Tight Bounds on the Rényi Entropy via Majorization with Applications to Guessing and Compression

Igal Sason

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel; sason@ee.technion.ac.il; Tel.: +972-4-8294699

check for updates

Abstract: This paper provides tight bounds on the Rényi entropy of a function of a discrete random variable with a finite number of possible values, where the considered function is not one to one. To that end, a tight lower bound on the Rényi entropy of a discrete random variable with a finite support is derived as a function of the size of the support, and the ratio of the maximal to minimal probability masses. This work was inspired by the recently published paper by Cicalese et al., which is focused on the Shannon entropy, and it strengthens and generalizes the results of that paper to Rényi entropies of arbitrary positive orders. In view of these generalized bounds and the works by Arikan and Campbell, non-asymptotic bounds are derived for guessing moments and lossless data compression of discrete memoryless sources.

Keywords: Majorization; Rényi entropy; Rényi divergence; cumulant generating functions; guessing moments; lossless source coding; fixed-to-variable source codes; Huffman algorithm; Tunstall codes

1. Introduction

Majorization theory is a simple and productive concept in the theory of inequalities, which also unifies a variety of familiar bounds [1,2]. These mathematical tools find various applications in diverse fields (see, e.g., [3]) such as economics [2,4,5], combinatorial analysis [2,6], geometric inequalities [2], matrix theory [2,6–8], Shannon theory [5,9–25], and wireless communications [26–33].

This work, which relies on the majorization theory, has been greatly inspired by the recent insightful paper by Cicalese et al. [12] (the research work in the present paper has been initialized while the author handled [12] as an associate editor). The work in [12] provides tight bounds on the Shannon entropy of a function of a discrete random variable with a finite number of possible values, where the considered function is not one to one. For that purpose, and while being of interest by its own right (see [12], Section 6), a tight lower bound on the Shannon entropy of a discrete random variable with a finite support was derived in [12] as a function of the size of the support, and the ratio of the maximal to minimal probability masses. The present paper aims to extend the bounds in [12] to Rényi entropies of arbitrary positive orders (note that the Shannon entropy is equal to the Rényi entropy of order 1), and to study the information-theoretic applications of these (non-trivial) generalizations in the context of non-asymptotic analysis of guessing moments and lossless data compression.

The motivation for this work is rooted in the diverse information-theoretic applications of Rényi measures [34]. These include (but are not limited to) asymptotically tight bounds on guessing moments [35], information-theoretic applications such as guessing subject to distortion [36], joint source-channel coding and guessing with application to sequential decoding [37], guessing with a prior access to a malicious oracle [38], guessing while allowing the guesser to give up and declare an error [39], guessing in secrecy problems [40,41], guessing with limited memory [42], and guessing under source uncertainty [43]; encoding tasks [44,45]; Bayesian hypothesis testing [9,22,23], and composite

hypothesis testing [46,47]; Rényi generalizations of the rejection sampling problem in [48], motivated by the communication complexity in distributed channel simulation, where these generalizations distinguish between causal and noncausal sampler scenarios [49]; Wyner's common information in distributed source simulation under Rényi divergence measures [50]; various other source coding theorems [23,39,51–58], channel coding theorems [23,58–64], including coding theorems in quantum information theory [65–67].

The presentation in this paper is structured as follows: Section 2 provides notation and essential preliminaries for the analysis in this paper. Sections 3 and 4 strengthen and generalize, in a non-trivial way, the bounds on the Shannon entropy in [12] to Rényi entropies of arbitrary positive orders (see Theorems 1 and 2). Section 5 relies on the generalized bound from Section 4 and the work by Arikan [35] to derive non-asymptotic bounds for guessing moments (see Theorem 3); Section 5 also relies on the generalized bound in Section 4 and the source coding theorem by Campbell [51] (see Theorem 4) for the derivation of non-asymptotic bounds for lossless compression of discrete memoryless sources (see Theorem 5).

2. Notation and Preliminaries

Let

- P be a probability mass function defined on a finite set $\mathcal{X}$;
- $p_{\max}$ and $p_{\min}$ be, respectively, the maximal and minimal positive masses of P;
- $G_P(k)$ be the sum of the k largest masses of P for $k \in \{1, \ldots, |\mathcal{X}|\}$ (note that $G_P(1) = p_{\max}$ and $G_P(|\mathcal{X}|) = 1$);
- $\mathcal{P}_n$, for an integer $n \geq 2$, be the set of all probability mass functions defined on $\mathcal{X}$ with $|\mathcal{X}| = n$; without any loss of generality, let $\mathcal{X} = \{1, \ldots, n\}$;
- $\mathcal{P}_n(\rho)$, for $\rho \geq 1$ and an integer $n \geq 2$, be the subset of all probability measures $P \in \mathcal{P}_n$ such that

$$\frac{p_{\max}}{p_{\min}} \leq \rho. \tag{1}$$

Definition 1 (Majorization). *Consider discrete probability mass functions P and Q defined on the same (finite or countably infinite) set $\mathcal{X}$. It is said that P is majorized by Q (or Q majorizes P), and it is denoted by $P \prec Q$, if $G_P(k) \leq G_Q(k)$ for all $k \in \{1, \ldots, |\mathcal{X}| - 1\}$ (recall that $G_P(|\mathcal{X}|) = G_Q(|\mathcal{X}|) = 1$). If P and Q are defined on finite sets of different cardinalities, then the probability mass function which is defined over the smaller set is first padded by zeros for making the cardinalities of these sets be equal.*

By Definition 1, a unit mass majorizes any other distribution; on the other hand, the equiprobable distribution on a finite set is majorized by any other distribution defined on the same set.

Definition 2 (Schur-convexity/concavity). *A function $f \colon \mathcal{P}_n \to \mathbb{R}$ is said to be Schur-convex if for every $P, Q \in \mathcal{P}_n$ such that $P \prec Q$, we have $f(P) \leq f(Q)$. Likewise, f is said to be Schur-concave if $-f$ is Schur-convex, i.e., $P, Q \in \mathcal{P}_n$ and $P \prec Q$ imply that $f(P) \geq f(Q)$.*

Definition 3 (Rényi entropy [34]). *Let X be a 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*

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

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

By its continuous extension,

$$H_0(X) = \log \big| \{ x \in \mathcal{X} : P_X(x) > 0 \} \big|, \tag{3}$$

$$H_1(X) = H(X), \tag{4}$$

$$H_\infty(X) = \log \frac{1}{p_{\max}} \tag{5}$$

where $H(X)$ is the (Shannon) entropy of X.

Proposition 1 (Schur-concavity of the Rényi entropy (Appendix F.3.a (p. 562) of [2]))**.** *The Rényi entropy of an arbitrary order $\alpha > 0$ is Schur-concave; in particular, for $\alpha = 1$, the Shannon entropy is Schur-concave.*

Remark 1. *[17] (Theorem 2) strengthens Proposition 1, though it is not needed for our analysis.*

Definition 4 (Rényi divergence [34])**.** *Let P and Q be probability mass functions defined on a finite or countably infinite set $\mathcal{X}$. The Rényi divergence of order $\alpha \in [0, \infty]$ is defined as follows:*

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

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \, \log \sum_{x \in \mathcal{X}} P^\alpha(x) \, Q^{1-\alpha}(x). \tag{6}$$

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

$$D_0(P\|Q) = \max_{\mathcal{A}: P(\mathcal{A})=1} \log \frac{1}{Q(\mathcal{A})}, \tag{7}$$

$$D_1(P\|Q) = D(P\|Q), \tag{8}$$

$$D_\infty(P\|Q) = \log \sup_{x \in \mathcal{X}} \frac{P(x)}{Q(x)}, \tag{9}$$

where $D(P\|Q)$ in the right side of (8) is the relative entropy (a.k.a. Kullback-Leibler divergence).

Throughout this paper, for $a \in \mathbb{R}$, $\lceil a \rceil$ denotes the ceiling of a (i.e., the smallest integer not smaller than the real number a), and $\lfloor a \rfloor$ denotes the flooring of a (i.e., the greatest integer not greater than a).

3. A Tight Lower Bound on the Rényi Entropy

We provide in this section a tight lower bound on the Rényi entropy, of an arbitrary order $\alpha > 0$, when the probability mass function of the discrete random variable is defined on a finite set of cardinality n, and the ratio of the maximal to minimal probability masses is upper bounded by an arbitrary fixed value $\rho \in [1, \infty)$. In other words, we derive the largest possible gap between the order-α Rényi entropies of an equiprobable distribution and a non-equiprobable distribution (defined on a finite set of the same cardinality) with a given value for the ratio of the maximal to minimal probability masses. The basic tool used for the development of our result in this section relies on the majorization theory. Our result strengthens the result in [12] (Theorem 2) for the Shannon entropy, and it further provides a generalization for the Rényi entropy of an arbitrary order $\alpha > 0$ (recall that the Shannon entropy is equal to the Rényi entropy of order $\alpha = 1$, see (4)). Furthermore, the approach for proving the main result in this section differs significantly from the proof in [12] for the Shannon entropy. The main result in this section is a key result for all what follows in this paper.

The following lemma is a restatement of [12] (Lemma 6).

Lemma 1. *Let* $P \in \mathcal{P}_n(\rho)$ *with* $\rho \geq 1$ *and an integer* $n \geq 2$, *and assume without any loss of generality that the probability mass function* P *is defined on the set* $\mathcal{X} = \{1, \ldots, n\}$. *Let* $Q \in \mathcal{P}_n$ *be defined on* $\mathcal{X}$ *as follows:*

$$
Q(j) = \begin{cases} \rho\, p_{\min}, & j \in \{1, \ldots, i\}, \\ 1 - (n + i\rho - i - 1)p_{\min}, & j = i + 1, \\ p_{\min}, & j \in \{i + 2, \ldots, n\} \end{cases} \tag{10}
$$

where

$$
i := \left\lfloor \frac{1 - np_{\min}}{(\rho - 1)\, p_{\min}} \right\rfloor. \tag{11}
$$

Then,

(1) $Q \in \mathcal{P}_n(\rho)$, *and* $Q(1) \geq Q(2) \geq \ldots \geq Q(n) > 0$;
(2) $P \prec Q$.

Proof. See [12] (p. 2236) (top of the second column).   □

Lemma 2. *Let* $\rho > 1$, $\alpha > 0$, *and* $n \geq 2$ *be an integer. For*

$$
\beta \in \left[\frac{1}{1 + (n-1)\rho}, \frac{1}{n} \right] := \Gamma_\rho^{(n)}, \tag{12}
$$

let $Q_\beta \in \mathcal{P}_n(\rho)$ *be defined on* $\mathcal{X} = \{1, \ldots, n\}$ *as follows:*

$$
Q_\beta(j) = \begin{cases} \rho\beta, & j \in \{1, \ldots, i_\beta\}, \\ 1 - (n + i_\beta\, \rho - i_\beta - 1)\beta, & j = i_\beta + 1, \\ \beta, & j \in \{i_\beta + 2, \ldots, n\} \end{cases} \tag{13}
$$

where

$$
i_\beta := \left\lfloor \frac{1 - n\beta}{(\rho - 1)\beta} \right\rfloor. \tag{14}
$$

Then, for every $\alpha > 0$,

$$
\min_{P \in \mathcal{P}_n(\rho)} H_\alpha(P) = \min_{\beta \in \Gamma_\rho^{(n)}} H_\alpha(Q_\beta). \tag{15}
$$

Proof. See Appendix A.   □

Lemma 3. *For* $\rho > 1$ *and* $\alpha > 0$, *let*

$$
c_\alpha^{(n)}(\rho) := \log n - \min_{P \in \mathcal{P}_n(\rho)} H_\alpha(P), \quad n = 2, 3, \ldots \tag{16}
$$

with $c_\alpha^{(1)}(\rho) := 0$. *Then, for every* $n \in \mathbb{N}$,

$$
0 \leq c_\alpha^{(n)}(\rho) \leq \log \rho, \tag{17}
$$

$$
c_\alpha^{(n)}(\rho) \leq c_\alpha^{(2n)}(\rho), \tag{18}
$$

and $c_\alpha^{(n)}(\rho)$ *is monotonically increasing in* $\alpha \in [0, \infty]$.

