumulant generating function* of the codeword lengths of P_X-optimal binary codes is given by

$$\Lambda^*(\rho) \triangleq \log \mathbb{E}\big[2^{\rho\,\ell(f_X^*(X))}\big], \quad \rho\in\mathbb{R}. \tag{134}$$

Remark 20: (134) is actually a scaled cumulant generating function. The cumulant generating function of a random variable X is given by

$$\Lambda_X(\rho) = \log_{\mathrm{e}} \mathbb{E}\big[e^{\rho X}\big], \quad \rho\in\mathbb{R} \tag{135}$$

whereas, following Campbell [13], it is more natural to study the function given by

$$\widetilde{\Lambda}_X(\rho) = \log \mathbb{E}\big[2^{\rho X}\big]. \tag{136}$$

Note, however, that (135) and (136) satisfy

$$\widetilde{\Lambda}_X(\rho) = \Lambda_X(\rho \log_{\mathrm{e}} 2)\, \log \mathrm{e}, \tag{137}$$

which implies that they can be obtained from each other by proper linear scalings of the axes.

As mentioned in the introduction, the cumulant generating function of the codeword lengths provides an important design criterion. In particular, it yields the average length via the equality

$$\lim_{\rho\to 0} \frac{\Lambda^*(\rho)}{\rho} = \mathbb{E}[\ell(f_X^*(X))]. \tag{138}$$

Theorem 12: [16, Theorem 1] If $\rho\in(-\infty,-1]$, then

$$H_\infty(X) - \log\log_2(1+|\mathcal{X}|) \le -\Lambda^*(\rho) \le H_\infty(X), \tag{139}$$

and, if $\rho\in(-1,0)\cup(0,\infty)$, then

$$H_{\frac{1}{1+\rho}}(X) - \log\log_2(1+|\mathcal{X}|) \le \frac{\Lambda^*(\rho)}{\rho} \le H_{\frac{1}{1+\rho}}(X). \tag{140}$$

By invoking the Chernoff bound and using Theorem 12, the following result holds.

Theorem 13: [16, Theorem 2] For all $H(X) < R < \log|\mathcal{X}|$

$$\log \frac{1}{\mathbb{P}[\ell(f_X^*(X)) \ge R]} \ge \sup_{\rho>0}\Big\{\rho R - \rho H_{\frac{1}{1+\rho}}(X)\Big\} \tag{141}$$

$$= D(X_\alpha \| X) \tag{142}$$

where $\alpha\in(0,1)$ is a function of R chosen so that $R = H(X_\alpha)$, and X_α has the scaled probability mass function

$$P_{X_\alpha}(x) = \frac{P_X^\alpha(x)}{\sum\limits_{a\in\mathcal{X}} P_X^\alpha(a)}, \quad x\in\mathcal{X}. \tag{143}$$

B. Improved bounds on the distribution of the optimal code lengths

We derive bounds on the cumulant generating function and the complementary cumulative distribution of optimal lengths for lossless compression of a random variable X which takes values on a finite set $\mathcal{X}$. These bounds improve those in Theorems 12 and 13, and in Section IV-C we use them to derive non-asymptotic bounds for optimal fixed-to-variable lossless codes.

We start by generalizing [16, Lemma 1] from $\beta = 1$ to arbitrary $\beta \in \mathbb{R}$.

Lemma 5: For an optimal binary code, and for all $\beta \in \mathbb{R}$

$$\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell(f_X^*(x))} = \begin{cases} (2^{\Delta}-1)s_\beta^m + \dfrac{1-s_\beta^m}{1-s_\beta}, & \beta \neq 1 \\ m + 2^{\Delta} - 1, & \beta = 1 \end{cases} \tag{144}$$

$$\triangleq t(\beta, |\mathcal{X}|) \tag{145}$$

where

$$s_\beta = 2^{1-\beta}, \tag{146}$$

$$m = \big\lfloor \log_2(1+|\mathcal{X}|) \big\rfloor, \tag{147}$$

$$\Delta = \log_2(1+|\mathcal{X}|) - m \in [0,1). \tag{148}$$

Proof: From Definition 9, there are 2^ℓ elements of $\mathcal{X}$ which are assigned codewords of length ℓ with $\ell \in \{0,1,\dots,m-1\}$, and $|\mathcal{X}| - 2^m + 1$ elements which are assigned codewords of length m. Hence, for $\beta \in \mathbb{R}$,

$$\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell(f_X^*(x))} = (|\mathcal{X}| - 2^m + 1)\, 2^{-\beta m} + \sum_{\ell=0}^{m-1} 2^{\ell}\, 2^{-\beta \ell}. \tag{149}$$

In view of (147) and (148), for $\beta = 1$, (144) follows from (149). If $\beta \neq 1$, then from (146)–(149)

$$\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell(f_X^*(x))} = \big[(1+|\mathcal{X}|)2^{-m} - 1\big]\; 2^{(1-\beta)m} + \frac{1-2^{(1-\beta)m}}{1-2^{1-\beta}} \tag{150}$$

$$= (2^{\Delta}-1)s_\beta^m + \frac{1-s_\beta^m}{1-s_\beta}. \tag{151}$$

∎

Lemma 6: Let X be a random variable taking values on a finite set $\mathcal{X}$, and let $\rho \neq 0$. Then, for an optimal binary code,

$$\frac{1}{\rho}\,\log \mathbb{E}\big[2^{\rho\,\ell(f_X^*(X))}\big] \geq \sup_{\beta\in(-\rho,\infty)\setminus\{0\}} \frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X) - \log\, t(\beta, |\mathcal{X}|)\right], \tag{152}$$

where $t(\cdot,\cdot)$ is defined in (145).

Proof: Let $g\colon \mathcal{X} \to \{1,2,4,8,\dots\}$ be defined by $g(x) = 2^{\ell(f_X^*(x))}$ for all $x \in \mathcal{X}$. The result in (152) follows by combining (33) and Lemma 5. ∎

Lemma 7: Let X be a random variable taking values on a finite set $\mathcal{X}$, and let g_X be a ranking function of X. Then, for every optimal binary code and for all $\rho > 0$,

$$2^{-\rho}\,\mathbb{E}[g_X^\rho(X)] < \mathbb{E}\big[2^{\rho\,\ell(f_X^*(X))}\big] \leq \mathbb{E}[g_X^\rho(X)]. \tag{153}$$

Proof: Let x_k be the k-th most probable outcome of X under a given tie-breaking rule. For an optimal compression code, the length of the assigned codeword of the element $x_k \in \mathcal{X}$ satisfies $\ell(f^*_X(x_k)) = \lfloor \log_2 k \rfloor$ ([26, (15)]). Hence, for $\rho > 0$,

$$\mathbb{E}\big[2^{\rho\, \ell(f^*_X(X))}\big] = \sum_k P_X(x_k)\, 2^{\rho\, \lfloor \log_2 k \rfloor} \tag{154}$$

$$\leq \sum_k P_X(x_k)\, k^{\rho} \tag{155}$$

$$= \mathbb{E}\big[g_X^{\rho}(X)\big]. \tag{156}$$

This proves the right side of (153). To show the left side of (153), note that

$$\lfloor \log_2 k \rfloor > \log_2 k - 1 \tag{157}$$

which implies from (154), (156) and (157) that for $\rho > 0$

$$\mathbb{E}\big[2^{\rho\, \ell(f^*_X(X))}\big] > \sum_k P_X(x_k)\, 2^{\rho\, (\log_2 k - 1)} \tag{158}$$

$$= 2^{-\rho}\, \mathbb{E}\big[g_X^{\rho}(X)\big]. \tag{159}$$

■

Theorem 14: Let X be a random variable taking values on a finite set $\mathcal{X}$. Then, for every optimal binary code, the cumulant generating function in (134) satisfies

$$\sup_{\beta \in (-\rho, \infty) \setminus \{0\}} \frac{1}{\beta} \left[H_{\frac{\beta}{\beta+\rho}}(X) - \log t(\beta, |\mathcal{X}|) \right]$$

$$\leq \frac{\Lambda^*(\rho)}{\rho} \tag{160}$$

$$\leq H_{\frac{1}{1+\rho}}(X) + \frac{1}{\rho} \log\!\left(\frac{1}{1+\rho} \left[1 - \exp\left(-\rho H_{\frac{1}{1+\rho}}(X) \right) \right] + \exp\left((\rho - 1)^+ H_{\frac{1}{\rho}}(X) - \rho H_{\frac{1}{1+\rho}}(X) \right) \right), \tag{161}$$

for all $\rho > 0$, where $t(\cdot)$ is given in (145). Moreover, (160) also holds for $\rho < 0$.

Proof: The lower bound in the left side of (160) is Lemma 6, and the upper bound in the right side of (161) follows from Theorem 4 and Lemma 7. ■

Remark 21: For $\rho \in (-1, 0) \cup (0, \infty)$, loosening the bound in the left side of (160) by the sub-optimal choice of $\beta = 1$ and invoking $t(1, |\mathcal{X}|) \leq \log_2(1 + |\mathcal{X}|)$ (in view of Lemma 5, and since $2^x - 1 \leq x$ for $x \in [0, 1]$) recovers the lower bound in (140).

Remark 22: In view of (70), the second term in the right side of (161) is non-positive for all $\rho \geq 1$; due to the non-negativity of the Rényi entropy, this also holds for $\rho \in (0, 1)$. Hence, for $\rho > 0$, the upper bound in (161) improves the bound in the right side of (140).

