# Encoding Tasks and Rényi Entropy

Christoph Bunte and Amos Lapidoth, *Fellow, IEEE*

***Abstract*— A task is randomly drawn from a finite set of tasks and is described using a fixed number of bits. All the tasks that share its description must be performed. Upper and lower bounds on the minimum $\rho$th moment of the number of performed tasks are derived. The case where a sequence of tasks is produced by a source and $n$ tasks are jointly described using $nR$ bits is considered. If $R$ is larger than the Rényi entropy rate of the source of order $1/(1+\rho)$ (provided it exists), then the $\rho$th moment of the ratio of performed tasks to $n$ can be driven to one as $n$ tends to infinity. If $R$ is smaller than the Rényi entropy rate, this moment tends to infinity. The results are generalized to account for the presence of side-information. In this more general setting, the key quantity is a conditional version of Rényi entropy that was introduced by Arimoto. For IID sources, two additional extensions are solved, one of a rate-distortion flavor and the other where different tasks may have different nonnegative costs. Finally, a divergence that was identified by Sundaresan as a mismatch penalty in the Massey-Arikan guessing problem is shown to play a similar role here.**

***Index Terms*— Divergence, Rényi entropy, Rényi entropy rate, mismatch, source coding, tasks.**

## I. Introduction

A TASK $X$ that is drawn from a finite set of tasks $\mathcal{X}$ according to some probability mass function (PMF) $P$ is to be described using a fixed number of bits. The least number of bits needed for an unambiguous description is the base-2 logarithm of the total number of tasks in $\mathcal{X}$ (rounded up to the nearest integer). When fewer bits are available, the classical source coding approach is to provide descriptions for the tasks with the largest or with the "typical" probabilities only. This has the obvious drawback that less common, or atypical, tasks will never be completed. For example, if $\mathcal{X}$ comprises all possible household chores, then "wash the dishes" will almost certainly occur more frequently than "take out the garbage", but most people would agree that the latter should not be neglected.

The classical approach is not so well-suited here because it does not take into account the fact that not performing an unlikely but critical task may have grave consequences, and that performing a superfluous task often causes little or no harm. A more natural approach in this context is to partition the set of tasks into subsets. If a particular task needs to be completed, then the subset containing it is described and all the tasks in this subset are performed. This approach has the disadvantage that tasks are sometimes completed superfluously, but it guarantees that critical tasks, no matter how atypical, are never neglected (provided that the number of subsets in the partition of $\mathcal{X}$ does not exceed $M$, when $\log M$ is the number of bits available to describe them). One way to partition the set of tasks is to provide distinct descriptions for the typical tasks and to group together the atypical ones. We will see, however, that this may not always be optimal.

If we assume for simplicity that all tasks require an equal amount of effort, then it seems reasonable to choose the subsets so as to minimize the expected number of performed tasks. Ideally, this expectation is close to one. More generally, we look at the $\rho$-th moment of the number of performed tasks, where $\rho$ may be any positive number. Phrased in mathematical terms, we consider encoders of the form

$$f\colon \mathcal{X} \to \{1,\ldots,M\}, \tag{1}$$

where $M$ is a given positive integer. Every such encoder gives rise to a partition of $\mathcal{X}$ into $M$ disjoint subsets

$$f^{-1}(m) = \bigl\{x \in \mathcal{X} : f(x) = m\bigr\}, \quad m \in \{1,\ldots,M\}. \tag{2}$$

Here $f(x)$ is the description of the task $x$, and the set $f^{-1}(f(x))$ comprises all the tasks sharing the same description as $x$, i.e., the set of tasks that are performed when $x$ is required.

We seek an $f$ that minimizes the $\rho$-th moment of the cardinality of $f^{-1}(f(X))$, i.e.,

$$\mathrm{E}\bigl[|f^{-1}(f(X))|^{\rho}\bigr] = \sum_{x\in\mathcal{X}} P(x)|f^{-1}(f(x))|^{\rho}. \tag{3}$$

This minimum is at least 1 because $X \in f^{-1}(f(X))$; it is nonincreasing in $M$ (because fewer tasks share the same description when $M$ grows); and it is equal to one for $M \geq |\mathcal{X}|$ (because then $\mathcal{X}$ can be partitioned into singletons). Our first result is a pair of lower and upper bounds on this minimum. The bounds are expressed in terms of the *Rényi entropy of X of order* $1/(1+\rho)$

$$H_{\frac{1}{1+\rho}}(X) = \frac{1+\rho}{\rho}\log\sum_{x\in\mathcal{X}} P(x)^{\frac{1}{1+\rho}}. \tag{4}$$

Throughout $\log(\cdot)$ stands for $\log_2(\cdot)$, the logarithm to base 2. For typographic reasons we henceforth use the notation

$$\tilde{\rho} = \frac{1}{1+\rho}, \quad \rho > 0. \tag{5}$$

*Theorem I.1: Let X be a chance variable taking value in a finite set $\mathcal{X}$, and let $\rho > 0$.*

1) *For all positive integers M and every $f\colon \mathcal{X} \to \{1,\ldots,M\}$,*

$$\mathrm{E}\bigl[|f^{-1}(f(X))|^{\rho}\bigr] \geq 2^{\rho(H_{\tilde{\rho}}(X)-\log M)}. \tag{6}$$

Manuscript received January 24, 2014; accepted May 6, 2014. Date of publication June 6, 2014; date of current version August 14, 2014. This paper was presented in part at the 2013 IEEE Information Theory Workshop and in part at the 2014 IEEE International Symposium on Information Theory.

The authors are with the Signal and Information Processing Laboratory, ETH Zurich, Zurich 8092, Switzerland (e-mail: bunte@isi.ee.ethz.ch; lapidoth@isi.ee.ethz.ch).