Proof. See Appendix B. □

Lemma 4. *For $\alpha > 0$ and $\rho > 1$, the limit*

$$c_\alpha^{(\infty)}(\rho) := \lim_{n\to\infty} c_\alpha^{(n)}(\rho) \tag{19}$$

exists, having the following properties:

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

$$c_\alpha^{(\infty)}(\rho) = \frac{1}{\alpha-1} \log\left(1 + \frac{1 + \alpha(\rho-1) - \rho^\alpha}{(1-\alpha)(\rho-1)}\right) - \frac{\alpha}{\alpha-1} \log\left(1 + \frac{1 + \alpha(\rho-1) - \rho^\alpha}{(1-\alpha)(\rho^\alpha-1)}\right), \tag{20}$$

　　and

$$\lim_{\alpha\to\infty} c_\alpha^{(\infty)}(\rho) = \log\rho. \tag{21}$$

(b)　*If $\alpha = 1$, then*

$$c_1^{(\infty)}(\rho) = \lim_{\alpha\to1} c_\alpha^{(\infty)}(\rho) = \frac{\rho\log\rho}{\rho-1} - \log\left(\frac{e\rho\log_e\rho}{\rho-1}\right). \tag{22}$$

(c)　*For all $\alpha > 0$,*

$$\lim_{\rho\downarrow1} c_\alpha^{(\infty)}(\rho) = 0. \tag{23}$$

- *For every $n \in \mathbb{N}$, $\alpha > 0$ and $\rho \geq 1$,*

$$0 \leq c_\alpha^{(n)}(\rho) \leq c_\alpha^{(2n)}(\rho) \leq c_\alpha^{(\infty)}(\rho) \leq \log\rho. \tag{24}$$

Proof. See Appendix C. □

　　In view of Lemmata 1–4, we obtain the following main result in this section:

Theorem 1. *Let $\alpha > 0$, $\rho > 1$, $n \geq 2$, and let $c_\alpha^{(n)}(\rho)$ in (16) designate the maximal gap between the order-α Rényi entropies of equiprobable and arbitrary distributions in $\mathcal{P}_n(\rho)$. Then,*

(a)　*The non-negative sequence $\{c_\alpha^{(n)}(\rho)\}_{n=2}^\infty$ can be calculated by the real-valued single-parameter optimization in the right side of (15).*

(b)　*The asymptotic limit as $n \to \infty$, denoted by $c_\alpha^{(\infty)}(\rho)$, admits the closed-form expressions in (20) and (22), and it satisfies the properties in (21), (23) and (24).*

Remark 2. *Setting $\alpha = 2$ in Theorem 1 gives that, for all $P \in \mathcal{P}_n(\rho)$ (with $\rho > 1$, and an integer $n \geq 2$),*

$$H_2(P) \geq \log n - c_2^{(n)}(\rho) \tag{25}$$

$$\geq \log n - c_2^{(\infty)}(\rho) \tag{26}$$

$$= \log \frac{4\rho n}{(1+\rho)^2} \tag{27}$$

where (25)–(27) hold, respectively, due to (16), (24) and (20). This strengthens the result in [68] (Proposition 2) which gives the same lower bound as in the right side of (27) for $H(P)$ rather than for $H_2(P)$ (recall that $H(P) \geq H_2(P)$).

For a numerical illustration of Theorem 1, Figure 1 provides a plot of $c_\alpha^{(\infty)}(\rho)$ in (20) and (22) as a function of $\rho \geq 1$, confirming numerically the properties in (21) and (23). Furthermore, Figure 2 provides plots of $c_\alpha^{(n)}(\rho)$ in (16) as a function of $\alpha > 0$, for $\rho = 2$ (left plot) and $\rho = 256$ (right plot), with several values of $n \geq 2$; the calculation of the curves in these plots relies on (15), (20) and (22), and they illustrate the monotonicity and boundedness properties in (24).
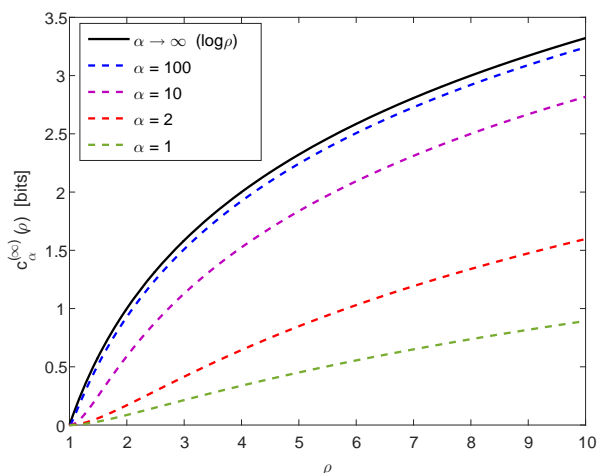

Figure 1. A plot of $c_\alpha^{(\infty)}(\rho)$ in (20) and (22) (log is on base 2) as a function of ρ, confirming numerically the properties in (21) and (23).
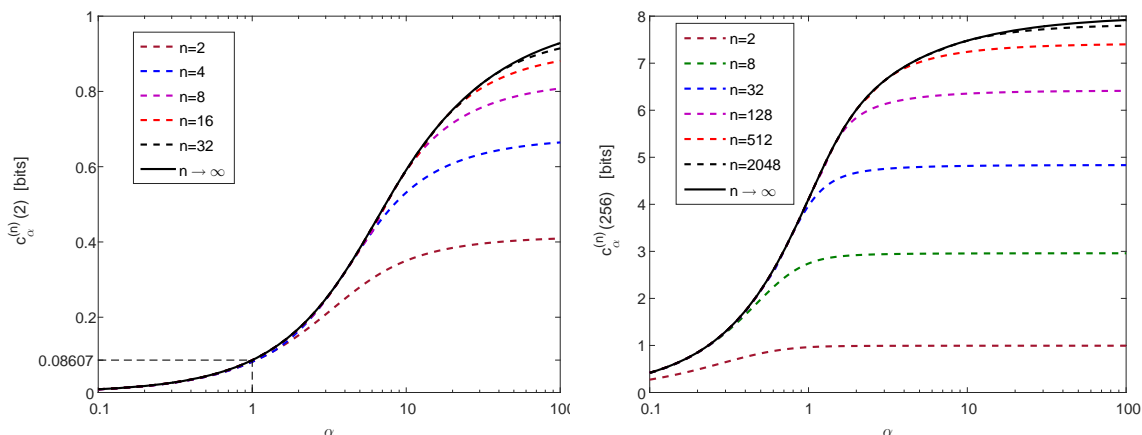

Figure 2. Plots of $c_\alpha^{(n)}(\rho)$ in (16) (log is on base 2) as a function of $\alpha > 0$, for $\rho = 2$ (left plot) and $\rho = 256$ (right plot), with several values of $n \geq 2$.

Remark 3. *Theorem 1 strengthens the result in [12] (Theorem 2) for the Shannon entropy (i.e., for $\alpha = 1$), in addition to its generalization to Rényi entropies of arbitrary orders $\alpha > 0$. This is because our lower bound on the Shannon entropy is given by*

$$H(P) \geq \log n - c_1^{(n)}(\rho), \quad \forall P \in \mathcal{P}_n(\rho), \tag{28}$$

whereas the looser bound in [12] is given by (see [12] ((7)) and (22) here)

$$H(P) \geq \log n - c_1^{(\infty)}(\rho), \quad \forall P \in \mathcal{P}_n(\rho), \tag{29}$$

and we recall that $0 \leq c_1^{(n)}(\rho) \leq c_1^{(\infty)}(\rho)$ (see (24)). Figure 3 shows the improvement in the new lower bound (28) over (29) by comparing $c_1^{(\infty)}(\rho)$ versus $c_1^{(n)}(\rho)$ for $\rho \in [1, 10^5]$ and with several values of n. It is reflected from Figure 3 that there is a very marginal improvement in the lower bound on the Shannon entropy (28)

over the bound in (29) if $\rho \leq 30$ (even for small values of n), whereas there is a significant improvement over the bound in (29) for large values of ρ; by increasing the value of n, also the value of ρ needs to be increased for observing an improvement of the lower bound in (28) over (29) (see Figure 3).
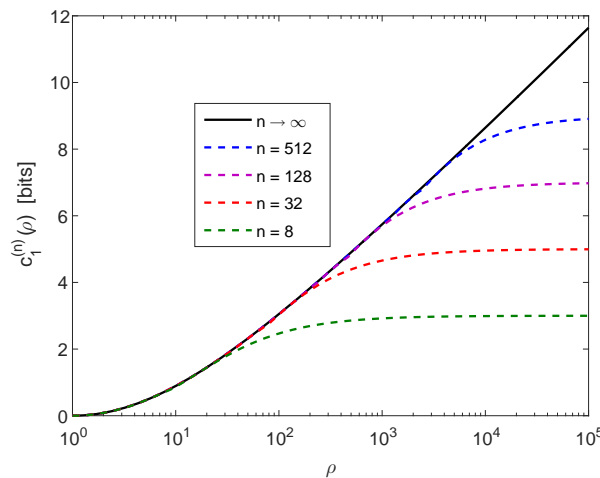

Figure 3. A plot of $c_1^{(\infty)}(\rho)$ in (22) versus $c_1^{(n)}(\rho)$ for finite n ($n = 512, 128, 32,$ and 8) as a function of ρ.

An improvement of the bound in (28) over (29) leads to a tightening of the upper bound in [12] (Theorem 4) on the compression rate of Tunstall codes for discrete memoryless sources, which further tightens the bound by Jelinek and Schneider in [69] (Equation (9)). More explicitly, in view of [12] (Section 6), an improved upper bound on the compression rate of these variable-to-fixed lossless source codes is obtained by combining [12] (Equations (36) and (38)) with a tightened lower bound on the entropy $H(W)$ of the leaves of the tree graph for Tunstall codes. From (28), the latter lower bound is given by $H(W) \geq \log_2 n - c_1^{(n)}(\rho)$ where $c_1^{(n)}(\rho)$ is expressed in bits, $\rho := \frac{1}{p_{\min}}$ is the reciprocal of the minimal positive probability of the source symbols, and n is the number of codewords (so, all codewords are of length $\lceil \log_2 n \rceil$ bits). This yields a reduction in the upper bound on the non-asymptotic compression rate R of Tunstall codes from $\frac{\lceil \log_2 n \rceil H(X)}{\log_2 n - c_1^{(\infty)}(\rho)}$ (see [12] (Equation (40)) and (22)) to $\frac{\lceil \log_2 n \rceil H(X)}{\log_2 n - c_1^{(n)}(\rho)}$ bits per source symbol where $H(X)$ denotes the source entropy (converging, in view of (17), to $H(X)$ as we let $n \to \infty$).