The Chernoff bound and (161) readily yield the following lower bound.

Theorem 15: Under the assumptions in Theorem 14, for $R < \log |\mathcal{X}|$,

$$\log \left(\frac{1}{\mathbb{P}\big[\ell(f^*_X(X)) > R\big]} \right) \geq \sup_{\rho > 0} \Bigg\{ \rho R - \rho H_{\frac{1}{1+\rho}}(X) - \log\!\bigg(\frac{1}{1+\rho} \left[1 - \exp\left(-\rho H_{\frac{1}{1+\rho}}(X) \right) \right] + \exp\left((\rho - 1)^+ H_{\frac{1}{\rho}}(X) - \rho H_{\frac{1}{1+\rho}}(X) \right) \bigg) \Bigg\}. \tag{162}$$

Remark 23: The bound in (162) is strictly tighter than the right side in (142) unless X is deterministic. To show this, note that in view of Remark 22 and (141)–(142), the bound in (162) cannot be looser than the right side in (142). Moreover, since the function $\rho R - \rho H_{\frac{1}{1+\rho}}(X)$ is concave in $\rho > 0$, its local (and hence global) maximum is achieved at $\rho^* > 0$; from the proof of [16, Theorem 2], this value of ρ^* satisfies

$$\alpha = \frac{1}{1+\rho^*}, \quad H(X_\alpha) = R, \tag{163}$$

$$\rho^* R - \rho^* H_{\frac{1}{1+\rho^*}}(X) = D(X_\alpha \| X) \tag{164}$$

where X_α has the scaled probability mass function in (143). Hence, by replacing the supremum over $\rho > 0$ with the value at $\rho = \rho^*$, (164) yields the following loosened bound:

$$\log\left(\frac{1}{\mathbb{P}\big[\ell(f_X^*(X)) > R\big]}\right) \geq D(X_\alpha \| X) - \log\Bigg(\frac{1}{1+\rho^*}\left[1 - \exp\left(-\rho^* H_{\frac{1}{1+\rho^*}}(X)\right)\right] + \exp\left((\rho^*-1)^+ H_{\frac{1}{\rho^*}}(X) - \rho^* H_{\frac{1}{1+\rho^*}}(X)\right)\Bigg). \tag{165}$$

Hence, it is enough to prove that (165) is strictly tighter than (142). This follows from the strict inequality in (70) whenever X is non-deterministic, thus implying that, for all $\rho > 0$,

$$\sum_{x \in \mathcal{X}} P_X^{\frac{1}{1+\rho}}(x) = \exp\left(\tfrac{\rho}{1+\rho}\, H_{\frac{1}{1+\rho}}(X)\right) > 1. \tag{166}$$

C. Non-asymptotic bounds for fixed-to-variable lossless source codes

We consider the fixed-to-variable-length lossless compression in Definition 9 where the object to be compressed $x^n = (x_1, \ldots, x_n) \in \mathcal{A}^n$ is a string of length n (n is known to both encoder and decoder), whose letters are drawn from a finite alphabet $\mathcal{X}$ according to the probability mass function $P_{X^n}(x^n) = \prod_{i=1}^n P_X(x_i)$ for all $x^n \in \mathcal{A}^n$. We consider the following non-asymptotic measures for optimal fixed-to-variable lossless compression:

- The cumulant generating function of the codeword lengths is given by

$$\Lambda_n(\rho) := \frac{1}{n}\log \mathbb{E}\left[2^{\rho\, \ell(f_{X^n}^*(X^n))}\right], \quad \rho \in \mathbb{R}. \tag{167}$$

- The non-asymptotic version of the source reliability function is given by

$$E_n(R) = \frac{1}{n}\log\left(\frac{1}{\mathbb{P}\big[\frac{1}{n}\,\ell(f_{X^n}^*(X^n)) \geq R\big]}\right). \tag{168}$$

Theorem 16: Consider a memoryless and stationary source of finite alphabet $\mathcal{A}$, and let $f_{X^n}^* \colon \mathcal{A}^n \to \{0,1\}^*$ be an optimal compression code. Then, the following bounds hold:

a) For all $\rho > 0$

$$\begin{aligned} &\sup_{\beta \in (-\rho,\infty)\setminus\{0\}} \frac{\rho}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X) - \frac{1}{n}\log t(\beta, |\mathcal{A}|^n)\right] \\ &\leq \Lambda_n(\rho) \qquad (169) \\ &\leq \rho H_{\frac{1}{1+\rho}}(X) + \frac{1}{n}\log\Bigg(\frac{1}{1+\rho}\left[1 - \exp\left(-n\rho H_{\frac{1}{1+\rho}}(X)\right)\right] + \exp\left(n\left[(\rho-1)^+ H_{\frac{1}{\rho}}(X) - \rho H_{\frac{1}{1+\rho}}(X)\right]\right)\Bigg) \qquad (170) \end{aligned}$$

where $t(\cdot)$ is as defined in (145).

b) For $R < \log|\mathcal{A}|$

$$E_n(R) \geq \sup_{\rho>0}\Bigg\{\rho R - \rho H_{\frac{1}{1+\rho}}(X) - \frac{1}{n}\log\Bigg(\frac{1}{1+\rho}\left[1-\exp\left(-n\rho H_{\frac{1}{1+\rho}}(X)\right)\right] + \exp\left(n\left[(\rho-1)^+ H_{\frac{1}{\rho}}(X) - \rho H_{\frac{1}{1+\rho}}(X)\right]\right)\Bigg)\Bigg\}. \tag{171}$$

Proof: Items a) and b) follow, respectively, from Theorems 14 and 15 with $|\mathcal{X}|$ replaced by $|\mathcal{A}|^n$, and since $H_\alpha(X^n) = nH_\alpha(X)$ holds for i.i.d. random variables and for all $\alpha > 0$ (see Lemma 1). ■

Remark 24: The non-asymptotic bounds on the cumulant generating function in (169)–(170) recover the known asymptotic result in [16, (29)] where for all $\rho > 0$

$$\Lambda(\rho) := \lim_{n\to\infty} \Lambda_n(\rho) = \rho H_{\frac{1}{1+\rho}}(X), \tag{172}$$

which, incidentally, coincides with Arikan's asymptotic fundamental limit for $\lim_{n\to\infty} \frac{1}{n}\log\mathbb{E}[g^{\rho}_{X^n}(X^n)]$ when X^n is i.i.d. [1]. To this end, note that $\lim_{n\to\infty}\frac{1}{n}\log t(1,|\mathcal{A}|^n) = 0$, and selecting $\beta = 1$ in the left side of (169) yields

$$\varliminf_{n\to\infty} \Lambda_n(\rho) \geq \rho H_{\frac{1}{1+\rho}}(X). \tag{173}$$

Moreover, since $H_\alpha(X)$ is monotonically non-increasing in α,

$$\rho H_{\frac{1}{1+\rho}}(X) - (\rho-1)^+ H_{\frac{1}{\rho}}(X) \geq \min\{\rho,1\}\, H_{\frac{1}{\rho}}(X) \tag{174}$$

and, if X is non-deterministic, then (170) and (174) yield

$$\varlimsup_{n\to\infty} \Lambda_n(\rho) \leq \rho H_{\frac{1}{1+\rho}}(X), \tag{175}$$

recovering (172) from (173) and (175). Furthermore, (171) and (174) imply that

$$E(R) := \varliminf_{n\to\infty} E_n(R) \tag{176}$$

$$\geq \sup_{\rho>0}\Big\{\rho R - \rho H_{\frac{1}{1+\rho}}(X)\Big\}. \tag{177}$$

Although, as noted in Remark 24, the improvement in the bounds afforded by Theorem 16 washes out asymptotically, the following example illustrates the improvement in the non-asymptotic regime.

Example 7: Consider a discrete memoryless source which emits a string of n letters from the alphabet $\mathcal{A} = \{a,b,c\}$ with $P_X(a) = \frac{4}{7}$, $P_X(b) = \frac{2}{7}$ and $P_X(c) = \frac{1}{7}$.

The bounds on the cumulant generating function in [16, Theorem 1] (see (140)) are given by

$$\rho\, H_{\frac{1}{1+\rho}}(X) - \tfrac{\rho}{n}\log\log_2(1+|\mathcal{A}|^n) \leq \Lambda_n(\rho) \leq \rho H_{\frac{1}{1+\rho}}(X) \tag{178}$$

for $\rho > 0$. Figure 3 compares (178) with the improved bounds in (169)–(170). For $n = 10$, the upper and lower bounds are compared to the exact normalized cumulant (see the left plot in Figure 3); this indicates that the lower bound in (169) can be tight even for small values of n. The match between the upper and lower bounds in (169)–(170) improves by increasing n, and the tightening obtained by the lower bound in (169) can be significant for small values of n.

Lower bounds on the non-asymptotic source reliability function $E_n(R)$ are plotted in Figure 4. In this figure, the lower bound in Theorem 13 is compared to the tighter bound in (171), illustrating the improvement for small to moderate values of n.