Communicated by R. Sundaresan, Associate Editor for Communications.
Digital Object Identifier 10.1109/TIT.2014.2329490

2) *For all integers $M > \log|\mathcal{X}| + 2$ there exists $f\colon \mathcal{X} \to \{1, \dots, M\}$ such that*

$$\mathrm{E}\big[|f^{-1}(f(X))|^{\rho}\big] < 1 + 2^{\rho(H_{\tilde{\rho}}(X) - \log \widetilde{M})}, \tag{7}$$

*where $\widetilde{M} = (M - \log|\mathcal{X}| - 2)/4$.*

A proof is provided in Section III. Theorem I.1 is particularly useful when applied to the case where a sequence of tasks is produced by a source $\{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X}$ and the first $n$ tasks $X^n = (X_1, \dots, X_n)$ are jointly described using $nR$ bits (the number $R$ is the *rate* of the description in bits per task and can be any nonnegative number):

*Theorem I.2: Let $\{X_i\}_{i=1}^{\infty}$ be a source with finite alphabet $\mathcal{X}$, and let $\rho > 0$.*

1) *If $R > \limsup_{n\to\infty} H_{\tilde{\rho}}(X^n)/n$, then there exist encoders $f_n\colon \mathcal{X}^n \to \{1, \dots, 2^{nR}\}$ such that*[1]

$$\lim_{n\to\infty} \mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] = 1. \tag{8}$$

2) *If $R < \liminf_{n\to\infty} H_{\tilde{\rho}}(X^n)/n$, then for any choice of encoders $f_n\colon \mathcal{X}^n \to \{1, \dots, 2^{nR}\}$,*

$$\lim_{n\to\infty} \mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] = \infty. \tag{9}$$

*Proof:* On account of Theorem I.1, for all $n$ large enough so that $2^{nR} > n\log|\mathcal{X}| + 2$,

$$\begin{aligned} 2^{n\rho\left(\frac{H_{\tilde{\rho}}(X^n)}{n} - R\right)} &\leq \min_{f_n\colon \mathcal{X}^n \to \{1,\dots,2^{nR}\}} \mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] \\ &< 1 + 2^{n\rho\left(\frac{H_{\tilde{\rho}}(X^n)}{n} - R + \delta_n\right)}, \end{aligned} \tag{10}$$

where $\delta_n \to 0$ as $n \to \infty$. □

When it exists, the limit

$$H_{\alpha}(\{X_i\}_{i=1}^{\infty}) \triangleq \lim_{n\to\infty} \frac{H_{\alpha}(X^n)}{n} \tag{11}$$

is called the *Rényi entropy rate of $\{X_i\}_{i=1}^{\infty}$ of order $\alpha$*. It exists for a large class of sources, including time-invariant Markov sources [1], [2].

If we assume that every $n$-tuple of tasks in $f_n^{-1}(f_n(X^n))$ is performed (even if this means that some tasks are performed multiple times) and thus that the total number of performed tasks is $n$ times $|f_n^{-1}(f_n(X^n))|$, then Theorem I.2 furnishes the following operational characterization of the Rényi entropy rate for all orders in $(0, 1)$. For all rates above the Rényi entropy rate of order $1/(1+\rho)$, the $\rho$-th moment of the ratio of performed tasks to $n$ can be driven to one as $n$ tends to infinity. For all rates below it, this moment grows to infinity. In fact, the proof of Theorem I.2 shows that for large $n$ it grows exponentially in $n$ with exponent approaching

$$\rho\big(H_{\tilde{\rho}}(\{X_i\}_{i=1}^{\infty}) - R\big). \tag{12}$$

More precisely, (10) shows that for all rates $R < H_{\tilde{\rho}}(\{X_i\}_{i=1}^{\infty})$,

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \log \min_{f_n\colon \mathcal{X}^n \to \{1,\dots,2^{nR}\}} \mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] \\ &\quad = \rho\big(H_{\tilde{\rho}}(\{X_i\}_{i=1}^{\infty}) - R\big). \end{aligned} \tag{13}$$

Note that for IID sources the Rényi entropy rate reduces to the Rényi entropy because in this case $H_{\tilde{\rho}}(X^n) = nH_{\tilde{\rho}}(X_1)$. Other operational characterizations of the Rényi entropy rate were given in [1]–[6], and of the Rényi entropy in [7]–[10].

The connection between the problem of encoding tasks and the Massey-Arikan guessing problem [10], [11] is explored in [12].

The operational characterization of Rényi entropy provided by Theorem I.2 (applied to IID sources) reveals many of the known properties of Rényi entropy (see [9], [13]). For example, it shows that $H_{\tilde{\rho}}(X)$ is nondecreasing in $\rho$ because $\xi^{\rho}$ is nondecreasing in $\rho$ when $\xi \geq 1$. It also shows that

$$H(X) \leq H_{\tilde{\rho}}(X) \leq \log|\mathrm{supp}(P)|, \tag{14}$$

where $H(X)$ denotes the Shannon entropy and $\mathrm{supp}(P) = \{x : P(x) > 0\}$ denotes the support of $P$. Indeed, if $R < H(X)$, then, by the converse part of the classical source-coding theorem [14, Th. 3.1.1]

$$\lim_{n\to\infty} \Pr\big(|f_n^{-1}(f_n(X^n))| \geq 2\big) = 1, \tag{15}$$

which implies that the $\rho$-th moment of $|f_n^{-1}(f_n(X^n))|$ cannot tend to one as $n$ tends to infinity. And if $R > \log|\mathrm{supp}(P)|$, then every $n$-tuple of tasks that occurs with positive probability can be given a distinct description so for every $n$

$$\min_{f_n\colon \mathcal{X}^n \to \{1,\dots,2^{nR}\}} \mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] = 1. \tag{16}$$

The limit

$$\lim_{\rho\to\infty} H_{\tilde{\rho}}(X) = \log|\mathrm{supp}(P)| \tag{17}$$

follows from our operational characterization of Rényi entropy as follows. If $R > \log|\mathrm{supp}(P)|$, then, by the pigeonhole-principle, for any choice of $f_n\colon \mathcal{X}^n \to \{1, \dots, 2^{nR}\}$ there must exist some $x_0^n \in \mathrm{supp}(P^n)$ for which

$$|f_n^{-1}(f_n(x_0^n))| \geq 2^{n(\log|\mathrm{supp}(P)| - R)}. \tag{18}$$

Since $P^n(x_0^n) \geq p_{\min}^n$, where $p_{\min}$ denotes the smallest nonzero probability of any source symbol, we have

$$\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^{\rho}\big] \geq P^n(x_0^n)|f_n^{-1}(f_n(x_0^n))|^{\rho} \tag{19}$$

$$\geq 2^{n\rho(\log|\mathrm{supp}(P)| - R + \rho^{-1}\log p_{\min})}. \tag{20}$$

For all sufficiently large $\rho$ the RHS tends to infinity as $n \to \infty$, which proves that $\lim_{\rho\to\infty} H_{\tilde{\rho}}(X) \geq \log|\mathrm{supp}(P)|$; the reverse inequality follows from (14).