Remark 4. *Equality (15) with the minimizing probability mass function of the form (13) holds, in general, by replacing the Rényi entropy with an arbitrary Schur-concave function (as it can be easily verified from the proof of Lemma 2 in Appendix A). However, the analysis leading to Lemmata 3–4 and Theorem 1 applies particularly to the Rényi entropy.*

4. Bounds on the Rényi Entropy of a Function of a Discrete Random Variable

This section relies on Theorem 1 and majorization for extending [12] (Theorem 1), which applies to the Shannon entropy, to Rényi entropies of any positive order. More explicitly, let $\alpha \in (0, \infty)$ and

- $\mathcal{X}$ and $\mathcal{Y}$ be finite sets of cardinalities $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$ with $n > m \geq 2$; without any loss of generality, let $\mathcal{X} = \{1, \ldots, n\}$ and $\mathcal{Y} = \{1, \ldots, m\}$;
- X be a random variable taking values on $\mathcal{X}$ with a probability mass function $P_X \in \mathcal{P}_n$;
- $\mathcal{F}_{n,m}$ be the set of deterministic functions $f \colon \mathcal{X} \to \mathcal{Y}$; note that $f \in \mathcal{F}_{n,m}$ is not one to one since $m < n$.

The main result in this section sharpens the inequality $H_\alpha(f(X)) \leq H_\alpha(X)$, for every deterministic function $f \in \mathcal{F}_{n,m}$ with $n > m \geq 2$ and $\alpha > 0$, by obtaining non-trivial upper and lower bounds on $\max_{f \in \mathcal{F}_{n,m}} H_\alpha(f(X))$. The calculation of the exact value of $\min_{f \in \mathcal{F}_{n,m}} H_\alpha(f(X))$ is much easier, and it is expressed in closed form by capitalizing on the Schur-concavity of the Rényi entropy.

The following main result extends [12] (Theorem 1) to Rényi entropies of arbitrary positive orders.

Theorem 2. *Let $X \in \{1, \ldots, n\}$ be a random variable which satisfies $P_X(1) \geq P_X(2) \geq \ldots \geq P_X(n)$.*

(a)　*For $m \in \{2, \ldots, n-1\}$, if $P_X(1) < \frac{1}{m}$, let $\widetilde{X}_m$ be the equiprobable random variable on $\{1, \ldots, m\}$; otherwise, if $P_X(1) \geq \frac{1}{m}$, let $\widetilde{X}_m \in \{1, \ldots, m\}$ be a random variable with the probability mass function*

$$P_{\widetilde{X}_m}(i) = \begin{cases} P_X(i), & i \in \{1, \ldots, n^*\}, \\ \dfrac{1}{m - n^*} \displaystyle\sum_{j=n^*+1}^{n} P_X(j), & i \in \{n^* + 1, \ldots, m\}, \end{cases} \tag{30}$$

where n^* is the maximal integer $i \in \{1, \ldots, m-1\}$ such that

$$P_X(i) \geq \frac{1}{m-i} \sum_{j=i+1}^{n} P_X(j). \tag{31}$$

Then, for every $\alpha > 0$,

$$\max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) \in \big[H_\alpha(\widetilde{X}_m) - v(\alpha), \, H_\alpha(\widetilde{X}_m)\big], \tag{32}$$

where

$$v(\alpha) := c_\alpha^{(\infty)}(2) = \begin{cases} \log\left(\dfrac{\alpha - 1}{2^\alpha - 2}\right) - \dfrac{\alpha}{\alpha - 1} \log\left(\dfrac{\alpha}{2^\alpha - 1}\right), & \alpha \neq 1, \\ \log\left(\dfrac{2}{e \ln 2}\right) \approx 0.08607 \, bits, & \alpha = 1. \end{cases} \tag{33}$$

(b)　*There exists an explicit construction of a deterministic function $f^* \in \mathcal{F}_{n,m}$ such that*

$$H_\alpha\big(f^*(X)\big) \in \big[H_\alpha(\widetilde{X}_m) - v(\alpha), \, H_\alpha(\widetilde{X}_m)\big] \tag{34}$$

where f^* is independent of α, and it is obtained by using Huffman coding (as in [12] for $\alpha = 1$).

(c)　*Let $\widetilde{Y}_m \in \{1, \ldots, m\}$ be a random variable with the probability mass function*

$$P_{\widetilde{Y}_m}(i) = \begin{cases} \displaystyle\sum_{k=1}^{n-m+1} P_X(k), & i = 1, \\ P_X(n - m + i), & i \in \{2, \ldots, m\}. \end{cases} \tag{35}$$

Then, for every $\alpha > 0$,

$$\min_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) = H_\alpha(\widetilde{Y}_m). \tag{36}$$

Remark 5. *Setting $\alpha = 1$ specializes Theorem 2 to [12] (Theorem 1) (regarding the Shannon entropy). This point is further elaborated in Remark 8, after the proof of Theorem 2.*

Remark 6. *Similarly to [12] (Lemma 1), an exact solution of the maximization problem in the left side of (32) is strongly NP-hard [70]; this means that, unless $P = NP$, there is no polynomial time algorithm which, for an arbitrarily small $\varepsilon > 0$, computes an admissible deterministic function $f_\varepsilon \in \mathcal{F}_{n,m}$ such that*

$$H_\alpha\big(f_\varepsilon(X)\big) \geq (1 - \varepsilon) \max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big). \tag{37}$$

This motivates the derivation of the bounds in (32), and the simple construction of a deterministic function $f^* \in \mathcal{F}_{n,m}$ achieving (34).

A proof of Theorem 2 relies on the following lemmata.

Lemma 5. *Let $X \in \{1, \ldots, n\}$, $m < n$ and $\alpha > 0$. Then,*

$$\max_{Q \in \mathcal{P}_m: P_X \prec Q} H_\alpha(Q) = H_\alpha(\widetilde{X}_m) \tag{38}$$

where the probability mass function of $\widetilde{X}_m$ is given in (30).

Proof. Since $P_X \prec P_{\widetilde{X}_m}$ (see [12] (Lemma 2)) with $P_{\widetilde{X}_m} \in \mathcal{P}_m$, and $P_{\widetilde{X}_m} \prec Q$ for all $Q \in \mathcal{P}_m$ such that $P_X \prec Q$ (see [12] (Lemma 4)), the result follows from the Schur-concavity of the Rényi entropy. $\square$

Lemma 6. *Let $X \in \{1, \ldots, n\}$, $\alpha > 0$, and $f \in \mathcal{F}_{n,m}$ with $m < n$. Then,*

$$H_\alpha\big(f(X)\big) \leq H_\alpha(\widetilde{X}_m). \tag{39}$$

Proof. Since f is a deterministic function in $\mathcal{F}_{n,m}$ with $m < n$, the probability mass function of $f(X)$ is an element in $\mathcal{P}_m$ which majorizes P_X (see [12] (Lemma 3)). Inequality (39) then follows from Lemma 5. $\square$

We are now ready to prove Theorem 2.

Proof. In view of (39),

$$\max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) \leq H_\alpha(\widetilde{X}_m). \tag{40}$$

We next construct a function $f^* \in \mathcal{F}_{n,m}$ such that, for all $\alpha > 0$,

$$H_\alpha\big(f^*(X)\big) \geq \max_{Q \in \mathcal{P}_m: P_X \prec Q} H_\alpha(Q) - v(\alpha) \tag{41}$$

$$\geq \max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) - v(\alpha) \tag{42}$$

where the function $v \colon (0, \infty) \to (0, \infty)$ in the right side of (41) is given in (33), and (42) holds due to (38) and (40). The function f^* in our proof coincides with the construction in [12], and it is, therefore, independent of α.

We first review and follow the concept of the proof of [12] (Lemma 5), and we then deviate from the analysis there for proving our result. The idea behind the proof of [12] (Lemma 5) relies on the following algorithm:

(1) Start from the probability mass function $P_X \in \mathcal{P}_n$ with $P_X(1) \geq \ldots \geq P_X(n)$;
(2) Merge successively pairs of probability masses by applying the Huffman algorithm;
(3) Stop the merging process in Step 2 when a probability mass function $Q \in \mathcal{P}_m$ is obtained (with $Q(1) \geq \ldots \geq Q(m)$);
(4) Construct the deterministic function $f^* \in \mathcal{F}_{n,m}$ by setting $f^*(k) = j \in \{1, \ldots, m\}$ for all probability masses $P_X(k)$, with $k \in \{1, \ldots, n\}$, being merged in Steps 2–3 into the node of $Q(j)$.

Let $i \in \{0, \ldots, m-1\}$ be the largest index such that $P_X(1) = Q(1), \ldots, P_X(i) = Q(i)$ (note that $i = 0$ corresponds to the case where each node $Q(j)$, with $j \in \{1, \ldots, m\}$, is constructed by merging at least two masses of the probability mass function P_X). Then, according to [12] (p. 2225),

$$Q(i+1) \leq 2 Q(m). \tag{43}$$

Let

$$S := \sum_{j=i+1}^{m} Q(j) \tag{44}$$

be the sum of the $m - i$ smallest masses of the probability mass function Q. In view of (43), the vector

$$\overline{Q} := \left(\frac{Q(i+1)}{S}, \ldots, \frac{Q(m)}{S} \right) \tag{45}$$

represents a probability mass function where the ratio of its maximal to minimal masses is upper bounded by 2.

At this point, our analysis deviates from [12] (p. 2225). Applying Theorem 1 to $\overline{Q}$ with $\rho = 2$ gives

$$H_\alpha(\overline{Q}) \geq \log(m - i) - c_\alpha^{(\infty)}(2) \tag{46}$$

with

$$c_\alpha^{(\infty)}(2) = \frac{1}{\alpha - 1} \log \left(1 + \frac{1 + \alpha - 2^\alpha}{1 - \alpha} \right) - \frac{\alpha}{\alpha - 1} \log \left(1 + \frac{1 + \alpha - 2^\alpha}{(1 - \alpha)(2^\alpha - 1)} \right) \tag{47}$$

$$= \log \left(\frac{\alpha - 1}{2^\alpha - 2} \right) - \frac{\alpha}{\alpha - 1} \log \left(\frac{\alpha}{2^\alpha - 1} \right) \tag{48}$$

$$= v(\alpha) \tag{49}$$

where (47) follows from (20); (48) is straightforward algebra, and (49) is the definition in (33).

For $\alpha \in (0, 1) \cup (1, \infty)$, we get

$$H_\alpha(Q) = \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^{m} Q^\alpha(j) \right) \tag{50}$$

$$= \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^{i} Q^\alpha(j) + \sum_{j=i+1}^{m} Q^\alpha(j) \right) \tag{51}$$

$$= \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^{i} Q^\alpha(j) + S^\alpha \exp\left((1 - \alpha) H_\alpha(\overline{Q})\right) \right) \tag{52}$$

$$\geq \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^{i} Q^\alpha(j) + S^\alpha \exp\left((1 - \alpha)\left(\log(m - i) - v(\alpha)\right)\right) \right) \tag{53}$$

$$= \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^{i} Q^\alpha(j) + S^\alpha (m - i)^{1 - \alpha} \exp\left((\alpha - 1) v(\alpha)\right) \right) \tag{54}$$

where (51) holds since $i \in \{0, \ldots, m - 1\}$; (52) follows from (2) and (45); (53) holds by (46)–(49).