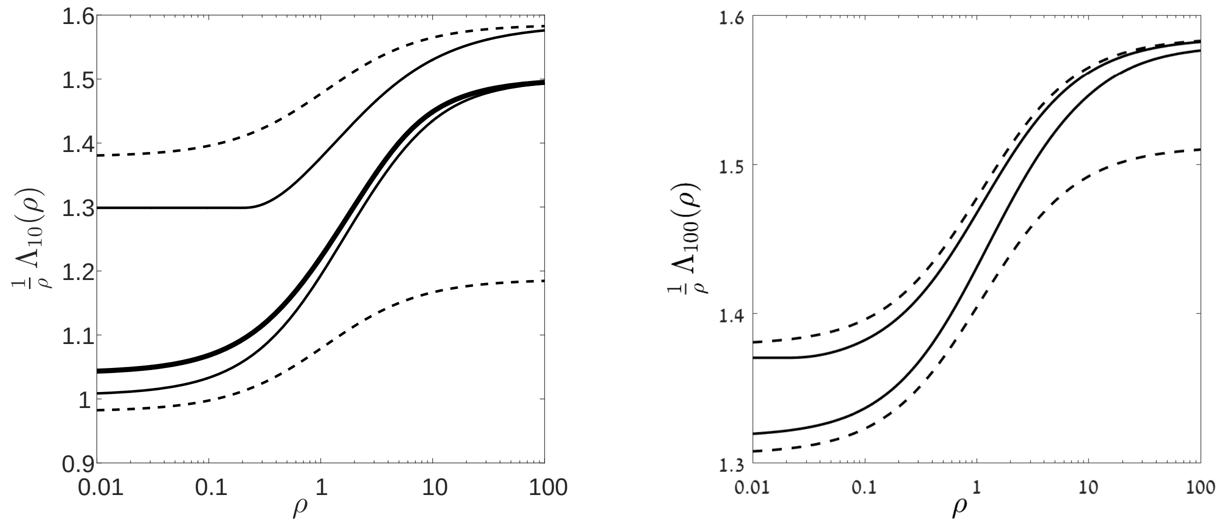

Fig. 3. Bounds on the normalized cumulant generating function, $\frac{\Lambda_n(\rho)}{\rho}$ (in bits), of the codeword lengths of optimal lossless compression of strings of length n emitted from the discrete memoryless source in Example 7. The dashed lines are the bounds in (178), and the thin solid lines refer to the improved bounds in (169)–(170). The left plot corresponds to $n = 10$, in which case we can compute the exact normalized cumulant (thick solid curve). The right plot corresponds to $n = 100$.

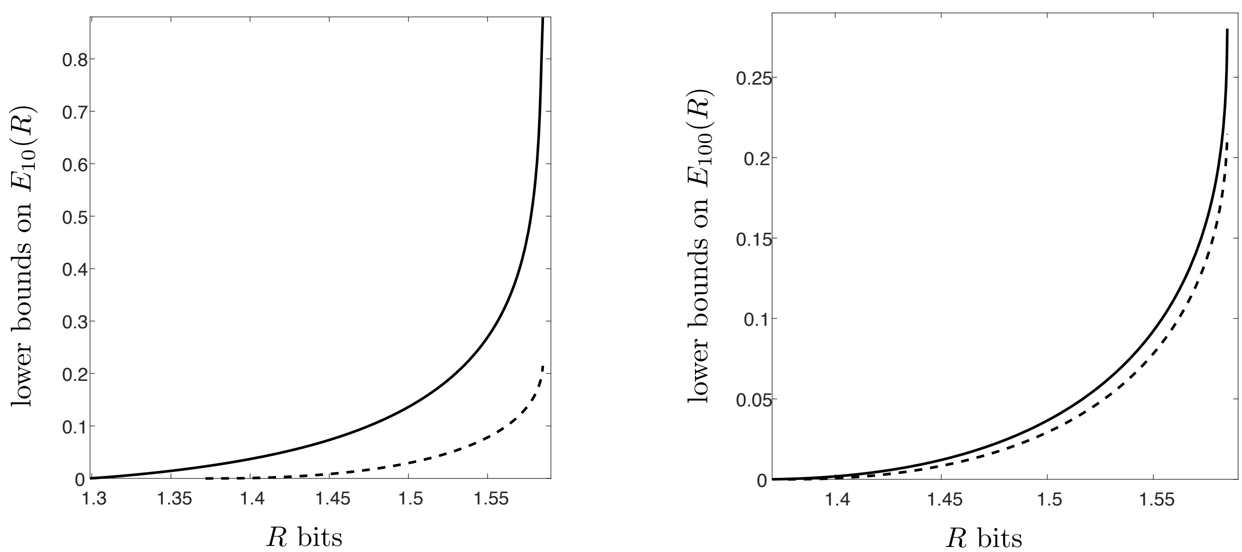

Fig. 4. Lower bounds on the non-asymptotic reliability function, $E_n(R)$ (base 2), for the discrete memoryless source in Example 7. In each plot, the dashed curve refers to the lower bound on $E_n(R)$ in Theorem 13, and the solid curve refers to the tighter lower bound in (171). The left and right plots correspond to $n = 10$ and $n = 100$, respectively.

V. Lower Bounds for Variable-Length Source Coding Allowing Errors

Following a recent study by Kuzuoka [28], this section applies our bounding techniques to derive improved lower bounds on the cumulant generating function of the codeword lengths for variable-length source coding allowing errors (which, in contrast to the conventional fixed-to-fixed paradigm, are not necessarily detectable by the decoder) by means of the smooth Rényi entropy in Definition 5.

In contrast to [28], the bounds in this subsection are derived for source codes without the prefix condition when either the maximal or average decoding error probabilities are limited not to exceed a given value $\varepsilon \in [0, 1)$.

Theorem 17: Let X take values on a finite set $\mathcal{X}$, and let $f\colon \mathcal{X} \to \mathcal{C}$ be an encoder (possibly stochastic) with a finite codebook $\mathcal{C} \subseteq \{0,1\}^*$, and let $\ell \colon \mathcal{C} \to \{0, 1, \ldots, \}$ be the length function of the codewords in $\mathcal{C}$. Fix $\varepsilon \in [0, 1)$ and $\rho > 0$.

1) If the *average* decoding error probability cannot be larger than ε, then

$$\frac{1}{\rho} \log \mathbb{E}\left[2^{\rho\, \ell(f(X))}\right] \geq \sup_{\beta > 0} \frac{1}{\beta}\left[H^{(\varepsilon)}_{\frac{\beta}{\beta+\rho}}(X) - \log t(\beta, |\mathcal{X}|)\right] \tag{179}$$

where $H^{(\varepsilon)}_{\alpha}(X)$ is the ε-smooth Rényi entropy of order α, and $t(\cdot)$ is given in (145).

2) If the *maximal* decoding error probability cannot be larger than ε, then also

$$\frac{1}{\rho} \log \mathbb{E}\left[2^{\rho\, \ell(f(X))}\right] \geq \sup_{\beta \in (-\rho, 0)} \frac{1}{\beta}\left[H_{\frac{\beta}{\beta+\rho}}(X) - \log t(\beta, |\mathcal{X}|)\right] - \frac{1}{\rho} \log \frac{1}{1-\varepsilon}. \tag{180}$$

Proof: We first derive (179), and then rely on its proof for the derivation of (180).

1) Let $Q \colon \mathcal{X} \to \mathcal{C}$ denote a transition probability matrix such that $Q(c|x)$ is the probability that a codeword $c \in \mathcal{C}$ is assigned to $x \in \mathcal{X}$ by the stochastic encoder. Let $\psi \colon \mathcal{C} \to \mathcal{X}$ be the deterministic decoding function. The average decoding error probability is given by

$$P_{\mathrm{e}} = \mathbb{P}[X \neq \psi(f(X))] \tag{181}$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \sum_{c \colon \psi(c) \neq x} Q(c|x). \tag{182}$$

In order to minimize P_{e} for a given (stochastic) encoder, the decision relies on a MAP decoder:

$$\psi(c) \in \arg\max_{x \in \mathcal{X}} Q(c|x) P_X(x), \quad c \in \mathcal{C} \tag{183}$$

where ties are arbitrarily resolved. Let

$$\gamma(x) \triangleq Q\big(\psi^{-1}(x)|x\big) \in (0, 1] \tag{184}$$

denote the probability that $x \in \mathcal{X}$ is assigned to a codeword that is decoded into x. Since the average decoding error probability satisfies $P_{\mathrm{e}} \leq \varepsilon$, it follows from (182) and (184) that

$$\sum_{x \in \mathcal{X}} P_X(x) \gamma(x) \geq 1 - \varepsilon. \tag{185}$$

For all $x \in \mathcal{X}$, let

$$\ell_\psi(x) \triangleq \min_{c \in \psi^{-1}(x)} \ell(c) \tag{186}$$

be the minimal length of the codewords for which the decoder chooses x, and let

$$\mu(x) \triangleq P_X(x) \gamma(x), \quad x \in \mathcal{X}. \tag{187}$$

From (184), (185) and (187)

$$0 \leq \mu(x) \leq P_X(x), \quad x \in \mathcal{X}, \tag{188}$$

$$\sum_{x \in \mathcal{X}} \mu(x) \geq 1 - \varepsilon, \tag{189}$$

which, by (19), yields

$$\mu \in \mathcal{B}^{(\varepsilon)}(P_X). \tag{190}$$

For all $\rho > 0$

$$\mathbb{E}\Big[2^{\rho\,\ell(f(X))}\Big] = \sum_{x\in\mathcal{X}} P_X(x) \sum_{c\in\mathcal{C}} Q(c|x)\,2^{\rho\,\ell(c)} \tag{191}$$