As to the limit when $\rho$ approaches zero, note that if $R > H(X)$, then the probability that the cardinality of $f_n^{-1}(f_n(X^n))$ exceeds one can be driven to zero exponentially fast in $n$ [15, Th. 2.15], say as $e^{-n\zeta}$ for some $\zeta > 0$ and sufficiently large $n$. And since $|f_n^{-1}(f_n(X^n))|$ is trivially upper-bounded by $2^{n\log|\mathcal{X}|}$, the $\rho$-th moment of $|f_n^{-1}(f_n(X^n))|$ will tend to one if $\rho < \zeta/\log|\mathcal{X}|$. Thus, $\lim_{\rho\to 0} H_{\tilde{\rho}}(X) \leq H(X)$ and, in view of (14),

$$\lim_{\rho\to 0} H_{\tilde{\rho}}(X) = H(X). \tag{21}$$

The rest of this paper is organized as follows. In Section II we introduce some notation. In Section III we prove Theorem I.1. In Section IV we consider a mismatched version of the direct part of Theorem I.1 (i.e., the upper bound), where

[1] Throughout $2^{nR}$ stands for $\lfloor 2^{nR} \rfloor$.

$f$ is designed based on the law $Q$ instead of $P$. We show that the penalty incurred by this mismatch can be expressed in terms of a divergence measure between $P$ and $Q$ that was proposed by Sundaresan [16]. In Section V we state and prove a universal version of the direct part of Theorem I.2 for IID sources. In Section VI we generalize Theorems I.1 and I.2 to account for the presence of side-information, where the key quantity is a conditional version of Rényi entropy. We also generalize the result from Section V. In Section VII we study a rate-distortion version of the problem for IID sources, where the key quantity is "Rényi's analog to the rate-distortion function" introduced by Arikan and Merhav [17]. In Section VIII we study the problem of encoding IID tasks when different tasks may have different costs.

## II. Notation and Preliminaries

We denote by $\mathbb{N}$ the set of positive integers. The cardinality of a finite set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. We use the notation $x^n = (x_1, \ldots, x_n)$ for $n$-tuples. If $P$ is a PMF on $\mathcal{X}$, then $P^n$ denotes the product PMF on $\mathcal{X}^n$

$$P^n(x^n) = \prod_{i=1}^{n} P(x_i), \quad x^n \in \mathcal{X}^n. \tag{22}$$

The support of $P$ is denoted by $\operatorname{supp}(P)$, so

$$\operatorname{supp}(P) = \{x \in \mathcal{X} : P(x) > 0\}. \tag{23}$$

If $\mathcal{A} \subseteq \mathcal{X}$, then we write $P(\mathcal{A})$ in lieu of $\sum_{x \in \mathcal{A}} P(x)$. If $W(\cdot|x)$ is a PMF on a finite set $\mathcal{Y}$ for every $x \in \mathcal{X}$ (i.e., a channel from $\mathcal{X}$ to $\mathcal{Y}$), then $P \circ W$ denotes the induced joint PMF on $\mathcal{X} \times \mathcal{Y}$

$$(P \circ W)(x, y) = P(x)W(y|x), \quad (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{24}$$

and $PW$ denotes the induced marginal PMF on $\mathcal{Y}$

$$(PW)(y) = \sum_{x \in \mathcal{X}} P(x)W(y|x), \quad y \in \mathcal{Y}. \tag{25}$$

The collection of all PMFs on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The collection of all channels from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\mathcal{P}(\mathcal{Y}|\mathcal{X})$.

For information-theoretic quantities (entropy, relative entropy, mutual information, etc.) we adopt the notation in [15]. We need basic results from the Method of Types as presented in [15, Ch. 2]. The set of types of sequences in $\mathcal{X}^n$ (i.e., the set of rational PMFs with denominator $n$) is denoted by $\mathcal{P}_n(\mathcal{X})$. The set of all $x^n \in \mathcal{X}^n$ of type $Q \in \mathcal{P}_n(\mathcal{X})$ (i.e., the type class of $Q$) is denoted by $T_Q^{(n)}$ or by $T_Q$ if $n$ is clear from the context. The $V$-shell of a sequence $x^n \in \mathcal{X}^n$ is denoted by $T_V(x^n)$.

The ceiling of a real number $\xi$ (i.e., the smallest integer no smaller than $\xi$) is denoted by $\lceil \xi \rceil$. We frequently use the inequality

$$\lceil \xi \rceil^\rho < 1 + 2^\rho \xi^\rho, \quad \xi \geq 0, \tag{26}$$

which is easily checked by considering separately the cases $0 \leq \xi \leq 1$ and $\xi > 1$. As mentioned in the introduction, $\log(\cdot)$ denotes the base-2 logarithm, and $\log_\alpha(\cdot)$ denotes the base-$\alpha$ logarithm for general $\alpha > 1$.

## III. Proof of Theorem I.1

### A. The Lower Bound (Converse)

The proof of the lower bound (6) in Theorem I.1 is inspired by the proof of [10, Th. 1]. It hinges on the following simple observation.

*Proposition III.1: If $\mathcal{L}_1, \ldots, \mathcal{L}_M$ is a partition of a finite set $\mathcal{X}$ into $M$ nonempty subsets, i.e.,*

$$\bigcup_{m=1}^{M} \mathcal{L}_m = \mathcal{X} \quad \text{and} \quad (\mathcal{L}_m \cap \mathcal{L}_{m'} = \emptyset \text{ iff } m' \neq m), \tag{27}$$

*and $L(x)$ is the cardinality of the subset containing $x$, i.e., $L(x) = |\mathcal{L}_m|$ if $x \in \mathcal{L}_m$, then*

$$\sum_{x \in \mathcal{X}} \frac{1}{L(x)} = M. \tag{28}$$

*Proof:*

$$\sum_{x \in \mathcal{X}} \frac{1}{L(x)} = \sum_{m=1}^{M} \sum_{x \in \mathcal{L}_m} \frac{1}{L(x)} \tag{29}$$

$$= \sum_{m=1}^{M} \sum_{x \in \mathcal{L}_m} \frac{1}{|\mathcal{L}_m|} \tag{30}$$

$$= M. \tag{31}$$

□

To prove the lower bound in Theorem I.1, fix any $f : \mathcal{X} \to \{1, \ldots, M\}$, and let $N$ denote the number of nonempty subsets in the partition $f^{-1}(1), \ldots, f^{-1}(M)$. Note that for this partition the cardinality of the subset containing $x$ is

$$L(x) = |f^{-1}(f(x))|, \quad x \in \mathcal{X}. \tag{32}$$

Recall Hölder's Inequality: If $a$ and $b$ are functions from $\mathcal{X}$ into the nonnegative reals, and $p$ and $q$ are real numbers larger than one satisfying $1/p + 1/q = 1$, then

$$\sum_{x \in \mathcal{X}} a(x)b(x) \leq \left(\sum_{x \in \mathcal{X}} a(x)^p\right)^{1/p} \left(\sum_{x \in \mathcal{X}} b(x)^q\right)^{1/q}. \tag{33}$$