In view of (44), let $Q^* \in \mathcal{P}_m$ be the probability mass function which is given by

$$Q^*(j) = \begin{cases} Q(j), & j = 1, \ldots, i \\ \dfrac{S}{m - i}, & j = i + 1, \ldots, m. \end{cases} \tag{55}$$

From (50)–(55), we get

$$H_\alpha(Q) \geq \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{i} (Q^*(j))^\alpha + \sum_{j=i+1}^{m} (Q^*(j))^\alpha \exp((\alpha-1)\,v(\alpha)) \right) \tag{56}$$

$$= \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{m} (Q^*(j))^\alpha + \sum_{j=i+1}^{m} (Q^*(j))^\alpha \left(\exp((\alpha-1)\,v(\alpha)) - 1 \right) \right) \tag{57}$$

$$= H_\alpha(Q^*) + \frac{1}{1-\alpha} \log \left(1 + T \left(\exp((\alpha-1)\,v(\alpha)) - 1 \right) \right) \tag{58}$$

with

$$T := \frac{\displaystyle\sum_{j=i+1}^{m} (Q^*(j))^\alpha}{\displaystyle\sum_{j=1}^{m} (Q^*(j))^\alpha} \in [0,1]. \tag{59}$$

Since $T \in [0,1]$ and $v(\alpha) > 0$ for $\alpha > 0$, it can be verified from (56)–(58) that for $\alpha \in (0,1) \cup (1,\infty)$

$$H_\alpha(Q) \geq H_\alpha(Q^*) - v(\alpha). \tag{60}$$

The validity of (60) is extended to $\alpha = 1$ by taking the limit $\alpha \to 1$ on both sides of this inequality, and due to the continuity of $v(\cdot)$ in (33) at $\alpha = 1$. Applying the majorization result $Q^* \prec P_{\widetilde{X}_m}$ in [12] ((31)), it follows from (60) and the Schur-concavity of the Rényi entropy that, for all $\alpha > 0$,

$$H_\alpha(Q) \geq H_\alpha(Q^*) - v(\alpha) \geq H_\alpha(\widetilde{X}_m) - v(\alpha), \tag{61}$$

which together with (40), prove Items a) and b) of Theorem 2 (note that, in view of the construction of the deterministic function $f^* \in \mathcal{F}_{n,m}$ in Step 4 of the above algorithm, we get $H_\alpha(f^*(X)) = H_\alpha(Q)$).

We next prove Item c). Equality (36) is due to the Schur-concavity of the Rényi entropy, and since we have

- $f(X)$ is an aggregation of X, i.e., the probability mass function $Q \in \mathcal{P}_m$ of $f(X)$ satisfies $Q(j) = \sum_{i \in I_j} P_X(i)$ ($1 \leq j \leq m$) where $I_1, \ldots, I_m$ partition $\{1, \ldots, n\}$ into m disjoint subsets as follows:

$$I_j := \{i \in \{1, \ldots, n\} : f(i) = j\}, \quad j = 1, \ldots, m; \tag{62}$$

- By the assumption $P_X(1) \geq P_X(2) \geq \ldots \geq P_X(n)$, it follows that $Q \prec P_{\widetilde{Y}_m}$ for every such $Q \in \mathcal{P}_m$;
- From (35), $\widetilde{Y}_m = \widetilde{f}(X)$ where the function $\widetilde{f} \in \mathcal{F}_{n,m}$ is given by $\widetilde{f}(k) := 1$ for all $k \in \{1, \ldots, \text{n-m+1}\}$, and $\widetilde{f}(n - m + i) := i$ for all $i \in \{2, \ldots, m\}$. Hence, $P_{\widetilde{Y}_m}$ is an element in the set of the probability mass functions of $f(X)$ with $f \in \mathcal{F}_{n,m}$ which majorizes every other element from this set.

□

Remark 7. *The solid line in the left plot of Figure 2 depicts $v(\alpha) := c_\alpha^{(\infty)}(2)$ in (33) for $\alpha > 0$. In view of Lemma 4, and by the definition in (33), the function $v \colon (0,\infty) \to (0,\infty)$ is indeed monotonically increasing and continuous.*

Remark 8. *Inequality (43) leads to the application of Theorem 1 with $\rho = 2$ (see (46)). In the derivation of Theorem 2, we refer to $v(\alpha) := c_\alpha^{(\infty)}(2)$ (see (47)–(49)) rather than referring to $c_\alpha^{(n)}(2)$ (although, from (24), we have $0 \leq c_\alpha^{(n)}(2) \leq v(\alpha)$ for all $\alpha > 0$). We do so since, for $n \geq 16$, the difference between the curves*

of $c_\alpha^{(n)}(2)$ (as a function of $\alpha > 0$) and the curve of $c_\alpha^{(\infty)}(2)$ is marginal (see the dashed and solid lines in the left plot of Figure 2), and also because the function v in (33) is expressed in a closed form whereas $c_\alpha^{(n)}(2)$ is subject to numerical optimization for finite n (see (15) and (16)). For this reason, Theorem 2 coincides with the result in [12] (Theorem 1) for the Shannon entropy (i.e., for $\alpha = 1$) while providing a generalization of the latter result for Rényi entropies of arbitrary positive orders α. Theorem 1, however, both strengthens the bounds in [12] (Theorem 2) for the Shannon entropy with finite cardinality n (see Remark 3), and it also generalizes these bounds to Rényi entropies of all positive orders.

Remark 9. *The minimizing probability mass function in (35) to the optimization problem (36), and the maximizing probability mass function in (30) to the optimization problem (38) are in general valid when the Rényi entropy of a positive order is replaced by an arbitrary Schur-concave function. However, the main results in (32)–(34) hold particularly for the Rényi entropy.*

Remark 10. *Theorem 2 makes use of the random variables denoted by $\widetilde{X}_m$ and $\widetilde{Y}_m$, rather than (more simply) X_m and Y_m respectively, because Section 5 considers i.i.d. samples $\{X_i\}_{i=1}^k$ and $\{Y_i\}_{i=1}^k$ with $X_i \sim P_X$ and $Y_i \sim P_Y$; note, however, that the probability mass functions of $\widetilde{X}_m$ and $\widetilde{Y}_m$ are different from P_X and P_Y, respectively, and for that reason we make use of tilted symbols in the left sides of (30) and (35).*

5. Information-Theoretic Applications: Non-Asymptotic Bounds for Lossless Compression and Guessing

Theorem 2 is applied in this section to derive non-asymptotic bounds for lossless compression of discrete memoryless sources and guessing moments. Each of the two subsections starts with a short background for making the presentation self-contained.

5.1. Guessing

5.1.1. Background

The problem of guessing discrete random variables has various theoretical and operational aspects in information theory (see [35–38,40,41,43,56,71–81]). 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 successively asking questions of the form "Is X equal to x?" until the value of X is guessed correctly. 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. The required number of guesses is therefore equal to $g(x)$ when $X = x$ with $x \in \mathcal{X}$.

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 [77] and by McEliece and Yu [78], followed by a derivation of improved upper and lower bounds by De Santis et al. [80]. More generally, given a probability mass function P_X on $\mathcal{X}$, it is of interest to minimize the generalized guessing moment $\mathbb{E}[g^\rho(X)] = \sum_{x \in \mathcal{X}} P_X(x) g^\rho(x)$ for $\rho > 0$. 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 $g_X(x) = \ell$ if $P_X(x)$ is the ℓ-th largest mass [77]. 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 $\rho > 0$. Upper and lower bounds on the ρ-th moment of ranking functions, expressed in terms of the Rényi entropies, were derived by Arikan [35], Boztaş [71], followed by recent improvements in the non-asymptotic regime by Sason and Verdú [56]. 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 $X^k = (X_1, \ldots, X_k)$ whose letters belong to a finite alphabet $\mathcal{X}$; computing all the probabilities of the mass function P_{X^k} over the set $\mathcal{X}^k$, and then sorting them in decreasing order

for the calculation of the ρ-th moment of the optimal guessing function for the elements of $\mathcal{X}^k$ becomes infeasible even for moderate values of k. In contrast, regardless of the value of k, bounds on guessing moments which depend on the Rényi entropy are readily computable if for example, $\{X_i\}_{i=1}^k$ are independent; in which case, the Rényi entropy of the vector is equal to the sum of the Rényi entropies of its components. Arikan's bounds in [35] are asymptotically tight for random vectors of length k as $k \to \infty$, thus providing the correct exponential growth rate of the guessing moments for sufficiently large k.

5.1.2. Analysis

We next analyze the following setup of guessing. Let $\{X_i\}_{i=1}^k$ be i.i.d. random variables where $X_1 \sim P_X$ takes values on a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$. To cluster the data [82] (see also [12] (Section 3.A) and references therein), suppose that each X_i is mapped to $Y_i = f(X_i)$ where $f \in \mathcal{F}_{n,m}$ is an arbitrary deterministic function (independent of the index i) with $m < n$. Consequently, $\{Y_i\}_{i=1}^k$ are i.i.d., and each Y_i takes values on a finite set $\mathcal{Y}$ with $|\mathcal{Y}| = m < |\mathcal{X}|$.

Let $g_{X^k} \colon \mathcal{X}^k \to \{1, \dots, n^k\}$ and $g_{Y^k} \colon \mathcal{Y}^k \to \{1, \dots, m^k\}$ be, respectively, the ranking functions of the random vectors $X^k = (X_1, \dots, X_k)$ and $Y^k = (Y_1, \dots, Y_k)$ by sorting in separate decreasing orders the probabilities $P_{X^k}(x^k) = \prod_{i=1}^k P_X(x_i)$ for $x^k \in \mathcal{X}^k$, and $P_{Y^k}(y^k) = \prod_{i=1}^k P_Y(y_i)$ for $y^k \in \mathcal{Y}^k$ where ties in both cases are resolved arbitrarily. In view of Arikan's bounds on the ρ-th moment of ranking functions (see [35] (Theorem 1) for the lower bound, and [35] (Proposition 4) for the upper bound), since $|\mathcal{X}^k| = n^k$ and $|\mathcal{Y}^k| = m^k$, the following bounds hold for all $\rho > 0$:

$$\rho H_{\frac{1}{1+\rho}}(X) - \frac{\rho \log(1 + k \ln n)}{k} \leq \frac{1}{k} \log \mathbb{E}\big[g_{X^k}^\rho(X^k)\big] \leq \rho H_{\frac{1}{1+\rho}}(X), \tag{63}$$

$$\rho H_{\frac{1}{1+\rho}}(Y) - \frac{\rho \log(1 + k \ln m)}{k} \leq \frac{1}{k} \log \mathbb{E}\big[g_{Y^k}^\rho(Y^k)\big] \leq \rho H_{\frac{1}{1+\rho}}(Y). \tag{64}$$

In the following, we rely on Theorem 2 and the bounds in (63) and (64) to obtain bounds on the exponential reduction of the ρ-th moment of the ranking function of X^k as a result of its mapping to Y^k. First, the combination of (63) and (64) yields

$$\rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(Y) \right] - \frac{\rho \log(1 + k \ln n)}{k}$$

$$\leq \frac{1}{k} \log \frac{\mathbb{E}\big[g_{X^k}^\rho(X^k)\big]}{\mathbb{E}\big[g_{Y^k}^\rho(Y^k)\big]} \tag{65}$$

$$\leq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(Y) \right] + \frac{\rho \log(1 + k \ln m)}{k}. \tag{66}$$

In view of Theorem 2-(a) and (65), it follows that for an arbitrary $f \in \mathcal{F}_{n,m}$ and $\rho > 0$

$$\frac{1}{k} \log \frac{\mathbb{E}\big[g_{X^k}^\rho(X^k)\big]}{\mathbb{E}\big[g_{Y^k}^\rho(Y^k)\big]} \geq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right] - \frac{\rho \log(1 + k \ln n)}{k} \tag{67}$$

where $\widetilde{X}_m$ is a random variable whose probability mass function is given in (30). Please note that

$$H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \leq H_{\frac{1}{1+\rho}}(X), \quad \frac{\rho \log(1 + k \ln n)}{k} \xrightarrow[k \to \infty]{} 0 \tag{68}$$

where the first inequality in (68) holds since $P_X \prec P_{\widetilde{X}_m}$ (see Lemma 5) and the Rényi entropy is Schur-concave.