$$\geq \sum_{x\in\mathcal{X}} P_X(x) \sum_{c\in\psi^{-1}(x)} Q(c|x)\,2^{\rho\,\ell(c)}, \tag{192}$$

and, for every $x \in \mathcal{X}$,

$$\sum_{c\in\psi^{-1}(x)} Q(c|x)\,2^{\rho\,\ell(c)} \geq \sum_{c\in\psi^{-1}(x)} Q(c|x)\,2^{\rho\,\ell_\psi(x)} \tag{193}$$

$$= \gamma(x)\,2^{\rho\,\ell_\psi(x)} \tag{194}$$

where (193) holds due to (186), and (194) follows from (184). Hence, for all $\rho > 0$,

$$\mathbb{E}\Big[2^{\rho\,\ell(f(X))}\Big] \geq \sum_{x\in\mathcal{X}} P_X(x)\gamma(x)\,2^{\rho\,\ell_\psi(x)} \tag{195}$$

$$= \sum_{x\in\mathcal{X}} \mu(x)\,2^{\rho\,\ell_\psi(x)} \tag{196}$$

where (195) follows from (191)–(194), and (196) is due to (187). The finite sets $\{\psi^{-1}(x)\}_{x\in\mathcal{X}}$ are disjoint. For $x \in \mathcal{X}$, let $c^*(x) \in \psi^{-1}(x)$ be a codeword which achieves the minimum in the right side of (186), i.e.,

$$\ell_\psi(x) = \ell(c^*(x)). \tag{197}$$

Since the codewords $\{c^*(x)\}_{x\in\mathcal{X}}$ are distinct, it follows from (197) and the proof of Lemma 5 that

$$\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell_\psi(x)} \leq t(\beta, |\mathcal{X}|), \quad \beta \geq 0, \tag{198}$$

and

$$\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell_\psi(x)} \geq t(\beta, |\mathcal{X}|), \quad \beta \leq 0 \tag{199}$$

since any perturbation of the set of codeword lengths of a P_X-optimal code necessarily shifts shorter to longer codewords. Let

$$\alpha \triangleq \frac{\beta+\rho}{\beta} \tag{200}$$

with $\rho > 0$ and $\beta \neq 0$, and define the following probability mass functions on $\mathcal{X}$:

$$R(x) = \frac{\mu^{\frac{1}{\alpha}}(x)}{\sum_{a\in\mathcal{X}} \mu^{\frac{1}{\alpha}}(a)}, \tag{201}$$

$$S(x) = \frac{2^{-\beta\,\ell_\psi(x)}}{\sum_{a\in\mathcal{X}} 2^{-\beta\,\ell_\psi(a)}}. \tag{202}$$

Straightforward calculation shows that

$$D_\alpha(R\|S) = \frac{1}{\alpha-1}\,\log \sum_{x\in\mathcal{X}} R^\alpha(x) S^{1-\alpha}(x) \tag{203}$$

$$= \frac{\beta}{\rho}\,\log\left(\sum_{x\in\mathcal{X}} \mu(x)\,2^{\rho\,\ell_\psi(x)}\right) + \log\left(\sum_{x\in\mathcal{X}} 2^{-\beta\,\ell_\psi(x)}\right) - \frac{\beta+\rho}{\rho}\,\log\left(\sum_{x\in\mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x)\right). \tag{204}$$

We now fix $\beta > 0$, and we proceed to upper bound each of the three terms in (204): the first term is upper bounded using (196), the second term is upper bounded by $\log t(\beta, |\mathcal{X}|)$ in view of (198), and the third term satisfies

$$H^{(\varepsilon)}_{\frac{\beta}{\beta+\rho}}(X) \leq \frac{\beta+\rho}{\rho} \, \log \sum_{x \in \mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x), \tag{205}$$

which holds in view of Definition 5 and (190). Plugging those bounds into the right side of (204), and recalling that $D_\alpha(R \| S) \geq 0$ (this follows from Lemma 2, and since $\alpha > 0$ in (200) for $\beta, \rho > 0$), we obtain

$$\frac{1}{\rho} \, \log \mathbb{E}\left[2^{\rho\, l(f(X))}\right] \geq \frac{1}{\beta} \left[H^{(\varepsilon)}_{\frac{\beta}{\beta+\rho}}(X) - \log t(\beta, |\mathcal{X}|) \right]. \tag{206}$$

Finally, (179) follows by supremizing (206) over the free parameter $\beta > 0$.

2) If the maximal decoding error probability does not exceed ε, then so is the average decoding error probability, and we can rely on the results in Item 1) of this proof. Fix $\beta \in (-\rho, 0)$. In view of (200), $\alpha < 0$ and $D_\alpha(R \| S) \leq 0$ (Lemma 2). From (204), we get

$$\frac{\beta}{\rho} \, \log \left(\sum_{x \in \mathcal{X}} \mu(x) \, 2^{\rho \, \ell_\psi(x)} \right) + \log \left(\sum_{x \in \mathcal{X}} 2^{-\beta \, \ell_\psi(x)} \right) - \frac{\beta+\rho}{\rho} \, \log \left(\sum_{x \in \mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x) \right) \leq 0. \tag{207}$$

Due to the above assumption on the maximal decoding error probability, (184) implies that for every $x \in \mathcal{X}$

$$\gamma(x) \in [1-\varepsilon, 1], \tag{208}$$

and, consequently, (187) implies that

$$(1-\varepsilon) P_X(x) \leq \mu(x) \leq P_X(x), \quad x \in \mathcal{X}. \tag{209}$$

Since $\frac{\beta}{\beta+\rho} < 0$ and $\frac{\rho}{\beta+\rho} > 0$, (2) (with negative orders of the Rényi entropy) and (209) yield

$$\begin{aligned} H_{\frac{\beta}{\beta+\rho}}(X) &\leq \frac{\beta+\rho}{\rho} \, \log \sum_{x \in \mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x) && (210) \\ &\leq H_{\frac{\beta}{\beta+\rho}}(X) - \frac{\beta}{\rho} \, \log \frac{1}{1-\varepsilon}. && (211) \end{aligned}$$

Consequently, we have

$$\begin{aligned} \frac{1}{\rho} \, \log \mathbb{E}\left[2^{\rho\, l(f(X))}\right] &\geq \frac{1}{\rho} \, \log \sum_{x \in \mathcal{X}} \mu(x) \, 2^{\rho \, \ell_\psi(x)} && (212) \\ &\geq \frac{1}{\beta} \left[\frac{\beta+\rho}{\rho} \, \log \left(\sum_{x \in \mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x) \right) - \log \left(\sum_{x \in \mathcal{X}} 2^{-\beta \, \ell_\psi(x)} \right) \right] && (213) \\ &\geq \frac{1}{\beta} \left[\frac{\beta+\rho}{\rho} \, \log \left(\sum_{x \in \mathcal{X}} \mu^{\frac{\beta}{\beta+\rho}}(x) \right) - \log t(\beta, |\mathcal{X}|) \right] && (214) \\ &\geq \frac{1}{\beta} \left[H_{\frac{\beta}{\beta+\rho}}(X) - \frac{\beta}{\rho} \, \log \frac{1}{1-\varepsilon} - \log t(\beta, |\mathcal{X}|) \right] && (215) \end{aligned}$$

where (212) holds due to (196); (213) follows from (207) and since $\beta < 0$; (214) holds due to (199); finally, (215) follows from (211). Finally, (180) follows by supremizing the right side of (215) over the free parameter $\beta \in (-\rho, 0)$.

■

Remark 25: Note that the lower bound in (179) is expressed in terms of the smooth Rényi entropy of orders $\alpha \in (0,1)$, and the lower bound in (180) is expressed in terms of the conventional Rényi entropy of negative orders. The calculation of the bound in (179) relies on Lemma 3a).

Remark 26: The combination of letting $\varepsilon = 0$ in (179) and (180) recovers the result in Lemma 6.

Remark 27: If the maximal decoding error probability cannot be larger than $\varepsilon \in (0,1)$, then neither of the bounds in (179) and (180) is superseded by the other, as can be verified by numerical experimentation.

Remark 28: For prefix codes, Kraft's inequality gives $\sum_{x \in \mathcal{X}} 2^{-\ell_\psi(x)} \leq 1$. Replacing $t(1,|\mathcal{X}|)$ in the right side of (198) by 1 recovers the result in [28, Theorem 2] from (179):

$$\frac{1}{\rho} \log \mathbb{E}\left[2^{\rho\, \ell(f(X))}\right] \geq H^{(\varepsilon)}_{\frac{1}{1+\rho}}(X), \quad \rho > 0 \tag{216}$$

where the average error probability cannot be larger than ε. An analogous result of (216), for a lower bound on the normalized cumulant generating function of the codeword lengths without imposing the prefix condition, is given by

$$\frac{1}{\rho} \log \mathbb{E}\left[2^{\rho\, \ell(f(X))}\right] \geq H^{(\varepsilon)}_{\frac{1}{1+\rho}}(X) - \log t(1,|\mathcal{X}|), \quad \rho > 0 \tag{217}$$

where $t(1,|\mathcal{X}|)$ is given in (145). Since (217) is obtained by replacing the supremization over $\beta > 0$ in the right side of (179) with its value at $\beta = 1$, Theorem 17 gives a better bound than (217).