Rearranging (33) gives

$$\sum_{x \in \mathcal{X}} a(x)^p \geq \left(\sum_{x \in \mathcal{X}} b(x)^q\right)^{-p/q} \left(\sum_{x \in \mathcal{X}} a(x)b(x)\right)^p. \tag{34}$$

Substituting $p = 1 + \rho$, $q = (1 + \rho)/\rho$, $a(x) = P(x)^{\frac{1}{1+\rho}} |f^{-1}(f(x))|^{\frac{\rho}{1+\rho}}$ and $b(x) = |f^{-1}(f(x))|^{-\frac{\rho}{1+\rho}}$ in (34), we obtain

$$\sum_{x \in \mathcal{X}} P(x)|f^{-1}(f(x))|^\rho$$

$$\geq \left(\sum_{x \in \mathcal{X}} \frac{1}{|f^{-1}(f(x))|}\right)^{-\rho} \left(\sum_{x \in \mathcal{X}} P(x)^{\frac{1}{1+\rho}}\right)^{1+\rho} \tag{35}$$

$$= 2^{\rho(H_{\tilde{\rho}}(X) - \log N)} \tag{36}$$

$$\geq 2^{\rho(H_{\tilde{\rho}}(X) - \log M)}, \tag{37}$$

where (36) follows from (4), (32), and Proposition III.1; and where (37) follows because $N \leq M$. □

### *B. The Upper Bound (Direct Part)*

The key to the upper bound in Theorem I.1 is the following reversed version of Proposition III.1; a proof is provided in Appendix A.

*Proposition III.2: If $\mathcal{X}$ is a finite set, $\lambda\colon \mathcal{X} \to \mathbb{N} \cup \{+\infty\}$ and*

$$\sum_{x\in\mathcal{X}} \frac{1}{\lambda(x)} = \mu \tag{38}$$

*(with the convention $1/\infty = 0$), then there exists a partition of $\mathcal{X}$ into at most*

$$\min_{\alpha>1} \lfloor \alpha\mu + \log_\alpha|\mathcal{X}| + 2 \rfloor \tag{39}$$

*subsets such that*

$$L(x) \le \min\{\lambda(x), |\mathcal{X}|\}, \quad \textit{for all } x \in \mathcal{X}, \tag{40}$$

*where $L(x)$ is the cardinality of the subset containing $x$.*

(Proposition III.1 cannot be fully reversed in the sense that (39) cannot be replaced with $\mu$. Indeed, consider $\mathcal{X} = \{a, b, c, d\}$ with $\lambda(a) = 1$, $\lambda(b) = 2$, and $\lambda(c) = \lambda(d) = 4$. In this example, $\mu$ equals 2, but we need 3 subsets to satisfy the cardinality constraints.)

Since Hölder's Inequality (33) holds with equality if, and only if, $a(x)^p$ is proportional to $b(x)^q$, it follows from the proof that the lower bound in Theorem I.1 holds with equality if, and only if, $|f^{-1}(f(x))|$ is proportional to $P(x)^{-1/(1+\rho)}$ and $f$ is surjective. We derive (7) by constructing a partition that approximately satisfies this relationship. To this end, we use Proposition III.2 with $\alpha = 2$ in (39) and

$$\lambda(x) = \begin{cases} \lceil \beta\, P(x)^{-\frac{1}{1+\rho}} \rceil & \text{if } P(x) > 0, \\ +\infty & \text{if } P(x) = 0, \end{cases} \tag{41}$$

where we choose $\beta$ just large enough to guarantee the existence of a partition of $\mathcal{X}$ into at most $M$ subsets satisfying (40). For $M > \log|\mathcal{X}| + 2$ this is accomplished by the choice

$$\beta = \frac{2\sum_{x\in\mathcal{X}} P(x)^{\frac{1}{1+\rho}}}{M - \log|\mathcal{X}| - 2}. \tag{42}$$

Indeed, with this choice

$$\begin{aligned} \mu &= \sum_{x\in\mathcal{X}} \frac{1}{\lambda(x)} && (43)\\ &\le \sum_{x\in\mathcal{X}} \frac{P(x)^{\frac{1}{1+\rho}}}{\beta} && (44)\\ &= \frac{M - \log|\mathcal{X}| - 2}{2}, && (45) \end{aligned}$$

and hence

$$2\mu + \log|\mathcal{X}| + 2 \le M. \tag{46}$$

Let then the partition $\mathcal{L}_1, \dots, \mathcal{L}_N$ with $N \le M$ be as promised by Proposition III.2. Construct an encoder $f\colon \mathcal{X} \to \{1, \dots, M\}$ by setting $f(x) = m$ if $x \in \mathcal{L}_m$. For this encoder,

$$\begin{aligned} &\sum_{x\in\mathcal{X}} P(x)|f^{-1}(f(x))|^\rho \\ &\quad= \sum_{x:P(x)>0} P(x)L(x)^\rho && (47)\\ &\quad\le \sum_{x:P(x)>0} P(x)\lambda(x)^\rho && (48)\\ &\quad= \sum_{x:P(x)>0} P(x)\lceil \beta\, P(x)^{-\frac{1}{1+\rho}} \rceil^\rho && (49)\\ &\quad< 1 + (2\beta)^\rho \sum_{x:P(x)>0} P(x)^{\frac{1}{1+\rho}} && (50)\\ &\quad= 1 + 2^{\rho(H_{\tilde{\rho}}(X) - \log \widetilde{M})}, && (51) \end{aligned}$$

where (50) follows from (26), and where $\widetilde{M}$ is as in Theorem I.1. □

## IV. Mismatch

The key to the upper bound in Theorem I.1 was to use Proposition III.2 with $\lambda$ as in (41)–(42) to obtain a partition of $\mathcal{X}$ for which the cardinality of the subset containing $x$ is approximately proportional to $P(x)^{-1/(1+\rho)}$. Evidently, this construction requires knowledge of the distribution $P$ of $X$. (But see Section V for a universal version of the direct part of Theorem I.2 for IID sources that does not require knowledge of the source's distribution.)