By the explicit construction of the function $f^* \in \mathcal{F}_{n,m}$ according to the algorithm in Steps 1–4 in the proof of Theorem 2 (based on the Huffman procedure), by setting $Y_i := f^*(X_i)$ for every $i \in \{1, \ldots, k\}$, it follows from (34) and (66) that for all $\rho > 0$

$$\frac{1}{k} \log \frac{\mathbb{E}\left[g_{X^k}^\rho (X^k)\right]}{\mathbb{E}\left[g_{Y^k}^\rho (Y^k)\right]} \leq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) + v\left(\frac{1}{1+\rho}\right) \right] + \frac{\rho \log(1 + k \ln m)}{k} \tag{69}$$

where the monotonically increasing function $v \colon (0, \infty) \to (0, \infty)$ is given in (33), and it is depicted by the solid line in the left plot of Figure 2. In view of (33), it can be shown that the linear approximation $v(\alpha) \approx v(1)\alpha$ is excellent for all $\alpha \in [0, 1]$, and therefore for all $\rho > 0$

$$v\left(\frac{1}{1+\rho}\right) \approx \frac{0.08607}{1+\rho} \text{ bits.} \tag{70}$$

Hence, for sufficiently large value of k, the gap between the lower and upper bounds in (67) and (69) is marginal, being approximately equal to $\frac{0.08607\,\rho}{1+\rho}$ bits for all $\rho > 0$.

The following theorem summarizes our result in this section.

Theorem 3. *Let*

- *$\{X_i\}_{i=1}^k$ be i.i.d. with $X_1 \sim P_X$ taking values on a set $\mathcal{X}$ with $|\mathcal{X}| = n$;*
- *$Y_i = f(X_i)$, for every $i \in \{1, \ldots, k\}$, where $f \in \mathcal{F}_{n,m}$ is a deterministic function with $m < n$;*
- *$g_{X^k} \colon \mathcal{X}^k \to \{1, \ldots, n^k\}$ and $g_{Y^k} \colon \mathcal{Y}^k \to \{1, \ldots, m^k\}$ be, respectively, ranking functions of the random vectors $X^k = (X_1, \ldots, X_k)$ and $Y^k = (Y_1, \ldots, Y_k)$.*

Then, for every $\rho > 0$,

- (a) *The lower bound in (67) holds for every deterministic function $f \in \mathcal{F}_{n,m}$;*
- (b) *The upper bound in (69) holds for the specific $f^* \in \mathcal{F}_{n,m}$, whose construction relies on the Huffman algorithm (see Steps 1–4 of the procedure in the proof of Theorem 2);*
- (c) *The gap between these bounds, for $f = f^*$ and sufficiently large k, is at most $\rho\, v\left(\frac{1}{1+\rho}\right) \approx \frac{0.08607\,\rho}{1+\rho}$ bits.*

5.1.3. Numerical Result

The following simple example illustrates the tightness of the achievable upper bound and the universal lower bound in Theorem 3, especially for sufficiently long sequences.

Example 1. *Let X be geometrically distributed restricted to $\{1, \ldots, n\}$ with the probability mass function*

$$P_X(j) = \frac{(1-a)\, a^{j-1}}{1 - a^n}, \quad j \in \{1, \ldots, n\} \tag{71}$$

where $a = \frac{24}{25}$ and $n = 128$. Assume that $X_1, \ldots, X_k$ are i.i.d. with $X_1 \sim P_X$, and let $Y_i = f(X_i)$ for a deterministic function $f \in \mathcal{F}_{n,m}$ with $n = 128$ and $m = 16$. We compare the upper and lower bounds in Theorem 3 for the two cases where the sequence $X^k = (X_1, \ldots, X_k)$ is of length $k = 100$ or $k = 1000$. The lower bound in (67) holds for an arbitrary deterministic $f \in \mathcal{F}_{n,m}$, and the achievable upper bound in (69) holds for the construction of the deterministic function $f = f^* \in \mathcal{F}_{n,m}$ (based on the Huffman algorithm) in Theorem 3. Numerical results are shown in Figure 4, providing plots of the upper and lower bounds on $\frac{1}{k} \log_2 \frac{\mathbb{E}\left[g_{X^k}^\rho (X^k)\right]}{\mathbb{E}\left[g_{Y^k}^\rho (Y^k)\right]}$ in Theorem 3, and illustrating the improved tightness of these bounds when the value of k is increased from 100 (left plot) to 1000 (right plot). From Theorem 3-(c), for sufficiently large k, the gap between the upper and lower bounds is less than 0.08607 bits (for all $\rho > 0$); this is consistent with the right plot of Figure 4 where $k = 1000$.
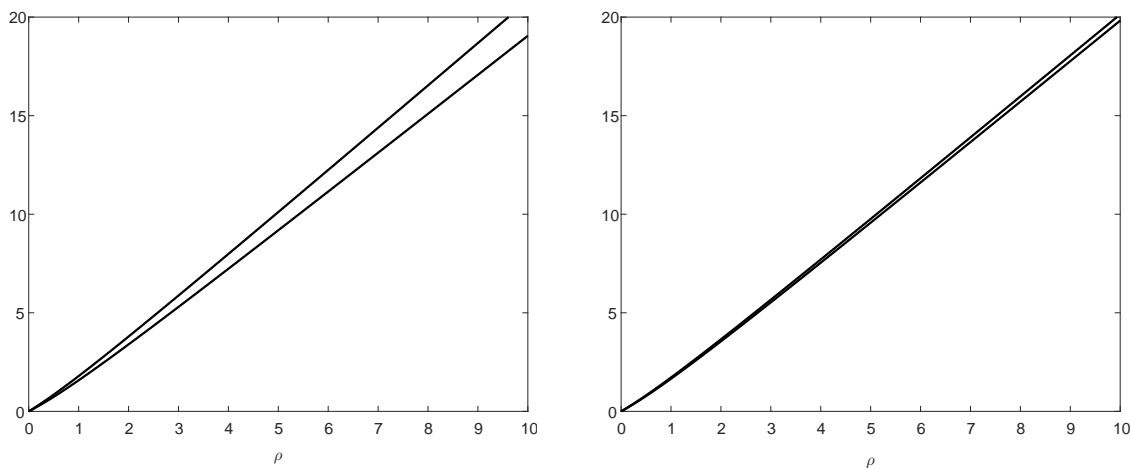
Figure 4. Plots of the upper and lower bounds on $\frac{1}{k} \log_2 \frac{\mathbb{E}\left[g^{\rho}_{X^k}(X^k)\right]}{\mathbb{E}\left[g^{\rho}_{Y^k}(Y^k)\right]}$ in Theorem 3, as a function of $\rho > 0$, for random vectors of length $k = 100$ (left plot) or $k = 1000$ (right plot) in the setting of Example 1. Each plot shows the universal lower bound for an arbitrary deterministic $f \in \mathcal{F}_{128, 16}$, and the achievable upper bound with the construction of the deterministic function $f = f^* \in \mathcal{F}_{128, 16}$ (based on the Huffman algorithm) in Theorem 3 (see, respectively, (67) and (69)).

5.2. Lossless Source Coding

5.2.1. Background

For uniquely decodable (UD) lossless source coding, Campbell [51,83] proposed the cumulant generating function of the codeword lengths as a generalization to the frequently used design criterion of average code length. Campbell's motivation in [51] 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 average code length per source symbol; on the other hand, 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.

We introduce the coding theorem by Campbell [51] for lossless compression of a discrete memoryless source (DMS) with UD codes, which serves for our analysis jointly with Theorem 2.

Theorem 4 (Campbell 1965, [51]). *Consider a DMS which emits symbols with a probability mass function P_X defined on a (finite or countably infinite) set $\mathcal{X}$. Consider a UD fixed-to-variable source code operating on source sequences of k symbols with an alphabet of the codewords of size D. Let $\ell(x^k)$ be the length of the codeword which corresponds to the source sequence $x^k := (x_1, \ldots, x_k) \in \mathcal{X}^k$. Consider the scaled cumulant generating function of the codeword lengths*

$$\Lambda_k(\rho) := \frac{1}{k} \log_D \left(\sum_{x^k \in \mathcal{X}^k} P_{X^k}(x^k) D^{\rho \, \ell(x^k)} \right), \quad \rho > 0 \tag{72}$$

where

$$P_{X^k}(x^k) = \prod_{i=1}^{k} P_X(x_i), \quad \forall \, x^k \in \mathcal{X}^k. \tag{73}$$

Then, for every $\rho > 0$, the following hold:

(a)   *Converse result:*

$$\frac{\Lambda_k(\rho)}{\rho} \geq \frac{1}{\log D} \, H_{\frac{1}{1+\rho}}(X). \tag{74}$$

(b)   *Achievability result: there exists a UD source code, for which*

$$\frac{\Lambda_k(\rho)}{\rho} \leq \frac{1}{\log D} \, H_{\frac{1}{1+\rho}}(X) + \frac{1}{k}. \tag{75}$$

The term scaled cumulant generating function is used in view of [56] (Remark 20). The bounds in Theorem 4, expressed in terms of the Rényi entropy, imply that for sufficiently long source sequences, it is possible to make the scaled cumulant generating function of the codeword lengths approach the Rényi entropy as closely as desired by a proper fixed-to-variable UD source code; moreover, the converse result shows that there is no UD source code for which the scaled cumulant generating function of its codeword lengths lies below the Rényi entropy. By invoking L'Hôpital's rule, one gets from (72)

$$\lim_{\rho \downarrow 0} \frac{\Lambda_k(\rho)}{\rho} = \frac{1}{k} \sum_{x^k \in \mathcal{X}^k} P_{X^k}(x^k) \, \ell(x^k) = \frac{1}{k} \, \mathbb{E}[\ell(X^k)]. \tag{76}$$

Hence, by letting ρ tend to zero in (74) and (75), it follows from (4) that Campbell's result in Theorem 4 generalizes the well-known bounds on the optimal average length of UD fixed-to-variable source codes (see, e.g., [84] ((5.33) and (5.37))):

$$\frac{1}{\log D} \, H(X) \leq \frac{1}{k} \, \mathbb{E}[\ell(X^k)] \leq \frac{1}{\log D} \, H(X) + \frac{1}{k}, \tag{77}$$

and (77) is satisfied by Huffman coding (see, e.g., [84] (Theorem 5.8.1)). Campbell's result therefore generalizes Shannon's fundamental result in [85] for the average codeword lengths of lossless compression codes, expressed in terms of the Shannon entropy.

Following the work by Campbell [51], Courtade and Verdú derived in [52] non-asymptotic bounds for the scaled cumulant generating function of the codeword lengths for P_X-optimal variable-length lossless codes [23,86]. These bounds were used in [52] to obtain simple proofs of the asymptotic normality of the distribution of codeword lengths, and the reliability function of memoryless sources allowing countably infinite alphabets. Sason and Verdú recently derived in [56] improved non-asymptotic bounds on the cumulant generating function of the codeword lengths for fixed-to-variable optimal lossless source coding without prefix constraints, and non-asymptotic bounds on the reliability function of a DMS, tightening the bounds in [52].

5.2.2. Analysis

The following analysis for lossless source compression with UD codes relies on a combination of Theorems 2 and 4.

Let $X_1, \ldots, X_k$ be i.i.d. symbols which are emitted from a DMS according to a probability mass function P_X whose support is a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$. Similarly to Section 5.1, to cluster the data, suppose that each symbol X_i is mapped to $Y_i = f(X_i)$ where $f \in \mathcal{F}_{n,m}$ is an arbitrary deterministic function (independent of the index i) with $m < n$. Consequently, the i.i.d. symbols $Y_1, \ldots, Y_k$ take values on a set $\mathcal{Y}$ with $|\mathcal{Y}| = m < |\mathcal{X}|$. Consider two UD fixed-to-variable source codes: one operating on the sequences $x^k \in \mathcal{X}^k$, and the other one operates on the sequences $y^k \in \mathcal{Y}^k$; let D be the size of the alphabets of both source codes. Let $\ell(x^k)$ and $\overline{\ell}(y^k)$ denote the length of the codewords for the source sequences x^k and y^k, respectively, and let $\Lambda_k(\cdot)$ and $\overline{\Lambda}_k(\cdot)$ denote their corresponding scaled cumulant generating functions (see (72)).