Example 8: Let $U_1, \ldots, U_n$ be i.i.d. random variables taking values in a set of cardinality 4, and having the probability mass function

$$P_{U_1} = \begin{bmatrix} \frac{4}{10} & \frac{3}{10} & \frac{2}{10} & \frac{1}{10} \end{bmatrix}. \tag{218}$$

Let

$$X_n = \sum_{i=1}^{n} U_i, \quad n \in \mathbb{N}. \tag{219}$$

The probability mass function P_{X_n} is equal to P_{U_1} convolved with itself $n-1$ times, and X_n takes $M_n = 3n+1$ values. Assume that a maximal error probability of $\varepsilon \in [0,1)$ is allowed in decoding X_n. Figure 5 compares the lower bound in (216), for binary prefix codes, with the lower bounds in Theorem 17 and (217) for binary codes without the prefix condition. The upper and lower plots in Figure 5 correspond to $\varepsilon = 0.01$ and $\varepsilon = 0$ (i.e., the lossless case), respectively; the left and right plots correspond, respectively, to $n = 1$ and $n = 100$ (note that, by the central limit theorem, X_n is close to Gaussian for large n). The gain of the combined lower bounds in (179) and (180) (the solid line in each plot) over the bound in (217) (the dashed lower line in the corresponding plot) is illustrated in Figure 5 by comparing the left and right plots; by increasing the value of n, the solid line becomes more steep at sufficiently large values of ρ.

APPENDIX A
COMPLETION OF THE PROOF OF THEOREM 2

We first prove (49a).

- For $\beta = 1$, (49a) holds due to the upper bound on the harmonic sum in [21, Theorem 1];

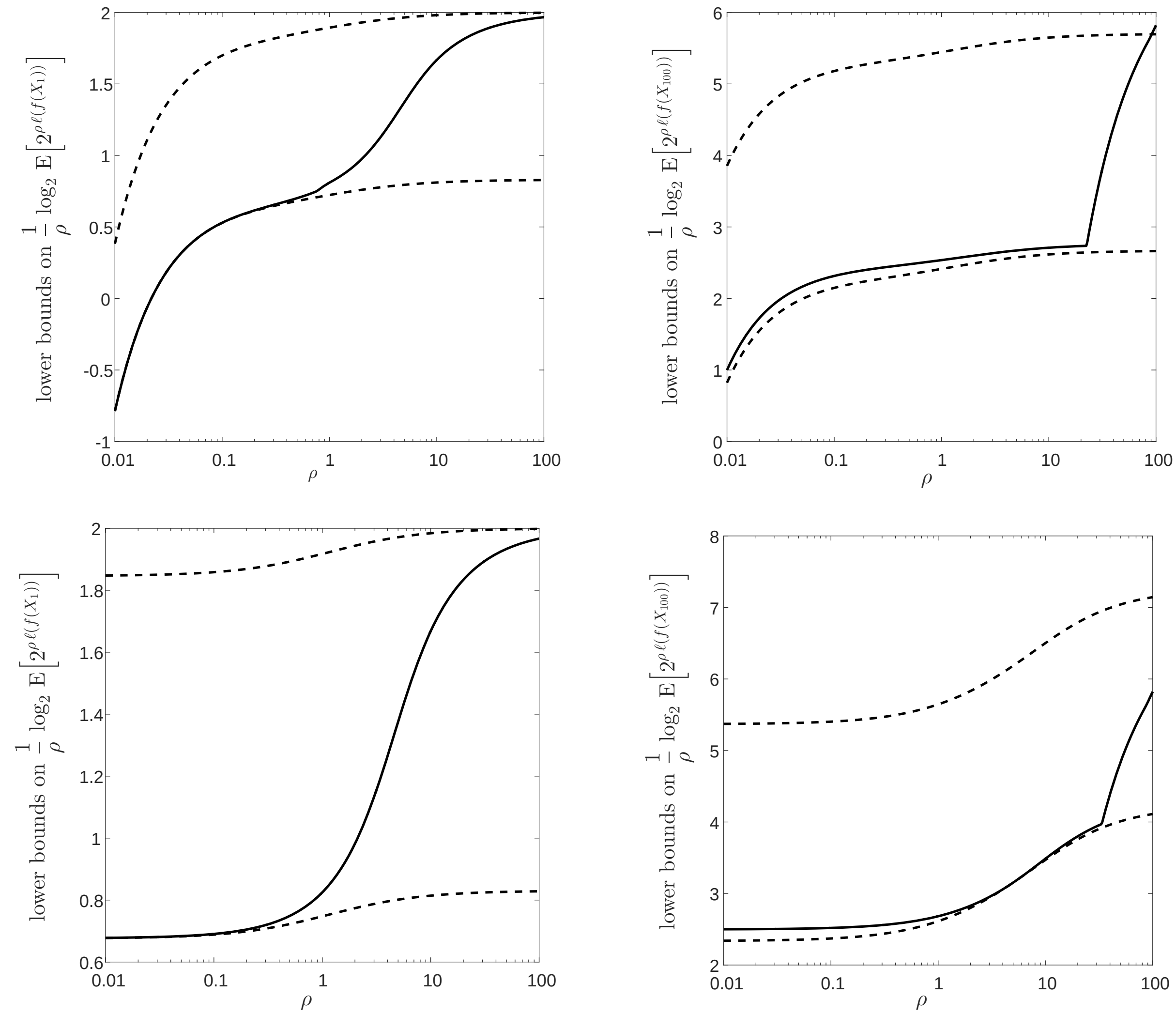

Fig. 5. Example 8: Lower bounds on $\frac{1}{\rho}\log_2 \mathbb{E}\left[2^{\rho\,\ell(f(X_n))}\right]$ with a maximal error probability of $\varepsilon = 0.01$ (upper plots), and the lossless case: $\varepsilon = 0$ (lower plots). The left and right plots correspond to $n = 1$ and $n = 100$, respectively. The upper and lower dashed lines in each plot refer, respectively, to (216) and (217); the solid lines refer to the combined lower bounds in (179) and (180) (Theorem 17).

- For $\beta > 1$

$$\sum_{j=1}^{M} \frac{1}{j^\beta} = \zeta(\beta) - \sum_{j=M+1}^{\infty} \frac{1}{j^\beta}$$

$$= \zeta(\beta) - \tfrac{1}{2}(M+1)^{-\beta} - \sum_{j=M+1}^{\infty} \tfrac{1}{2}\left(j^{-\beta} + (j+1)^{-\beta}\right) \tag{220}$$

where $\zeta(\beta) = \sum_{n=1}^{\infty} \frac{1}{n^\beta}$ for $\beta > 1$ denotes the Riemann zeta function; due to the convexity of $f_\beta(t) = t^{-\beta}$ in $(0, \infty)$, for all $j \in \mathbb{N}$,

$$\tfrac{1}{2}\left(j^{-\beta} + (j+1)^{-\beta}\right) \geq \int_j^{j+1} t^{-\beta}\,\mathrm{d}t. \tag{221}$$

Hence, from (220) and (221), for all $\beta > 1$,

$$\sum_{j=1}^{M} \frac{1}{j^\beta} \leq \zeta(\beta) - \tfrac{1}{2}(M+1)^{-\beta} - \int_{M+1}^{\infty} t^{-\beta}\,\mathrm{d}t \tag{222}$$

$$= \zeta(\beta) - \tfrac{1}{2}(M+1)^{-\beta} - \frac{(M+1)^{1-\beta}}{\beta-1} \tag{223}$$

and

$$\sum_{j=1}^{M} \frac{1}{j^\beta} \leq \sum_{j=1}^{M} \frac{1}{j} \leq u_M(1). \tag{224}$$

Combining (222)–(224) proves (49a) for $\beta > 1$.

- For $\beta > 0$, from the convexity of $f_\beta(t) = t^{-\beta}$ in $(0, \infty)$, Jensen's inequality yields

$$\int_{j-\frac{1}{2}}^{j+\frac{1}{2}} t^{-\beta}\,\mathrm{d}t \geq \frac{1}{j^\beta} \tag{225}$$

for all $j \in \mathbb{N}$, which implies that

$$\sum_{j=1}^{M} \frac{1}{j^\beta} \leq 1 + \int_{\frac{3}{2}}^{M+\frac{1}{2}} t^{-\beta}\,\mathrm{d}t \tag{226}$$

$$= 1 + \tfrac{1}{1-\beta}\left[\left(M+\tfrac{1}{2}\right)^{1-\beta} - \left(\tfrac{3}{2}\right)^{1-\beta}\right]. \tag{227}$$

This proves (49a) for $\beta \in (0,1)$. It can be verified numerically that the upper bound in (223) supersedes the bound (227) for $\beta > 1$; for this reason, we ignore the bound in (227) for the derivation of $u_M(\beta)$ for $\beta > 1$.

We next prove (49b).