In this section, we study the penalty when $P$ is replaced with $Q$ in (41) and (42). Since it is then still true that

$$\mu \le \frac{M - \log|\mathcal{X}| - 2}{2}, \tag{52}$$

Proposition III.2 guarantees the existence of a partition of $\mathcal{X}$ into at most $M$ subsets satisfying (40). Constructing an encoder $f$ from this partition as in Section III-B and following steps similar to (47)–(51) yields

$$\sum_{x\in\mathcal{X}} P(x)|f^{-1}(f(x))|^\rho < 1 + 2^{\rho(H_{\tilde{\rho}}(X) + \Delta_{\tilde{\rho}}(P||Q) - \log \widetilde{M})}, \tag{53}$$

where $\widetilde{M}$ is as in Theorem I.1, and where

$$\begin{aligned} &\Delta_\alpha(P||Q) \\ &\quad\triangleq \log \frac{\sum_{x\in\mathcal{X}} Q(x)^\alpha}{\left(\sum_{x\in\mathcal{X}} P(x)^\alpha\right)^{\frac{1}{1-\alpha}}} \left( \sum_{x\in\mathcal{X}} \frac{P(x)}{Q(x)^{1-\alpha}} \right)^{\frac{\alpha}{1-\alpha}}. && (54) \end{aligned}$$

The parameter $\alpha$ can be any positive number not equal to one. We use the convention $0/0 = 0$ and $a/0 = +\infty$ if $a > 0$. Thus, $\Delta_{\tilde{\rho}}(P||Q) < \infty$ only if the support of $P$ is contained in the support of $Q$.

The penalty in the exponent on the RHS of (53) when compared to the upper bound in Theorem I.1 is thus given by $\Delta_{\tilde{\rho}}(P||Q)$. To reinforce this, note further that

$$\Delta_\alpha(P^n||Q^n) = n\Delta_\alpha(P||Q). \tag{55}$$

Consequently, if the source $\{X_i\}_{i=1}^\infty$ is IID $P$ and we construct $f_n\colon \mathcal{X}^n \to \{1, \dots, 2^{nR}\}$ based on $Q^n$ instead of $P^n$, we obtain the bound

$$\begin{aligned}&\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^\rho\big]\\&\quad< 1 + 2^{n\rho(H_{\tilde\rho}(X_1)+\Delta_{\tilde\rho}(P||Q)-R+\delta_n)},\end{aligned} \tag{56}$$

where $\delta_n \to 0$ as $n \to \infty$. The RHS of (56) tends to one provided that $R > H_{\tilde\rho}(X_1) + \Delta_{\tilde\rho}(P||Q)$. Thus, in the IID case $\Delta_{\tilde\rho}(P||Q)$ is the rate penalty incurred by the mismatch between $P$ and $Q$.

The family of divergence measures $\Delta_\alpha(P||Q)$ was first identified by Sundaresan [16] who showed that it plays a similar role in the Massey-Arikan guessing problem [10], [11]. We conclude this section with some properties of $\Delta_\alpha(P||Q)$. Properties 1–3 (see below) were given in [16]; we repeat them here for completeness. Note that Rényi's divergence (see [9])

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log\sum_{x\in\mathcal{X}} P(x)^\alpha Q(x)^{1-\alpha}, \tag{57}$$

satisfies Properties 1 and 3 but none of the others in general.

*Proposition IV.1: The functional $\Delta_\alpha(P||Q)$ has the following properties.*

1) $\Delta_\alpha(P||Q) \geq 0$ *with equality if, and only if,* $P = Q$.
2) $\Delta_\alpha(P||Q) = \infty$ *if, and only if,* $\mathrm{supp}(P) \not\subseteq \mathrm{supp}(Q)$ *or* ($\alpha > 1$ *and* $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.)
3) $\lim_{\alpha\to 1}\Delta_\alpha(P||Q) = D(P||Q)$.
4) $\lim_{\alpha\to 0}\Delta_\alpha(P||Q) = \log\frac{|\mathrm{supp}(Q)|}{|\mathrm{supp}(P)|}$ *if* $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$.
5) $\lim_{\alpha\to\infty}\Delta_\alpha(P||Q) = \log\frac{\max_{x\in\mathcal{X}}P(x)}{\frac{1}{|\mathcal{Q}|}\sum_{x\in\mathcal{Q}}P(x)}$, *where* $\mathcal{Q} = \big\{x\in\mathcal{X} : Q(x) = \max_{x'\in\mathcal{X}} Q(x')\big\}$.

*Proof:* Property 2 follows by inspection of (54). Properties 3–5 follow by simple calculus. As to Property 1, consider first the case where $0 < \alpha < 1$. In view of Property 2, we may assume that $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$. Hölder's Inequality (33) with $p = 1/\alpha$ and $q = 1/(1-\alpha)$ gives

$$\begin{aligned}&\sum_{x\in\mathcal{X}} P(x)^\alpha\\&\quad= \sum_{x\in\mathrm{supp}(P)}\frac{P(x)^\alpha}{Q(x)^{\alpha(1-\alpha)}}Q(x)^{\alpha(1-\alpha)} &(58)\\&\quad\leq \bigg(\sum_{x\in\mathrm{supp}(P)}\frac{P(x)}{Q(x)^{1-\alpha}}\bigg)^\alpha\bigg(\sum_{x\in\mathrm{supp}(P)}Q(x)^\alpha\bigg)^{1-\alpha} &(59)\\&\quad\leq \bigg(\sum_{x\in\mathcal{X}}\frac{P(x)}{Q(x)^{1-\alpha}}\bigg)^\alpha\bigg(\sum_{x\in\mathcal{X}}Q(x)^\alpha\bigg)^{1-\alpha}. &(60)\end{aligned}$$

Dividing by $\sum_x P(x)^\alpha$ and taking $(1-\alpha)$-th roots shows that $\Delta_\alpha(P||Q) \geq 0$. The condition for equality in Hölder's Inequality implies that equality holds if, and only if, $P = Q$. Consider next the case where $\alpha > 1$. We may assume $\mathrm{supp}(P)\cap\mathrm{supp}(Q) \neq \emptyset$ (Property 2). Hölder's Inequality with $p = \alpha$ and $q = \alpha/(\alpha-1)$ gives

$$\begin{aligned}\sum_{x\in\mathcal{X}}\frac{P(x)}{Q(x)^{1-\alpha}} &= \sum_{x\in\mathcal{X}}P(x)Q(x)^{\alpha-1} &(61)\\&\leq \bigg(\sum_{x\in\mathcal{X}}P(x)^\alpha\bigg)^{\frac{1}{\alpha}}\bigg(\sum_{x\in\mathcal{X}}Q(x)^\alpha\bigg)^{\frac{\alpha-1}{\alpha}}. &(62)\end{aligned}$$

Dividing by $\sum_x P(x)/Q(x)^{1-\alpha}$ and raising to the power of $\alpha/(\alpha-1)$ shows that $\Delta_\alpha(P||Q) \geq 0$. Equality holds if, and only if, $P = Q$. □

## V. Universal Encoders for IID Sources

In Section I the direct part of Theorem I.2 is proved using the upper bound in Theorem I.1. It is pointed out in Section IV that the construction of the encoder in the proof of this upper bound requires knowledge of the distribution of $X$. As the next result shows, however, for IID sources we do not need to know the distribution of the source to construct good encoders.