In view of Theorem 4-(b), for every $\rho > 0$, there exists a UD source code for the sequences in $\mathcal{X}^k$ such that the scaled cumulant generating function of its codeword lengths satisfies (75). Furthermore, from Theorem 4-(a), we get

$$\frac{\overline{\Lambda}_k(\rho)}{\rho} \geq \frac{1}{\log D} \, H_{\frac{1}{1+\rho}}(Y). \tag{78}$$

From (75), (78) and Theorem 2 (a) and (b), for every $\rho > 0$, there exist a UD source code for the sequences in $\mathcal{X}^k$, and a construction of a deterministic function $f \in \mathcal{F}_{n,m}$ (as specified by Steps 1–4 in the proof of Theorem 2, borrowed from [12]) such that the difference between the two scaled cumulant generating functions satisfies

$$\Lambda_k(\rho) - \overline{\Lambda}_k(\rho) \leq \frac{\rho}{\log D} \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) + v\left(\frac{1}{1+\rho}\right) \right] + \frac{\rho}{k}, \tag{79}$$

where (79) holds for every UD source code operating on the sequences in $\mathcal{Y}^k$ with $Y_i = f(X_i)$ (for $i = 1, \ldots, k$) and the specific construction of $f \in \mathcal{F}_{n,m}$ as above, and $\widetilde{X}_m$ in the right side of (79) is a random variable whose probability mass function is given in (30). The right side of (79) can be very well approximated (for all $\rho > 0$) by using (70).

We proceed with a derivation of a lower bound on the left side of (79). In view of Theorem 4, it follows that (74) is satisfied for every UD source code which operates on the sequences in $\mathcal{X}^k$; furthermore, Theorems 2 and 4 imply that, for every $f \in \mathcal{F}_{n,m}$, there exists a UD source code which operates on the sequences in $\mathcal{Y}^k$ such that

$$\frac{\overline{\Lambda}_k(\rho)}{\rho} \leq \frac{1}{\log D} \, H_{\frac{1}{1+\rho}}(Y) + \frac{1}{k}, \tag{80}$$

$$\leq \frac{1}{\log D} \, H_{\frac{1}{1+\rho}}(\widetilde{X}_m) + \frac{1}{k}, \tag{81}$$

where (81) is due to (39) since $Y_i = f(X_i)$ (for $i = 1, \ldots, k$) with an arbitrary deterministic function $f \in \mathcal{F}_{n,m}$, and $Y_i \sim P_Y$ for every i; hence, from (74), (80) and (81),

$$\Lambda_k(\rho) - \overline{\Lambda}_k(\rho) \geq \frac{\rho}{\log D} \left(H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right) - \frac{\rho}{k}. \tag{82}$$

We summarize our result as follows.

Theorem 5. *Let*

- $X_1, \ldots, X_k$ *be i.i.d. symbols which are emitted from a DMS according to a probability mass function P_X whose support is a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$;*
- *Each symbol X_i be mapped to $Y_i = f(X_i)$ where $f \in \mathcal{F}_{n,m}$ is the deterministic function (independent of the index i) with $m < n$, as specified by Steps 1–4 in the proof of Theorem 2 (borrowed from [12]);*
- *Two UD fixed-to-variable source codes be used: one code encodes the sequences $x^k \in \mathcal{X}^k$, and the other code encodes their mappings $y^k \in \mathcal{Y}^k$; let the common size of the alphabets of both codes be D;*
- $\Lambda_k(\cdot)$ *and $\overline{\Lambda}_k(\cdot)$ be, respectively, the scaled cumulant generating functions of the codeword lengths of the k-length sequences in $\mathcal{X}^k$ (see (72)) and their mapping to $\mathcal{Y}^k$.*

Then, for every $\rho > 0$, the following holds for the difference between the scaled cumulant generating functions $\Lambda_k(\cdot)$ and $\overline{\Lambda}_k(\cdot)$:

(a) *There exists a UD source code for the sequences in $\mathcal{X}^k$ such that the upper bound in (79) is satisfied for every UD source code which operates on the sequences in $\mathcal{Y}^k$;*

(b) There exists a UD source code for the sequences in $\mathcal{Y}^k$ such that the lower bound in (82) holds for every UD source code for the sequences in $\mathcal{X}^k$; furthermore, the lower bound in (82) holds in general for every deterministic function $f \in \mathcal{F}_{n,m}$;

(c) The gap between the upper and lower bounds in (79) and (82), respectively, is at most $\frac{\rho}{\log D} \, v\left(\frac{1}{1+\rho}\right) + \frac{2\rho}{k}$ (the function $v \colon (0, \infty) \to (0, \infty)$ is introduced in (33)), which is approximately $\frac{0.08607\rho \, \log_D 2}{1+\rho} + \frac{2\rho}{k}$;

(d) The UD source codes in Items (a) and (b) for the sequences in $\mathcal{X}^k$ and $\mathcal{Y}^k$, respectively, can be constructed to be prefix codes by the algorithm in Remark 11.

Remark 11 (An Algorithm for Theorem 5 (d)). *A construction of the UD source codes for the sequences in $\mathcal{X}^k$ and $\mathcal{Y}^k$, whose existence is assured by Theorem 5 (a) and (b) respectively, is obtained by the following algorithm (of three steps) which also constructs them as prefix codes:*

(1) As a preparatory step, we first calculate the probability mass function P_Y from the given probability mass function P_X and the deterministic function $f \in \mathcal{F}_{n,m}$ which is obtained by Steps 1–4 in the proof of Theorem 2; accordingly, $P_Y(y) = \sum_{x \in \mathcal{X}: f(x)=y} P_X(x)$ for all $y \in \mathcal{Y}$. We then further calculate the probability mass functions for the i.i.d. sequences in $\mathcal{X}^k$ and $\mathcal{Y}^k$ (see (73)); recall that the number of types in $\mathcal{X}^k$ and $\mathcal{Y}^k$ is polynomial in k (being upper bounded by $(k+1)^{n-1}$ and $(k+1)^{m-1}$, respectively), and the values of these probability mass functions are fixed over each type;

(2) The sets of codeword lengths of the two UD source codes, for the sequences in $\mathcal{X}^k$ and $\mathcal{Y}^k$, can (separately) be designed according to the achievability proof in Campbell's paper (see [51] (p. 428)). More explicitly, let $\alpha := \frac{1}{1+\rho}$; for all $x^k \in \mathcal{X}^k$, let $\ell(x^k) \in \mathbb{N}$ be given by

$$\ell(x^k) = \left\lceil -\alpha \log_D P_{X^k}(x^k) + \log_D Q_k \right\rceil \tag{83}$$

with

$$Q_k := \sum_{x^k \in \mathcal{X}^k} P_{X^k}^\alpha(x^k) = \left(\sum_{x \in \mathcal{X}} P_X^\alpha(x) \right)^k, \tag{84}$$

and let $\bar{\ell}(y^k) \in \mathbb{N}$, for all $y^k \in \mathcal{Y}^k$, be given similarly to (83) and (84) by replacing P_X with P_Y, and P_{X^k} with P_{Y^k}. This suggests codeword lengths for the two codes which fulfil (75) and (80), and also, both satisfy Kraft's inequality;

(3) The separate construction of two prefix codes (a.k.a. instantaneous codes) based on their given sets of codeword lengths $\{\ell(x^k)\}_{x^k \in \mathcal{X}^k}$ and $\{\bar{\ell}(y^k)\}_{y^k \in \mathcal{Y}^k}$, as determined in Step 2, is standard (see, e.g., the construction in the proof of [84] (Theorem 5.2.1)).

Theorem 5 is of interest since it provides upper and lower bounds on the reduction in the cumulant generating function of close-to-optimal UD source codes because of clustering data, and Remark 11 suggests an algorithm to construct such UD codes which are also prefix codes. For long enough sequences (as $k \to \infty$), the upper and lower bounds on the difference between the scaled cumulant generating functions of the suggested source codes for the original and clustered data almost match (see (79) and (82)), being roughly equal to $\rho \left(H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right)$ (with logarithms on base D, which is the alphabet size of the source codes); as $k \to \infty$, the gap between these upper and lower bounds is less than $0.08607 \log_D 2$. Furthermore, in view of (76),

$$\lim_{\rho \downarrow 0} \frac{\Lambda_k(\rho) - \overline{\Lambda}_k(\rho)}{\rho} = \frac{1}{k} \left(\mathbb{E}[\ell(X^k)] - \mathbb{E}[\bar{\ell}(Y^k)] \right), \tag{85}$$

so, it follows from (4), (33), (79) and (82) that the difference between the average code lengths (normalized by k) of the original and clustered data satisfies

$$\frac{H(X) - H(\widetilde{X}_m)}{\log D} - \frac{1}{k} \le \frac{\mathbb{E}[\ell(X^k)] - \mathbb{E}[\overline{\ell}(Y^k)]}{k} \le \frac{H(X) - H(\widetilde{X}_m) + 0.08607 \log 2}{\log D}, \tag{86}$$

where the gap between the upper and lower bounds is equal to $0.08607 \log_D 2 + \frac{1}{k}$.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Lemma 2

We first find the extreme values of $p_{\min}$ under the assumption that $P \in \mathcal{P}_n(\rho)$. If $\frac{p_{\max}}{p_{\min}} = 1$, then P is the equiprobable distribution on $\mathcal{X}$ and $p_{\min} = \frac{1}{n}$. On the other hand, if $\frac{p_{\max}}{p_{\min}} = \rho$, then the minimal possible value of $p_{\min}$ is obtained when P is the one-odd-mass distribution with $n - 1$ masses equal to $\rho p_{\min}$ and a smaller mass equal to $p_{\min}$. The latter case yields $p_{\min} = \frac{1}{1+(n-1)\rho}$.

Let $\beta := p_{\min}$, so β can get any value in the interval $\left[\frac{1}{1+(n-1)\rho}, \frac{1}{n}\right] := \Gamma_\rho^{(n)}$. From Lemma 1, $P \prec Q_\beta$ and $Q_\beta \in \mathcal{P}_n(\rho)$, and the Schur-concavity of the Rényi entropy yields $H_\alpha(P) \ge H_\alpha(Q_\beta)$ for all $P \in \mathcal{P}_n(\rho)$ with $p_{\min} = \beta$. Minimizing $H_\alpha(P)$ over $P \in \mathcal{P}_n(\rho)$ can be therefore restricted to minimizing $H_\alpha(Q_\beta)$ over $\beta \in \Gamma_\rho^{(n)}$.

Appendix B. Proof of Lemma 3

The sequence $\{c_\alpha^{(n)}(\rho)\}_{n \in \mathbb{N}}$ is non-negative since $H_\alpha(P) \le \log n$ for all $P \in \mathcal{P}_n$. To prove (17),

$$0 \le c_\alpha^{(n)}(\rho) = \log n - \min_{P \in \mathcal{P}_n(\rho)} H_\alpha(P) \tag{A1}$$

$$\le \log n - \min_{P \in \mathcal{P}_n(\rho)} H_\infty(P) \tag{A2}$$

$$\le \log n - \log \frac{n}{\rho} = \log \rho \tag{A3}$$

where (A2) holds since $H_\alpha(P)$ is monotonically decreasing in α, and (A3) is due to (5) and $p_{\max} \le \frac{\rho}{n}$.