- For $\beta \in [-1, 0)$, (49b) follows from the concavity of $f(t) = t^{-\beta}$ for $t > 0$; by Jensen's inequality, we obtain the opposite inequalities in (225) and (226) for $\beta \in [-1, 0)$.
- For $\beta = -1$, $\sum_{j=1}^{M} \frac{1}{j^\beta} = \frac{1}{2}\, M(M+1)$.
- For $\beta \in (-\infty, -1)$, due to the convexity of $f(t) = t^{-\beta}$ for $t > 0$,

$$\sum_{j=1}^{M} \frac{1}{j^\beta} = \tfrac{1}{2} + \sum_{j=1}^{M-1} \tfrac{1}{2}\big(j^{-\beta} + (j+1)^{-\beta}\big) + \tfrac{1}{2}\, M^{-\beta} \tag{228}$$

$$\geq \tfrac{1}{2} + \sum_{j=1}^{M-1} \int_{j}^{j+1} t^{-\beta}\,\mathrm{d}t + \tfrac{1}{2}\, M^{-\beta} \tag{229}$$

$$= \int_{1}^{M} t^{-\beta}\,\mathrm{d}t + \tfrac{1}{2}\big(1 + M^{-\beta}\big) \tag{230}$$

$$= \frac{M^{1-\beta} - 1}{1-\beta} + \tfrac{1}{2}\big(1 + M^{-\beta}\big). \tag{231}$$

Finally, Theorem 2 follows from Theorem 1, (49a) and (49b).

Appendix B
Proof of Theorem 5

We first prove (73) by relying on the following result:

Lemma 8: For $\rho \in (0,1)$ and $u \geq 1$

$$u^\rho \leq \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + \frac{\rho}{1+\rho}\, 1\{1 \leq u < 2\} + \left(2^\rho - \frac{2^{1+\rho}-1}{1+\rho}\right) 1\{u \geq 2\}. \tag{232}$$

Proof: For $\rho \in (0,1)$, let $f_{1,\rho} \colon [1,\infty) \to \mathbb{R}$ and $f_{2,\rho} \colon [1,\infty) \to \mathbb{R}$ be given by

$$f_{1,\rho}(u) = \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + \frac{\rho}{1+\rho} - u^\rho, \tag{233}$$

$$f_{2,\rho}(u) = \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + 2^\rho - \frac{2^{1+\rho}-1}{1+\rho} - u^\rho. \tag{234}$$

For $\rho \in (0,1)$ and $u \in [1,\infty)$

$$\begin{aligned} f'_{1,\rho}(u) = f'_{2,\rho}(u) &= u^\rho - (u-1)^\rho - \rho u^{\rho-1} && (235)\\ &= \rho c^{\rho-1} - \rho u^{\rho-1}, \quad c \in (u-1, u) && (236)\\ &> 0, && (237)\end{aligned}$$

where (236) holds by the mean value theorem of calculus; moreover, since $f_{1,\rho}(1) = f_{2,\rho}(2) = 0$,

$$\begin{aligned} f_{1,\rho}(u) \geq 0, &\quad u \in [1,\infty), && (238)\\ f_{2,\rho}(u) \geq 0, &\quad u \in [2,\infty). && (239)\end{aligned}$$

This gives

$$\begin{aligned} 0 &\leq f_{1,\rho}(u)\, 1\{1 \leq u < 2\} + \min\{f_{1,\rho}(u), f_{2,\rho}(u)\}\, 1\{u \geq 2\} && (240)\\ &= f_{1,\rho}(u)\, 1\{1 \leq u < 2\} + f_{2,\rho}(u)\, 1\{u \geq 2\} && (241)\end{aligned}$$

where (240) follows from (238) and (239), and (241) follows from (233), (234), and since

$$2^\rho - \frac{2^{1+\rho}-1}{1+\rho} - \frac{\rho}{1+\rho} = \frac{(\rho-1)(2^\rho-1)}{1+\rho} < 0 \tag{242}$$

for $\rho \in (0,1)$. Finally, (232) is equivalent to the non-negativity of the right side of (241). ■

From Lemma 8, for $\rho \in (0,1)$,

$$\begin{aligned} \mathbb{E}\left[g_X^\rho(X)\right] &\leq \frac{1}{1+\rho}\, \mathbb{E}\left[g_X^{1+\rho}(X) - \big(g_X(X)-1\big)^{1+\rho}\right] \\ &\quad + \frac{\rho\, \mathbb{P}[g_X(X)=1]}{1+\rho} + \left(2^\rho - \frac{2^{1+\rho}-1}{1+\rho}\right) \mathbb{P}[g_X(X) \geq 2] && (243)\\ &= \frac{1}{1+\rho}\, \mathbb{E}\left[g_X^{1+\rho}(X) - \big(g_X(X)-1\big)^{1+\rho}\right] + \frac{\rho\, p_{\max}}{1+\rho} + \left(2^\rho - \frac{2^{1+\rho}-1}{1+\rho}\right)(1-p_{\max}) && (244)\\ &\leq \frac{1}{1+\rho}\, \exp\Big(\rho H_{\frac{1}{1+\rho}}(X)\Big) + \frac{\rho\, p_{\max}}{1+\rho} + \left(2^\rho - \frac{2^{1+\rho}-1}{1+\rho}\right)(1-p_{\max}) && (245)\end{aligned}$$

where (243) follows from (232) by substituting $u = g_X(X)$, and taking expectations on both sides of the inequality; (244) holds since $\mathbb{P}[g_X(X)=1] = p_{\max}$ (the first guess of X is a mode of P_X); (245) follows from (60), which is then simplified to (73).

We next prove (74).

Lemma 9: If $\rho \in [1,2]$ and $u \geq 1$, then

$$u^\rho \leq \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + \frac{u^\rho - (u-1)^\rho}{\rho} + \frac{\rho^2 - \rho - 1}{\rho(1+\rho)}. \tag{246}$$

Proof: For $\rho \in [1,2]$, let $f \colon [1,\infty) \to \mathbb{R}$ be given by

$$f_\rho(u) = \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + \frac{u^\rho - (u-1)^\rho}{\rho} - u^\rho + \frac{\rho^2 - \rho - 1}{\rho(1+\rho)}, \quad u \in [1,\infty). \tag{247}$$

For all $u \in [1,\infty)$,

$$\begin{aligned} f'_\rho(u) &= u^\rho - (u-1)^\rho + u^{\rho-1} - (u-1)^{\rho-1} - \rho u^{\rho-1} & (248)\\ &\geq 1 + \rho(u-1)^{\rho-1} + u^{\rho-1} - (u-1)^{\rho-1} - \rho u^{\rho-1} & (249)\\ &= 1 + (\rho-1)\left((u-1)^{\rho-1} - u^{\rho-1}\right) & (250)\\ &\geq 2 - \rho \geq 0 & (251) \end{aligned}$$

where (249) follows from the convexity of $u \mapsto u^\rho$ in $(0,\infty)$ for $\rho \geq 1$, and (251) holds since

$$0 \leq u^{\rho-1} - (u-1)^{\rho-1} \leq 1 \tag{252}$$

for $\rho \in [1,2]$ and $u \in [1,\infty)$. It can be verified that $f_\rho(1) = 0$, which implies that $f_\rho(u) \geq 0$ for $\rho \in [1,2]$ and $u \geq 1$. ■

Replacing x in (246) with $g_X(X) \geq 1$, and taking expectations on both sides of (246) yield, via (60) and (65), the result in (74).

If X is deterministic, then $H_\alpha(X) = 0$ for all $\alpha > 0$, $g_X(X) = 1$, and $p_{\max} = 1$, which imply that (73) and (74) hold with equality (note that $\frac{1}{1+\rho} + \frac{1}{\rho} + \frac{\rho^2-\rho-1}{\rho(\rho+1)} = 1$ holds for $\rho \neq -1, 0$).

APPENDIX C
PROOF OF THEOREM 6

Lemma 10: Under the assumptions in Theorem 3, if $\rho \geq 2$, then

$$\mathbb{E}[g_X^\rho(X)] \leq \frac{1}{1+\rho} \exp\Bigl(\rho H_{\frac{1}{1+\rho}}(X)\Bigr) + \frac{\rho}{2}\, \mathbb{E}[g_X^{\rho-1}(X)] - \frac{\rho(\rho-1)}{2(1+\rho)} \tag{253}$$

with equality if X is deterministic.