*Theorem V.1: Let $\mathcal{X}$ be a finite set. For every rate $R > 0$ there exist encoders $f_n : \mathcal{X} \to \{1,\ldots,2^{nR}\}$ such that for every IID source $\{X_i\}_{i=1}^\infty$ with alphabet $\mathcal{X}$ and every $\rho > 0$,*

$$\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^\rho\big] < 1 + 2^{-n\rho(R-H_{\tilde\rho}(X_1)-\delta_n)}, \tag{63}$$

*where*

$$\delta_n = \frac{1+(1+\rho^{-1})|\mathcal{X}|\log(n+1)}{n}. \tag{64}$$

*In particular,*

$$\lim_{n\to\infty}\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^\rho\big] = 1, \tag{65}$$

*whenever $H_{\tilde\rho}(X_1) < R$.*

*Proof:* We first partition $\mathcal{X}^n$ into the different type classes $T_Q$, of which there are less than $(n+1)^{|\mathcal{X}|}$. We then partition each $T_Q$ into $2^{n(R-\delta'_n)}$ subsets of cardinality at most $\lceil |T_Q| 2^{-n(R-\delta'_n)}\rceil$ where $\delta'_n = n^{-1}|\mathcal{X}|\log(n+1)$. Since $|T_Q| \leq 2^{nH(Q)}$, each $x^n \in T_Q$ thus ends up in a subset of cardinality at most

$$\big\lceil 2^{n(H(Q)-R+\delta'_n)}\big\rceil. \tag{66}$$

Note that the total number of subsets in the partition does not exceed $2^{nR}$. We construct $f_n : \mathcal{X} \to \{1,\ldots,2^{nR}\}$ by enumerating the subsets in the partition with the numbers in $\{1,\ldots,2^{nR}\}$ and by mapping to $m \in \{1,\ldots,2^{nR}\}$ the $x^n$'s that comprise the $m$-th subset. Suppose now that $\{X_i\}_{i=1}^\infty$ is IID $P$ with alphabet $\mathcal{X}$ and observe that

$$\begin{aligned}&\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|^\rho\big]\\&\quad= \sum_{x^n\in\mathcal{X}^n}P^n(x^n)|f_n^{-1}(f_n(x^n))|^\rho &(67)\\&\quad\leq \sum_{Q\in\mathcal{P}_n(\mathcal{X})}\sum_{x^n\in T_Q}P^n(x^n)\big\lceil 2^{n(H(Q)-R+\delta'_n)}\big\rceil^\rho &(68)\\&\quad< 1 + 2^\rho\sum_{Q\in\mathcal{P}_n(\mathcal{X})}2^{n\rho(H(Q)-R+\delta'_n)}\sum_{x^n\in T_Q}P^n(x^n) &(69)\\&\quad\leq 1 + 2^\rho\sum_{Q\in\mathcal{P}_n(\mathcal{X})}2^{-n\rho(R-H(Q)+\rho^{-1}D(Q||P)-\delta'_n)} &(70)\\&\quad\leq 1 + 2^{-n\rho(R-H_{\tilde\rho}(X_1)-\delta_n)}. &(71)\end{aligned}$$

Here (68) follows from the construction of $f_n$; (69) follows from (26); (70) follows because the probability of the source emitting a sequence of type $Q$ is at most $2^{-nD(Q||P)}$; and (71) follows from the identity (see [10])

$$H_{\tilde\rho}(X_1) = \max_{Q\in\mathcal{P}(\mathcal{X})} H(Q) - \rho^{-1}D(Q||P), \tag{72}$$

and the fact that $|\mathcal{P}_n(\mathcal{X})| < (n+1)^{|\mathcal{X}|}$. □

## VI. Tasks With Side-Information

In this section we generalize Theorems I.1, I.2, and V.1 to account for side-information: A task $X$ and side-information $Y$ are drawn according to a joint PMF $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$, where both $\mathcal{X}$ and $\mathcal{Y}$ are finite, and where the side-information is available to both the task describer (encoder) and the tasks performer. The encoder is now of the form

$$f\colon \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, M\}. \tag{73}$$

If the realization of $(X, Y)$ is $(x, y)$ and $f(x, y) = m$, then all the tasks in the set

$$f^{-1}(m, y) \triangleq \{x' \in \mathcal{X} : f(x', y) = m\} \tag{74}$$

are performed. As in Section I, we seek to minimize for a given $M$ the $\rho$-th moment of the number of performed tasks

$$\begin{aligned} &\mathrm{E}\big[|f^{-1}(f(X, Y), Y)|^{\rho}\big] \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x, y)|f^{-1}(f(x, y), y)|^{\rho}. \end{aligned} \tag{75}$$

The key quantity here is a conditional version of Rényi entropy (proposed by Arimoto [18]):

$$H_{\tilde{\rho}}(X|Y) = \frac{1}{\rho} \log \sum_{y\in\mathcal{Y}} \left( \sum_{x\in\mathcal{X}} P_{X,Y}(x, y)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \tag{76}$$

Theorem I.1 can be generalized as follows.