Let U_n denote the equiprobable probability mass function on $\{1, \ldots, n\}$. By the identity

$$D_\alpha(P\|U_n) = \log n - H_\alpha(P), \tag{A4}$$

and since, by Lemma 2, $H_\alpha(\cdot)$ attains its minimum over the set of probability mass functions $\mathcal{P}_n(\rho)$, it follows that $D_\alpha(\cdot\|U_n)$ attains its maximum over this set. Let $P^* \in \mathcal{P}_n(\rho)$ be the probability measure which achieves the minimum in $c_\alpha^{(n)}(\rho)$ (see (16)), then from (A4)

$$c_\alpha^{(n)}(\rho) = \max_{P \in \mathcal{P}_n(\rho)} D_\alpha(P\|U_n) \tag{A5}$$

$$= D_\alpha(P^*\|U_n). \tag{A6}$$

Let Q^* be the probability mass function which is defined on $\{1, \ldots, 2n\}$ as follows:

$$Q^*(i) = \begin{cases} \frac{1}{2} P^*(i), & i \in \{1, \ldots, n\}, \\ \frac{1}{2} P^*(i-n), & i \in \{n+1, \ldots, 2n\}. \end{cases} \tag{A7}$$

Since by assumption $P^* \in \mathcal{P}_n(\rho)$, it is easy to verify from (A7) that

$$Q^* \in \mathcal{P}_{2n}(\rho). \tag{A8}$$

Furthermore, from (A7),

$$D_\alpha(Q^* \| U_{2n}) = \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{2n} (Q^*(i))^\alpha \left(\frac{1}{2n} \right)^{1-\alpha} \right) \tag{A9}$$

$$= \frac{1}{\alpha - 1} \log \left(\frac{1}{2} \sum_{i=1}^{n} (P^*(i))^\alpha \left(\frac{1}{n} \right)^{1-\alpha} + \frac{1}{2} \sum_{i=n+1}^{2n} (P^*(i-n))^\alpha \left(\frac{1}{n} \right)^{1-\alpha} \right) \tag{A10}$$

$$= \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{n} (P^*(i))^\alpha \left(\frac{1}{n} \right)^{1-\alpha} \right) \tag{A11}$$

$$= D_\alpha(P^* \| U_n). \tag{A12}$$

Combining (A5)–(A12) yields

$$c_\alpha^{(2n)}(\rho) = \max_{Q \in \mathcal{P}_{2n}(\rho)} D_\alpha(Q \| U_{2n}) \tag{A13}$$

$$\geq D_\alpha(Q^* \| U_{2n}) \tag{A14}$$

$$= D_\alpha(P^* \| U_n) \tag{A15}$$

$$= c_\alpha^{(n)}(\rho), \tag{A16}$$

proving (18). Finally, in view of (A5), $c_\alpha^{(n)}(\rho)$ is monotonically increasing in α since so is the Rényi divergence of order α (see [87] (Theorem 3)).

Appendix C. Proof of Lemma 4

From Lemma 2, the minimizing distribution of H_α is given by $Q_\beta \in \mathcal{P}_n(\rho)$ where

$$Q_\beta = \Big(\underbrace{\rho\beta, \ldots, \rho\beta}_{i}, \ 1 - (n + i\rho - i - 1)\beta, \ \underbrace{\beta, \beta, \ldots, \beta}_{n-i-1} \Big) \tag{A17}$$

with $\beta \in \left[\frac{1}{1+(n-1)\rho}, \frac{1}{n} \right]$, and $1 - (n + i\rho - i - 1)\beta \leq \rho\beta \leq \frac{\rho}{n}$. It therefore follows that the influence of the middle probability mass of Q_β on $H_\alpha(Q_\beta)$ tends to zero as $n \to \infty$. Therefore, in this asymptotic case, one can instead minimize $H_\alpha(\widetilde{Q}_m)$ where

$$\widetilde{Q}_m = \Big(\underbrace{\rho\beta, \ldots, \rho\beta}_{m}, \ \underbrace{\beta, \beta, \ldots, \beta}_{n-m} \Big) \tag{A18}$$

with the free parameter $m \in \{1, \ldots, n\}$ and $\beta = \frac{1}{n+m(\rho-1)}$ (so that the total mass of $\widetilde{Q}_m$ is equal to 1).

For $\alpha \in (0,1) \cup (1,\infty)$, straightforward calculation shows that

$$
\begin{aligned}
H_\alpha(\widetilde{Q}_m) &= \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{n} \widetilde{Q}_m^\alpha(j) \right) \\
&= \log n - \frac{1}{\alpha - 1} \log \left(\frac{1 + \frac{m}{n}(\rho^\alpha - 1)}{\left(1 + \frac{m}{n}(\rho - 1)\right)^\alpha} \right),
\end{aligned}
\tag{A19}
$$

and by letting $n \to \infty$, the limit of the sequence $\{c_\alpha^{(n)}(\rho)\}_{n \in \mathbb{N}}$ exists, and it is equal to

$$
\begin{aligned}
c_\alpha^{(\infty)}(\rho) &:= \lim_{n \to \infty} c_\alpha^{(n)}(\rho) \\
&= \lim_{n \to \infty} \left(\log n - \min_{m \in \{1,\dots,n\}} H_\alpha(\widetilde{Q}_m) \right) \\
&= \lim_{n \to \infty} \max_{m \in \{1,\dots,n\}} \left\{ \frac{1}{\alpha - 1} \log \left(\frac{1 + \frac{m}{n}(\rho^\alpha - 1)}{\left(1 + \frac{m}{n}(\rho - 1)\right)^\alpha} \right) \right\} \\
&= \max_{x \in [0,1]} \left\{ \frac{1}{\alpha - 1} \log \left(\frac{1 + (\rho^\alpha - 1)x}{(1 + (\rho - 1)x)^\alpha} \right) \right\}.
\end{aligned}
\tag{A20}
$$

Let $f_\alpha : [0,1] \to \mathbb{R}$ be given by

$$
f_\alpha(x) = \frac{1 + (\rho^\alpha - 1)x}{(1 + (\rho - 1)x)^\alpha}, \quad x \in [0,1].
\tag{A21}
$$

Then, $f_\alpha(0) = f_\alpha(1) = 1$, and straightforward calculation shows that the derivative f_α' vanishes if and only if

$$
x = x^* := \frac{1 + \alpha(\rho - 1) - \rho^\alpha}{(1 - \alpha)(\rho - 1)(\rho^\alpha - 1)}.
\tag{A22}
$$

We rely here on a specialized version of the mean value theorem, known as Rolle's theorem, which states that any real-valued differentiable function that attains equal values at two distinct points must have a point somewhere between them where the first derivative at this point is zero. By Rolle's theorem, and due to the uniqueness of the point x^* in (A22), it follows that $x^* \in (0,1)$. Substituting (A22) into (A20) gives (20). Taking the limit of (20) when $\alpha \to \infty$ gives the result in (21).

In the limit where $\alpha \to 1$, the Rényi entropy of order α tends to the Shannon entropy. Hence, letting $\alpha \to 1$ in (20), it follows that for the Shannon entropy

$$
\begin{aligned}
c_1^{(\infty)}(\rho) &= \lim_{\alpha \to 1} c_\alpha^{(\infty)}(\rho) \\
&= \lim_{\alpha \to 1} \left\{ \frac{1}{\alpha - 1} \log \left(1 + \frac{1 + \alpha(\rho - 1) - \rho^\alpha}{(1 - \alpha)(\rho - 1)} \right) - \frac{\alpha}{\alpha - 1} \log \left(1 + \frac{1 + \alpha(\rho - 1) - \rho^\alpha}{(1 - \alpha)(\rho^\alpha - 1)} \right) \right\} \\
&= \frac{\rho \log \rho}{\rho - 1} - \log e - \log \left(\frac{\rho \log_e \rho}{\rho - 1} \right),
\end{aligned}
\tag{A23}
$$

where (A23) follows by invoking L'Hôpital's rule. This proves (22).

From (17)–(19), we get $0 \le c_\alpha^{(n)}(\rho) \le c_\alpha^{(\infty)}(\rho)$. Since $c_\alpha^{(n)}(\rho)$ is monotonically increasing in $\alpha \in [0, \infty]$, for every $n \in \mathbb{N}$, so is $c_\alpha^{(\infty)}(\rho)$; hence, (21) yields $c_\alpha^{(\infty)}(\rho) \le \log \rho$. This proves (24).

References

1. Hardy, G.H.; Littlewood, J.E.; Pólya, G. *Inequalities*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1952.
2. Marshall, A.W.; Olkin, I.; Arnold, B.C. *Inequalities: Theory of Majorization and Its Applications*, 2nd ed.; Springer: New York, NY, USA, 2011.
3. Arnold, B.C. Majorization: Here, there and everywhere. *Stat. Sci.* **2007**, *22*, 407–413. [CrossRef]
4. Arnold, B.C.; Sarabia, J.M. *Majorization and the Lorenz Order with Applications in Applied Mathematics and Economics*; Statistics for Social and Behavioral Sciences; Springer: New York, NY, USA, 2018.
5. Cicalese, F.; Gargano, L.; Vaccaro, U. Information theoretic measures of distances and their econometric applications. In Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 409–413.
6. Steele, J.M. *The Cauchy-Schwarz Master Class*; Cambridge University Press: Cambridge, UK, 2004.

7.　Bhatia, R. *Matrix Analysis Graduate Texts in Mathematics*; Springer: New York, NY, USA, 1997.

8.　Horn, R.A.; Johnson, C.R. *Matrix Analysis*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013.

9.　Ben-Bassat, M.; Raviv, J. Rényi's entropy and probability of error. *IEEE Trans. Inf. Theory* **1978**, *24*, 324–331. [CrossRef]

10.　Cicalese, F.; Vaccaro, U. Bounding the average length of optimal source codes via majorization theory. *IEEE Trans. Inf. Theory* **2004**, *50*, 633–637. [CrossRef]

11.　Cicalese, F.; Gargano, L.; Vaccaro, U. How to find a joint probability distribution with (almost) minimum entropy given the marginals. In Proceedings of the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 2178–2182.

12.　Cicalese, F.; Gargano, L.; Vaccaro, U. Bounds on the entropy of a function of a random variable and their applications. *IEEE Trans. Inf. Theory* **2018**, *64*, 2220–2230. [CrossRef]

13.　Cicalese, F.; Vaccaro, U. Maximum entropy interval aggregations. In Proceedings of the 2018 IEEE International Symposium on Information Theory, Vail, CO, USA, 17–20 June 2018; pp. 1764–1768.

14.　Harremoës, P. A new look on majorization. In Proceedings of the 2004 IEEE International Symposium on Information Theory and Its Applications, Parma, Italy, 10–13 October 2004; pp. 1422–1425.

15.　Ho, S.W.; Yeung, R.W. The interplay between entropy and variational distance. *IEEE Trans. Inf. Theory* **2010**, *56*, 5906–5929. [CrossRef]

16.　Ho, S.W.; Verdú S. On the interplay between conditional entropy and error probability. *IEEE Trans. Inf. Theory* **2010**, *56*, 5930–5942. [CrossRef]

17.　Ho, S.W.; Verdú, S. Convexity/concavity of the Rényi entropy and α-mutual information. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong Kong, China, 14–19 June 2015; pp. 745–749.

18.　Joe, H. Majorization, entropy and paired comparisons. *Ann. Stat.* **1988**, *16*, 915–925. [CrossRef]

19.　Joe, H. Majorization and divergence. *J. Math. Anal. Appl.* **1990**, *148*, 287–305. [CrossRef]

20.　Koga, H. Characterization of the smooth Rényi entropy using majorization. In Proceedings of the 2013 IEEE Information Theory Workshop, Seville, Spain, 9–13 September 2013; pp. 604–608.