Proof: If $\rho \geq 2$ and $u \geq 1$, then

$$u^\rho \leq \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} + \frac{\rho}{2}\, u^{\rho-1} - \frac{\rho(\rho-1)}{2(1+\rho)}. \tag{254}$$

To prove (254), let $\xi \colon [1,\infty) \to \mathbb{R}$ be given by

$$\xi(u) = \frac{u^{1+\rho} - (u-1)^{1+\rho}}{1+\rho} - u^\rho + \frac{\rho u^{\rho-1}}{2} - \frac{\rho(\rho-1)}{2(1+\rho)} \tag{255}$$

for $u \geq 1$, and similarly to (62), we denote $v(u) = u^\rho$ for $u \geq 0$. Then, for $u \geq 1$,

$$\begin{aligned} \xi'(u) &= u^\rho - (u-1)^\rho - \rho u^{\rho-1} + \tfrac{1}{2}\,\rho(\rho-1) u^{\rho-2} & (256)\\ &= v(u) - v(u-1) - v'(u) + \tfrac{1}{2} v''(u). & (257) \end{aligned}$$

By a Taylor series expansion of $v(\cdot)$ around u,

$$v(u-1) = v(u) - v'(u) + \tfrac{1}{2} v''(u) - \tfrac{1}{6} v^{(3)}(c) \tag{258}$$

for some $c \in (u-1, u) \subseteq (0, \infty)$. For $\rho \geq 2$, we have $v^{(3)}(c) = \rho(\rho-1)(\rho-2)c^{\rho-3} \geq 0$, which implies from (257) that $\xi'(u) \geq 0$ for all $u \geq 1$. It can be verified from (255) that $\xi(1) = 0$, which implies that $\xi(\cdot) \geq 0$ in $[1, \infty)$; this gives (254) from (255). By substituting $u = g_X(X)$ in (254) and taking expectations on both sides of the inequality, it follows that for $\rho \geq 2$

$$\mathbb{E}[g_X^\rho(X)] \leq \frac{\mathbb{E}\big[g_X^{1+\rho}(X)\big] - \mathbb{E}\big[\big(g_X(X)-1\big)^{1+\rho}\big]}{1+\rho} + \frac{\rho}{2}\,\mathbb{E}[g_X^{\rho-1}(X)] - \frac{\rho(\rho-1)}{2(1+\rho)} \tag{259}$$

which, from (60), yields (253). If X is deterministic, then $g_X(X) = 1$ and $H_\alpha(X) = 0$ for $\alpha > 0$, which implies equality in (253) since $\frac{1}{1+\rho} + \frac{\rho}{2} - \frac{\rho(\rho-1)}{2(1+\rho)} = 1$ holds for $\rho \neq -1$. ■

We proceed to prove Theorem 6. Combining the recursive upper bound in Lemma 10 with Theorem 5-b) gives, after some straightforward algebra, an upper bound on $\mathbb{E}[g_X^\rho(X)]$ for $\rho \geq 2$ which is of the form

$$\mathbb{E}[g_X^\rho(X)] \leq \sum_{j=0}^{\lfloor \rho \rfloor} c_j(\rho) \exp\Big((\rho - j)\, H_{\frac{1}{1+\rho-j}}(X)\Big) + d(\rho), \tag{260}$$

where the sequence $\{c_j(\rho)\}$ is given in (76), and $d(\rho)$ is an additive term which only depends on ρ (but it does not on the distribution of X). Since the results in Theorem 5-b) and Lemma 10 are satisfied with equalities if X is deterministic, then it follows that also (260) holds with equality in this special case. For such X, we have $g_X(X) = 1$ and $H_\alpha(X) = 0$ for all $\alpha > 0$, which therefore implies that

$$d(\rho) = 1 - \sum_{j=0}^{\lfloor \rho \rfloor} c_j(\rho). \tag{261}$$

The bound in (75) is obtained by combining (260) and (261).

Appendix D
Auxiliary Results for Section III-E

Lemmas 11–13 are used to derive the bounds on the optimal generalized guessing moment in (91) and (92), and Lemma 14 refers to the proof of Theorem 11.

Lemma 11: Let $\ell \in \mathbb{N}$, $p_1 \geq p_2 \geq \ldots \geq p_\ell \geq 0$ with $\sum\limits_{i=1}^{\ell} p_i = 1$, and let $f \colon \{1, \ldots, \ell\} \to \mathbb{R}$ satisfy

$$\frac{1}{j} \sum_{k=l-j+1}^{\ell} f(k) \geq f(\ell - j), \tag{262}$$

for all $j \in \{1, \ldots, \ell\}$. Then,

$$\sum_{i=1}^{\ell} p_i f(i) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} f(i). \tag{263}$$

Furthermore, if the inequality in (262) is strict for all j, then (263) holds with equality if and only if $p_i = \frac{1}{\ell}$ for all $i \in \{1, \ldots, \ell\}$.

Proof: Denote

$$u_j = \sum_{i=1}^{\ell-j} p_i f(i) + \frac{1}{j} \left(\sum_{i=\ell-j+1}^{\ell} p_i \right) \left(\sum_{i=\ell-j+1}^{\ell} f(i) \right), \quad j \in \{1, \ldots, \ell\}. \tag{264}$$

Its first and last terms are equal to the side of (263):

$$u_1 = \sum_{i=1}^{\ell} p_i f(i), \tag{265}$$

$$u_\ell = \frac{1}{\ell} \sum_{i=1}^{\ell} f(i) \tag{266}$$

so, proving that $\{u_j\}$ is monotonically increasing is sufficient to show (263). By its definition in (264), straightforward calculation shows that for $j \in \{1, \ldots, \ell - 1\}$

$$u_{j+1} - u_j = \frac{1}{j(j+1)} \left(j\, p_{\ell-j} - \sum_{i=\ell-j+1}^{\ell} p_i \right) \left(\sum_{i=\ell-j+1}^{\ell} f(i) - j f(\ell - j) \right) \tag{267}$$

$$\geq 0 \tag{268}$$

where (268) holds in view of $p_1 \geq \ldots \geq p_\ell$ and (262). Note that (267) and (268) imply that $u_1 = u_l$ if and only if $u_{j+1} = u_j$ for all $j \in \{1, \ldots, l-1\}$. Hence, if the inequality in (262) is strict, then (267) implies that $u_{j+1} = u_j$ for all j if and only if the monotonic sequence $\{p_i\}$ is fixed, i.e., $p_1 = p_2 = \ldots = p_\ell = \frac{1}{\ell}$ (since, by assumption, this sequence sums to 1). ∎

Lemma 12: Let $l \in \mathbb{N}$, $q_1 \geq \ldots \geq q_\ell \geq 0$, and let $f \colon \{1, \ldots, \ell\} \to \mathbb{R}$ be a strictly monotonically increasing function. Then,

$$\sum_{i=1}^{\ell} q_i\, f(1) \leq \sum_{i=1}^{\ell} q_i f(i) \leq \frac{1}{\ell} \left(\sum_{i=1}^{\ell} q_i \right) \left(\sum_{i=1}^{\ell} f(i) \right) \tag{269}$$

with equality in the left inequality if and only if $q_i = 0$ for all $i \in \{2, \ldots, \ell\}$, and equality in the right inequality if and only if $q_1 = \ldots q_\ell$.

Proof: Since by assumption $f \colon \{1, \ldots, \ell\} \to \mathbb{R}$ is a strictly monotonically increasing function, the inequality in (262) is strict for all $j \in \{1, \ldots, \ell - 1\}$. Let $\{p_i\}$ be the normalized version of the non-negative, monotonically decreasing sequence $\{q_i\}$ such that $\sum_{i=1}^{\ell} p_i = 1$. Hence,

$$p_i = \frac{q_i}{\sum_{j=1}^{\ell} q_j}, \quad i \in \{1, \ldots, \ell\}, \tag{270}$$

and $p_1 \geq p_2 \geq \ldots \geq p_\ell \geq 0$. Lemma 11 and (270) give

$$\sum_{i=1}^{\ell} q_i f(i) = \left(\sum_{i=1}^{\ell} p_i f(i) \right) \left(\sum_{i=1}^{\ell} q_i \right) \tag{271}$$

$$\leq \frac{1}{\ell} \left(\sum_{i=1}^{\ell} f(i) \right) \left(\sum_{i=1}^{\ell} q_i \right), \tag{272}$$

and, due to Lemma 11, (272) holds with equality if and only if $p_1 = \ldots = p_\ell = \frac{1}{\ell}$; due to (270), this holds if and only if $q_1 = \ldots = q_\ell$. Moreover, the assumptions on $f \colon \{1, \ldots, \ell\} \to \mathbb{R}$ and $\{q_i\}$ imply that

$$\sum_{i=1}^{\ell} q_i f(i) \geq f(1) \sum_{i=1}^{\ell} q_i, \tag{273}$$

where (273) holds with equality if and only if $q_i = 0$ for all $i \in \{2, \ldots, \ell\}$. ∎

Lemma 13: The function $f_\rho\colon [0,1) \to [0,\infty)$ defined in (93)–(94) is convex for all $\rho > 0$.

Proof: From (94), the value of $k_u \in \mathbb{N}$ is fixed in each interval $[1-\frac{1}{m}, 1-\frac{1}{m+1})$ with $m \in \mathbb{N}$. Hence, f_ρ in (93) is a linear function in each such interval; its positive slope, denoted by $s_\rho(m)$, is given by

$$s_\rho(m) = k_u(k_u+1)^\rho - \sum_{j=1}^{k_u} j^\rho. \tag{274}$$

By a transition from an interval $[1-\frac{1}{m}, 1-\frac{1}{m+1})$ to the successive interval $[1-\frac{1}{m+1}, 1-\frac{1}{m+2})$, the value of the positive integer k_u is increased by 1 (see (94)); consequently, it can be verified from (274) that for $\rho > 0$

$$s_\rho(m+1) > s_\rho(m), \quad m \in \mathbb{N}. \tag{275}$$

Hence, the slope of the linear function obtained by restricting f_ρ to the interval $[1-\frac{1}{m+1}, 1-\frac{1}{m+2})$ is larger than its slope in the interval $[1-\frac{1}{m}, 1-\frac{1}{m+1})$. Hence, f_ρ can be decomposed by linear functions in each interval $[1-\frac{1}{m}, 1-\frac{1}{m+1})$ whose slopes are monotonically increasing in $m \in \mathbb{N}$. It can be also verified from (93) that the function f_ρ is continuous at the endpoints of these intervals, which therefore yields its convexity on $[0,1) = \bigcup\limits_{m\in\mathbb{N}} [1-\frac{1}{m}, 1-\frac{1}{m+1})$. ■

Lemma 14: The identity in (127) holds for every integer $M \geq 2$.