*Theorem VI.1: Let $(X, Y)$ be a pair of chance variables taking value in the finite set $\mathcal{X} \times \mathcal{Y}$, and let $\rho > 0$.*

1) *For all positive integers $M$ and every $f\colon \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, M\}$,*

$$\mathrm{E}\big[|f^{-1}(f(X, Y), Y)|^{\rho}\big] \geq 2^{\rho(H_{\tilde{\rho}}(X|Y)-\log M)}. \tag{77}$$

2) *For all integers $M > \log|\mathcal{X}|+2$ there exists $f\colon \mathcal{X}\times\mathcal{Y} \to \{1, \ldots, M\}$ such that*

$$\mathrm{E}\big[|f^{-1}(f(X, Y), Y)|^{\rho}\big] < 1 + 2^{\rho(H_{\tilde{\rho}}(X|Y)-\log \widetilde{M})}, \tag{78}$$

*where $\widetilde{M} = (M - \log|\mathcal{X}| - 2)/4$.*

As a corollary we obtain a generalization of Theorem I.2.

*Theorem VI.2: Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be any source with finite alphabet $\mathcal{X} \times \mathcal{Y}$, and let $\rho > 0$.*

1) *If $R > \limsup_{n\to\infty} H_{\tilde{\rho}}(X^n|Y^n)/n$, then there exist $f_n\colon \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ such that*

$$\lim_{n\to\infty} \mathrm{E}\big[|f_n^{-1}(f_n(X^n, Y^n), Y^n)|^{\rho}\big] = 1. \tag{79}$$

2) *If $R < \liminf_{n\to\infty} H_{\tilde{\rho}}(X^n|Y^n)/n$, then for any choice of $f_n\colon \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$*

$$\lim_{n\to\infty} \mathrm{E}\big[|f_n^{-1}(f_n(X^n, Y^n), Y^n)|^{\rho}\big] = \infty. \tag{80}$$

To prove (77) fix $M$ and $f\colon \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, M\}$. Note that for every $y \in \mathcal{Y}$ the sets $f^{-1}(1, y), \ldots, f^{-1}(M, y)$ form a partition of $\mathcal{X}$, and the cardinality of the subset containing $x$ is $|f^{-1}(f(x, y), y)|$. Following steps similar to (35)–(37), we obtain

$$\begin{aligned} &\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)|f^{-1}(f(x, y), y)|^{\rho} \\ &\geq 2^{-\rho \log M} \left( \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}} \right)^{1+\rho}, \quad y \in \mathcal{Y}. \end{aligned} \tag{81}$$

Multiplying both sides by $P_Y(y)$ and summing over all $y \in \mathcal{Y}$ establishes (77).

To prove (78) fix some $y \in \mathcal{Y}$ and replace $P(x)$ with $P_{X|Y}(x|y)$ everywhere in the proof of the upper bound in Theorem I.1 (see Section III-B) to obtain an encoder $f_y\colon \mathcal{X} \to \{1, \ldots, M\}$ satisfying

$$\begin{aligned} &\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)|f_y^{-1}(f_y(x))|^{\rho} \\ &< 1 + 2^{-\rho \log \widetilde{M}} \left( \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \end{aligned} \tag{82}$$

Setting $f(x, y) = f_y(x)$, multiplying both sides of (82) by $P_Y(y)$, and summing over all $y \in \mathcal{Y}$ establishes (78). □

One may also generalize Theorem V.1:

*Theorem VI.3: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and let $\rho > 0$. For every rate $R > 0$ there exist encoders $f_n\colon \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, 2^{nR}\}$ such that for every IID source $\{(X_i, Y_i)\}_{i=1}^{\infty}$ with alphabet $\mathcal{X} \times \mathcal{Y}$,*

$$\begin{aligned} &\mathrm{E}\big[|f_n^{-1}(f_n(X^n, Y^n), Y^n)|^{\rho}\big] \\ &< 1 + 2^{-n\rho(R-H_{\tilde{\rho}}(X_1|Y_1)-\delta_n)}, \end{aligned} \tag{83}$$

*where*

$$\delta_n = \frac{1 + (1 + \rho^{-1})|\mathcal{X}||\mathcal{Y}| \log(n + 1) + \rho^{-1}|\mathcal{Y}| \log(n + 1)}{n}. \tag{84}$$

*In particular,*

$$\lim_{n\to\infty} \mathrm{E}\big[|f_n^{-1}(f_n(X^n, Y^n), Y^n)|^{\rho}\big] = 1, \tag{85}$$

*whenever $H_{\tilde{\rho}}(X_1|Y_1) < R$.*

*Proof:* We fix an arbitrary $y^n \in \mathcal{Y}^n$ and partition $\mathcal{X}^n$ into the different $V$-shells $T_V(y^n)$ (see [15, Ch. 2]) of which there are less than $(n + 1)^{|\mathcal{X}||\mathcal{Y}|}$. We then partition each $V$-shell into $2^{n(R-\delta'_n)}$ subsets of cardinality at most $\lceil|T_V(y^n)|2^{-n(R-\delta'_n)}\rceil$ where $\delta'_n = n^{-1}|\mathcal{X}||\mathcal{Y}| \log(n + 1)$. Since $|T_V(y^n)| \leq 2^{nH(V|P_{y^n})}$, where $P_{y^n}$ denotes the type of $y^n$, each $x^n \in T_V(y^n)$ will end up in a subset of cardinality at most

$$\lceil 2^{n(H(V|P_{y^n})-R+\delta'_n)} \rceil. \tag{86}$$

From this partition we construct $f_n(\cdot, y^n)\colon \mathcal{X} \to \{1, \ldots, 2^{nR}\}$ by enumerating the subsets with the numbers 1 through $2^{nR}$ and by mapping to each $m \in \{1, \ldots, 2^{nR}\}$ the $x^n$'s that comprise the $m$-th subset. Carrying out this construction for every $y^n \in \mathcal{Y}^n$ yields an encoder $f_n\colon \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, 2^{nR}\}$. Suppose now that $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is IID $P_{X,Y}$ with alphabet $\mathcal{X} \times \mathcal{Y}$ and observe that for every $y^n \in \mathcal{Y}^n$ with $P_Y^{(n)}(y^n) > 0$,

$$\begin{aligned} &\sum_{x^n\in\mathcal{X}^n} P_{X|Y}^{(n)}(x^n|y^n)\big|f_n^{-1}\big(f_n(x^n, y^n), y^n\big)\big|^{\rho} \\ &\leq \sum_{V:T_V(y^n)\neq\emptyset} \sum_{x^n\in T_V(y^n)} P_{X|Y}^{(n)}(x^n|y^n)\lceil 2^{n(H(V|P_{y^n})-R+\delta'_n)}\rceil^{\rho} \end{aligned} \tag{87}$$

$$< 1 + 2^{\rho} \sum_{V:T_V(y^n)\neq\emptyset} 2^{n\rho(H(V|P_{y^n})-R+\delta'_n)} \sum_{x^n\in T_V(y^n)} P_{X|Y}^{(n)}(x^n|y^n) \tag{88}$$

$$< 1 + 2^{\rho} \sum_{V: T_V(y^n) \neq \emptyset} 2^{-nD(V||P_{X|Y}|P_{y^n})} 2^{n\rho(H(V|P_{y^n}) - R + \delta'_n)}. \tag{89}$$