21.　Puchala, Z.; Rudnicki, L.; Zyczkowski, K. Majorization entropic uncertainty relations. *J. Phys. A Math. Theor.* **2013**, *46*, 1–12. [CrossRef]

22.　Sason, I.; Verdú S. Arimoto-Rényi conditional entropy and Bayesian *M*-ary hypothesis testing. *IEEE Trans. Inf. Theory* **2018**, *64*, 4–25. [CrossRef]

23.　Verdú S. *Information Theory*; 2018, in preparation.

24.　Witsenhhausen, H.S. Some aspects of convexity useful in information theory. *IEEE Trans. Inf. Theory* **1980**, *26*, 265–271. [CrossRef]

25.　Xi, B.; Wang, S.; Zhang, T. Schur-convexity on generalized information entropy and its applications. In *Information Computing and Applications*; Lecture Notes in Computer Science; Springer: New York, NY, USA, 2011; Volume 7030, pp. 153–160.

26.　Inaltekin, H.; Hanly, S.V. Optimality of binary power control for the single cell uplink. *IEEE Trans. Inf. Theory* **2012**, *58*, 6484–6496. [CrossRef]

27.　Jorshweick, E.; Bosche, H. Majorization and matrix-monotone functions in wireless communications. *Found. Trends Commun. Inf. Theory* **2006**, *3*, 553–701. [CrossRef]

28.　Palomar, D.P.; Jiang, Y. MIMO transceiver design via majorization theory. *Found. Trends Commun. Inf. Theory* **2006**, *3*, 331–551. [CrossRef]

29.　Roventa, I. *Recent Trends in Majorization Theory and Optimization: Applications to Wireless Communications*; Editura Pro Universitaria & Universitaria Craiova: Bucharest, Romania, 2015.

30.　Sezgin, A.; Jorswieck, E.A. Applications of majorization theory in space-time cooperative communications. In *Cooperative Communications for Improved Wireless Network Transmission: Framework for Virtual Antenna Array Applications*; Uysal, M., Ed.; Information Science Reference; IGI Global: Hershey, PA, USA, 2010; pp. 429–470.

31.　Viswanath, P.; Anantharam, V. Optimal sequences and sum capacity of synchronous CDMA systems. *IEEE Trans. Inf. Theory* **1999**, *45*, 1984–1993. [CrossRef]

32.　Viswanath, P.; Anantharam, V.; Tse, D.N.C. Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers. *IEEE Trans. Inf. Theory* **1999**, *45*, 1968–1983. [CrossRef]

33. Viswanath, P.; Anantharam, V. Optimal sequences for CDMA under colored noise: A Schur-saddle function property. *IEEE Trans. Inf. Theory* **2002**, *48*, 1295–1318. [CrossRef]

34. Rényi, A. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Probability Theory and Mathematical Statistics, Berkeley, CA, USA, 8–9 August 1961; pp. 547–561.

35. Arikan, E. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Inf. Theory* **1996**, *42*, 99–105. [CrossRef]

36. Arikan, E.; Merhav, N. Guessing subject to distortion. *IEEE Trans. Inf. Theory* **1998**, *44*, 1041–1056. [CrossRef]

37. Arikan, E.; Merhav, N. Joint source-channel coding and guessing with application to sequential decoding. *IEEE Trans. Inf. Theory* **1998**, *44*, 1756–1769. [CrossRef]

38. Burin, A.; Shayevitz, O. Reducing guesswork via an unreliable oracle. *IEEE Trans. Inf. Theory* **2018**, *64*, 6941–6953. [CrossRef]

39. Kuzuoka, S. On the conditional smooth Rényi entropy and its applications in guessing and source coding. *arXiv* **2018**, arXiv:1810.09070.

40. Merhav, N.; Arikan, E. The Shannon cipher system with a guessing wiretapper. *IEEE Trans. Inf. Theory* **1999**, *45*, 1860–1866. [CrossRef]

41. Sundaresan, R. Guessing based on length functions. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 716–719.

42. Salamatian, S.; Beirami, A.; Cohen, A.; Médard, M. Centralized versus decentralized multi-agent guesswork. In Proceedings of the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 2263–2267.

43. Sundaresan, R. Guessing under source uncertainty. *IEEE Trans. Inf. Theory* **2007**, *53*, 269–287. [CrossRef]

44. Bracher, A.; Lapidoth, A.; Pfister, C. Distributed task encoding. In Proceedings of the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 1993–1997.

45. Bunte, C.; Lapidoth, A. Encoding tasks and Rényi entropy. *IEEE Trans. Inf. Theory* **2014**, *60*, 5065–5076. [CrossRef]

46. Shayevitz, O. On Rényi measures and hypothesis testing. In Proceedings of the 2011 IEEE International Symposium on Information Theory, Saint Petersburg, Russia, 31 July–5 August 2011; pp. 800–804.

47. Tomamichel, M.; Hayashi, M. Operational interpretation of Rényi conditional mutual information via composite hypothesis testing against Markov distributions. In Proceedings of the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 585–589.

48. Harsha, P.; Jain, R.; McAllester, D.; Radhakrishnan, J. The communication complexity of correlation. *IEEE Trans. Inf. Theory* **2010**, *56*, 438–449. [CrossRef]

49. Liu, J.; Verdú S. Rejection sampling and noncausal sampling under moment constraints. In Proceedings of the 2018 IEEE International Symposium on Information Theory, Vail, CO, USA, 17–22 June 2018; pp. 1565–1569.

50. Yu, L.; Tan, V.Y.F. Wyner's common information under Rényi divergence measures. *IEEE Trans. Inf. Theory* **2018**, *64*, 3616–3632. [CrossRef]

51. Campbell, L.L. A coding theorem and Rényi's entropy. *Inf. Control* **1965**, *8*, 423–429. [CrossRef]

52. Courtade, T.; Verdú S. Cumulant generating function of codeword lengths in optimal lossless compression. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2494–2498.

53. Courtade, T.; Verdú S. Variable-length lossy compression and channel coding: Non-asymptotic converses via cumulant generating functions. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2499–2503.

54. Hayashi, M.; Tan, V.Y.F. Equivocations, exponents, and second-order coding rates under various Rényi information measures. *IEEE Trans. Inf. Theory* **2017**, *63*, 975–1005. [CrossRef]

55. Kuzuoka, S. On the smooth Rényi entropy and variable-length source coding allowing errors. In Proceedings of the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 745–749.

56. Sason, I.; Verdú S. Improved bounds on lossless source coding and guessing moments via Rényi measures. *IEEE Trans. Inf. Theory* **2018**, *64*, 4323–4346. [CrossRef]

57. Tan, V.Y.F.; Hayashi, M. Analysis of ramaining uncertainties and exponents under various conditional Rényi entropies. *IEEE Trans. Inf. Theory* **2018**, *64*, 3734–3755. [CrossRef]

58. Tyagi, H. Coding theorems using Rényi information measures. In Proceedings of the 2017 IEEE Twenty-Third National Conference on Communications, Chennai, India, 2–4 March 2017; pp. 1–6.

59. Csiszár, I. Generalized cutoff rates and Rényi information measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 26–34. [CrossRef]

60. Polyanskiy, Y.; Verdú S. Arimoto channel coding converse and Rényi divergence. In Proceedings of the Forty-Eighth Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, USA, 29 September–1 October 2010; pp. 1327–1333.

61. Sason, I. On the Rényi divergence, joint range of relative entropies, and a channel coding theorem. *IEEE Trans. Inf. Theory* **2016**, *62*, 23–34. [CrossRef]

62. Yu, L.; Tan, V.Y.F. Rényi resolvability and its applications to the wiretap channel. In *Lecture Notes in Computer Science, Proceedings of the 10th International Conference on Information Theoretic Security, Hong Kong, China, 29 November–2 December 2017*; Springer: New York, NY, USA, 2017; Volume 10681, pp. 208–233.

63. Arimoto, S. On the converse to the coding theorem for discrete memoryless channels. *IEEE Trans. Inf. Theory* **1973**, *19*, 357–359. [CrossRef]

64. Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. In Proceedings of the 2nd Colloquium on Information Theory, Keszthely, Hungary, 25–30 August 1975; Csiszár, I., Elias, P., Eds.; Colloquia Mathematica Societatis János Bolyai: Amsterdam, The Netherlands, 1977; Volume 16, pp. 41–52.

65. Dalai, M. Lower bounds on the probability of error for classical and classical-quantum channels. *IEEE Trans. Inf. Theory* **2013**, *59*, 8027–8056. [CrossRef]

66. Leditzky, F.; Wilde, M.M.; Datta, N. Strong converse theorems using Rényi entropies. *J. Math. Phys.* **2016**, *57*, 1–33. [CrossRef]

67. Mosonyi, M.; Ogawa, T. Quantum hypothesis testing and the operational interpretation of the quantum Rényi relative entropies. *Commun. Math. Phys.* **2015**, *334*, 1617–1648. [CrossRef]

68. Simic, S. Jensen's inequality and new entropy bounds. *Appl. Math. Lett.* **2009**, *22*, 1262–1265. [CrossRef]

69. Jelineck, F.; Schneider, K.S. On variable-length-to-block coding. *IEEE Trans. Inf. Theory* **1972**, *18*, 765–774. [CrossRef]

70. Garey, M.R.; Johnson, D.S. *Computers and Intractability: A Guide to the Theory of NP-Completness*; W. H. Freedman and Company: New York, NY, USA, 1979.

71. Boztaş S. Comments on "An inequality on guessing and its application to sequential decoding". *IEEE Trans. Inf. Theory* **1997**, *43*, 2062–2063. [CrossRef]

72. Bracher, A.; Hof, E.; Lapidoth, A. Guessing attacks on distributed-storage systems. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong-Kong, China, 14–19 June 2015; pp. 1585–1589.

73. Christiansen, M.M.; Duffy, K.R. Guesswork, large deviations, and Shannon entropy. *IEEE Trans. Inf. Theory* **2013**, *59*, 796–802. [CrossRef]

74. Hanawal, M.K.; Sundaresan, R. Guessing revisited: A large deviations approach. *IEEE Trans. Inf. Theory* **2011**, *57*, 70–78. [CrossRef]

75. Hanawal, M.K.; Sundaresan, R. The Shannon cipher system with a guessing wiretapper: General sources. *IEEE Trans. Inf. Theory* **2011**, *57*, 2503–2516. [CrossRef]

76. Huleihel, W.; Salamatian, S.; Médard, M. Guessing with limited memory. In Proceedings of the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 2258–2262.

77. Massey, J.L. Guessing and entropy. In Proceedings of the 1994 IEEE International Symposium on Information Theory, Trondheim, Norway, 27 June–1 July 1994; p. 204.

78. McEliece, R.J.; Yu, Z. An inequality on entropy. In Proceedings of the 1995 IEEE International Symposium on Information Theory, Whistler, BC, Canada, 17–22 September 1995; p. 329.

79. Pfister, C.E.; Sullivan, W.G. Rényi entropy, guesswork moments and large deviations. *IEEE Trans. Inf. Theory* **2004**, *50*, 2794–2800. [CrossRef]

80. De Santis, A.; Gaggia, A.G.; Vaccaro, U. Bounds on entropy in a guessing game. *IEEE Trans. Inf. Theory* **2001**, *47*, 468–473. [CrossRef]

81. Yona, Y.; Diggavi, S. The effect of bias on the guesswork of hash functions. In Proceedings of the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 2253–2257.

82. Gan, G.; Ma, C.; Wu, J. *Data Clustering: Theory, Algorithms, and Applications*; ASA-SIAM Series on Statistics and Applied Probability; SIAM: Philadelphia, PA, USA, 2007.

83. Campbell, L.L. Definition of entropy by means of a coding problem. *Probab. Theory Relat. Field* **1966**, *6*, 113–118. [CrossRef]

84. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.

85. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

86. Kontoyiannis, I.; Verdú S. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inf. Theory* **2014**, *60*, 777–795. [CrossRef]

87. Van Erven, T.; Harremoës, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]