Proof: The result in (127) is trivial for $M = 2$. Let $M \geq 3$, then

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ 0 & \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ 0 & \log_{\mathrm{e}}^2 2 & \cdots & \log_{\mathrm{e}}^2 M \\ 0 & \vdots & \vdots & \vdots \\ 0 & \log_{\mathrm{e}}^{M-1} 2 & \cdots & \log_{\mathrm{e}}^{M-1} M \end{vmatrix} = \begin{vmatrix} \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ \log_{\mathrm{e}}^2 2 & \cdots & \log_{\mathrm{e}}^2 M \\ \vdots & \vdots & \vdots \\ \log_{\mathrm{e}}^{M-1} 2 & \cdots & \log_{\mathrm{e}}^{M-1} M \end{vmatrix} \tag{276}$$

$$= \left(\prod_{k=2}^{M} \log_{\mathrm{e}} k\right) \begin{vmatrix} 1 & \cdots & 1 \\ \log_{\mathrm{e}} 2 & \cdots & \log_{\mathrm{e}} M \\ \vdots & \vdots & \vdots \\ \log_{\mathrm{e}}^{M-2} 2 & \cdots & \log_{\mathrm{e}}^{M-2} M \end{vmatrix} \tag{277}$$

$$= \prod_{k=2}^{M} \log_{\mathrm{e}} k \prod_{2 \leq i < j \leq M} (\log_{\mathrm{e}} j - \log_{\mathrm{e}} i) \tag{278}$$

where (276) holds by expanding according to the first column; (277) holds by factoring $\log_{\mathrm{e}} k$ from the $(k-1)$-th row for $k = 2, \ldots, M$; finally, (278) relies on the Vandermonde determinant (e.g., [29, p. 155]). ■

References

[1] 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.

[2] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. on Information Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[3] E. Arikan and N. Merhav, "Joint source-channel coding and guessing with application to sequential decoding," *IEEE Trans. on Information Theory*, vol. 44, no. 5, pp. 1756–1769, September 1998.

[4] 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 János Bolyai (I. Csiszár and P. Elias editors), Amsterdam, the Netherlands: North Holland, vol. 16, pp. 41-52, 1977.

[5] R. Atar, K. Chowdhary and P. Dupuis, "Robust bounds on risk-sensitive functionals via Rényi divergence," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 3, no. 1, pp. 18–33, January 2015.

[6] S. Boztaş, "Comments on "An inequality on guessing and its application to sequential decoding"," *IEEE Trans. on Information Theory*, vol. 43, no. 6, pp. 2062–2063, November 1997.

[7] 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.

[8] A. Bracher, A. Lapidoth and C. Pfister, "Distributed task encoding," *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pp. 1993–1997, Aachen, Germany, June 2017.

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

[10] C. Bunte and A. Lapidoth, "Maximizing Rényi entropy rate," *Proceedings of the 2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel*, pp. 1–4, December 2014, Eilat, Israel.

[11] A. Burin and O. Shayevitz, "Reducing guesswork via an unreliable oracle," March 2017. [Online]. Available at https://arxiv.org/abs/1703.01672.

[12] C. Cachin, "Smooth entropy and Rényi entropy," *Lecture Notes in Computer Science*, vol. 1233, pp. 193–208, Springer, 1997.

[13] L. L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, no. 4, pp. 423–429, August 1965.

[14] L. L. Campbell, "Definition of entropy by means of a coding problem," *Probability Theory and Related Fields*, vol. 6, no. 2, pp. 113–118, June 1966.

[15] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and Shannon entropy," *IEEE Trans. on Information Theory*, vol. 59, no. 2, pp. 796–802, February 2013.

[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. Courtade and S. Verdú, "Variable-length lossy compression and channel coding: Non-asymptotic converses via cumulant generating functions," *Proceedings of the 2014 IEEE International Symposium on Information Theory*, pp. 2499–2503, Honolulu, Hawaii, USA, July 2014.

[18] M. Drmota and W. Szpankowski, "Redundancy of lossless data compression for known sources by analytic methods," *Foundations and Trends in Communications and Information Theory*, vol. 13, no. 4, pp. 277–417, 2016.

[19] 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.

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

[21] B. N. Guo and F. Qi, "Sharp bounds for harmonic numbers," *Applied Mathematics and Computation*, vol. 218, no. 3, pp. 991–995, October 2011.

[22] M. K. Hanawal and R. Sundaresan, "Guessing revisited: a large deviations approach," *IEEE Trans. on Information Theory*, vol. 57, no. 1, pp. 70–78, January 2011.

[23] W. Huleihel, S. Salamatian, and M. Médard, "Guessing with limited memory," *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pp. 2258–2262, Aachen, Germany, June 2017.

[24] H. Koga and H. Yamamoto, "Asymptotic properties on codeword lengths of an optimal FV code for general sources," *IEEE Trans. on Information Theory*, vol. 51, no. 4, pp. 1546–1555, April 2005.

[25] H. Koga, "Characterization of the smooth Rényi entropy using majorization," *Proceedings of the 2013 IEEE Information Theory Workshop*, pp. 604–608, Seville, Spain, September 2013.

[26] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: non-asymptotics and asymptotics," *IEEE Trans. on Information Theory*, vol. 60, no. 2, pp. 777–795, February 2014.

[27] V. Kostina, Y. Polyanskiy and S. Verdú, "Variable-length compression allowing errors," *IEEE Trans. on Information Theory*, vol. 61, no. 8, pp. 4316–4330, August 2015.

[28] S. Kuzuoka, "On the smooth Rényi entropy and variable-length source coding allowing errors," *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pp. 745–749, Barcelona, Spain, July 2016.

[29] S. Lang, *Linear Algebra*, Springer, Third edition, 1987.

[30] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. on Information Theory*, vol. 50, no. 3, pp. 525–526, March 2004.

[31] J. L. Massey, "Guessing and entropy," *Proceedings of the 1994 IEEE International Symposium on Information Theory*, p. 204, Trondheim, Norway, June 1994.

[32] R. J. McEliece and Z. Yu, "An inequality on entropy," *Proceedings of the 1995 IEEE International Symposium on Information Theory*, p. 329, Whistler, Canada, September 1995.

[33] N. Merhav and E. Arikan, "The Shannon cipher system with a guessing wiretapper," *IEEE Trans. on Information Theory*, vol. 45, no. 6, pp. 1860–1866, September 1999.

[34] N. Merhav, “On optimum strategies for minimizing exponential moments for a given cost function,” *Communications in Information Systems*, vol. 11, no. 4, pp. 343–368, 2011.

[35] N. Merhav, “On empirical cumulant generating functions of code lengths for individual sequences,” *IEEE Trans. on Information Theory*, vol. 63, no. 12, pp. 7729–7736, December 2017.

[36] R. Renner and S. Wolf, “Smooth Rényi entropy and applications,” *Proceedings of the 2004 IEEE International Symposium on Information Theory*, p. 232, Chicago, Illinois, USA, July 2004.

[37] R. Renner and S. Wolf, “Simple and tight bounds for information reconciliation and privacy amplification,” *Advances in Cryptology - ASIACRYPT 2005*, Lecture Notes in Computer Science, vol. 3788, pp. 199–216, Springer, 2005.

[38] A. Rényi, “On measures of entropy and information,” *Proceedings of the 4th Berkeley Symposium on Probability Theory and Mathematical Statistics*, pp. 547–561, Berkeley, California, USA, 1961.

[39] Y. Sakai and K. Iwata, “Sharp bounds on Arimoto’s conditional Rényi entropies between two distinct orders,” *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pp. 2985–2989, Aachen, Germany, June 2017.

[40] S. Salamatian, A. Beirami, A. Cohen and M. Médard, “Centralized versus decentralized multi-agent guesswork,” *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pp. 2263–2267, Aachen, Germany, June 2017.

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

[42] I. Sason and S. Verdú, “Arimoto-Rényi conditional entropy and Bayesian M-Ary hypothesis testing,” *IEEE Trans. on Information Theory*, vol. 64, no. 1, pp. 4–25, January 2018.

[43] 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.

[44] 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.

[45] V. Strassen, “Asymptotische abschäzungen in Shannon’s informations-theorie,” *Transactions of the Third Prague Conference on Information Theory, Statistics, Decision Functions, Random Processes (Liblice, 1962).* Prague: Publ. House Czech. Acad. Sci., pp. 689–723, 1964.

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

[47] R. Sundaresan, “Guessing based on length functions,” *Proceedings of the 2007 IEEE International Symposium on Information Theory*, pp. 716–719, Nice, France, June 2007.

[48] W. Szpankowski and S. Verdú, “Minimum expected length of fixed-to-variable lossless compression without prefix constraints,” *IEEE Trans. on Information Theory*, vol. 57, no. 7, pp. 4017–4025, July 2011.

[49] S. Verdú, “Shannon’s inequality,” *2011 Workshop on Information Theory and Applications*, San Diego, California, USA, February 2011. [Online]. Available: http://ita.ucsd.edu/workshop/11/files/paper/paper_374.pdf.

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

[51] Y. Yona and S. Diggavi, “The effect of bias on the guesswork of hash functions,” *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pp. 2253–2257, Aachen, Germany, June 2017.