Here (87) follows from the construction of $f_n$; (88) follows from (26); (89) follows because conditional on $Y^n = y^n$ the probability that $X^n$ is in the $V$-shell of $y^n$ is at most $2^{-nD(V||P_{X|Y}|P_{y^n})}$. Noting that whether $T_V(y^n)$ is nonempty depends on $y^n$ only via its type, it follows that the sum in (89) depends on $y^n$ only via $P_{y^n}$. Noting further that the probability that $Y^n$ is of type $Q \in \mathcal{P}_n(\mathcal{Y})$ is at most $2^{-nD(Q||P_Y)}$ it follows from (87)–(89) upon taking expectation with respect to $Y^n$ that

$$\mathrm{E}\big[|f_n^{-1}(f_n(X^n, Y^n), Y^n)|^{\rho}\big] < 1 + 2^{\rho} \sum_{Q \in \mathcal{P}_n(\mathcal{Y})} 2^{-nD(Q||P_Y)} \times \sum_{V} 2^{-nD(V||P_{X|Y}|Q)} 2^{n\rho(H(V|Q) - R + \delta'_n)}, \tag{90}$$

where for a given $Q \in \mathcal{P}_n(\mathcal{Y})$ the inner sum extends over all $V$ such that $T_V(y^n)$ is not empty for some (and hence all) $y^n$ of type $Q$. In Appendix B it is shown that

$$H_{\tilde{\rho}}(X_1|Y_1) = \max_{\substack{Q \in \mathcal{P}(\mathcal{Y}) \\ V \in \mathcal{P}(\mathcal{X}|\mathcal{Y})}} H(V|Q) - \rho^{-1} D(Q \circ V || P_{X,Y}). \tag{91}$$

Using (91), the identity

$$D(Q \circ V || P_{X,Y}) = D(Q||P_Y) + D(V||P_{X|Y}|Q), \tag{92}$$

and the fact that the number of types of sequences in $\mathcal{Y}^n$ is less than $(n+1)^{|\mathcal{Y}|}$ and the number of conditional types $V$ is less than $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$, it follows that the RHS of (90) is upper-bounded by the RHS of (83). □

## VII. Coding for Tasks With a Fidelity Criterion

In this section we study a rate-distortion version of the problem described in Section I. We only treat IID sources and single-letter distortion functions. Suppose that the source $\{X_i\}_{i=1}^{\infty}$ generates tasks from a finite set of tasks $\mathcal{X}$ IID according to $P$. Let $\hat{\mathcal{X}}$ be some other finite set of tasks, and let $d \colon \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a function that measures the dissimilarity, or distortion, between any pair of tasks in $\mathcal{X} \times \hat{\mathcal{X}}$. The distortion function $d$ extends to $n$-tuples of tasks in the usual way:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i), \quad (x^n, \hat{x}^n) \in \mathcal{X}^n \times \hat{\mathcal{X}}^n. \tag{93}$$

We assume that for every $x \in \mathcal{X}$ there is some $\hat{x} \in \hat{\mathcal{X}}$ for which $d(x, \hat{x}) = 0$, i.e.,

$$\min_{\hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) = 0, \quad x \in \mathcal{X}. \tag{94}$$

We describe the first $n$ tasks $X^n$ using $nR$ bits with an encoder

$$f \colon \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}. \tag{95}$$

Subsequently, the description $f(X^n)$ of $X^n$ is decoded into a subset of $\hat{\mathcal{X}}^n$ by a decoder

$$\varphi \colon \{1, \ldots, 2^{nR}\} \to 2^{\hat{\mathcal{X}}^n}, \tag{96}$$

where $2^{\hat{\mathcal{X}}^n}$ denotes the collection of all subsets of $\hat{\mathcal{X}}^n$. We require that the subset produced by the decoder always contain at least one $n$-tuple of tasks within distortion $\Delta$ of $X^n$, i.e., we require

$$\min_{\hat{x}^n \in \varphi(f(x^n))} d(x^n, \hat{x}^n) \leq \Delta, \quad x^n \in \mathcal{X}^n. \tag{97}$$

Here $\Delta$ is a fixed nonnegative number. All $n$-tuples of tasks in the set $\varphi(f(X^n))$ are performed. The next theorem shows that the infimum of all rates $R$ for which the $\rho$-th moment of the ratio of performed tasks to $n$ can be driven to one as $n$ tends to infinity subject to the constraint (97) is given by

$$R_{\rho}(P, \Delta) \triangleq \max_{Q \in \mathcal{P}(\mathcal{X})} R(Q, \Delta) - \rho^{-1} D(Q||P), \tag{98}$$

where $R(Q, \Delta)$ is the classical rate-distortion function (see [15, Ch. 7]) evaluated at the distortion level $\Delta$ for an IID Q source and distortion function $d$. The function $R_{\rho}(P, \Delta)$ (multiplied by $\rho$) has previously appeared in [17] in the context of guessing.

*Theorem VII.1: Let $\{X_i\}_{i=1}^{\infty}$ be an IID $P$ source with finite alphabet $\mathcal{X}$, and let $\Delta \geq 0$ and $\rho > 0$.*

1) *If $R > R_{\rho}(P, \Delta)$, then there exist $(f_n, \varphi_n)$ as in* (95) *and* (96) *satisfying* (97) *such that*

$$\lim_{n \to \infty} \mathrm{E}\big[|\varphi_n(f_n(X^n))|^{\rho}\big] = 1. \tag{99}$$

2) *If $R < R_{\rho}(P, \Delta)$, then for any $(f_n, \varphi_n)$ as in* (95) *and* (96) *satisfying* (97),

$$\lim_{n \to \infty} \mathrm{E}\big[|\varphi_n(f_n(X^n))|^{\rho}\big] = \infty. \tag{100}$$

It follows immediately from (98) that $R_{\rho}(P, \Delta)$ is nonnegative and nondecreasing in $\rho > 0$. Some other properties are (see [17] for proofs):

1) $R_{\rho}(P, \Delta)$ is nonincreasing, continuous and convex in $\Delta \geq 0$.
2) $R_{\rho}(P, 0) = H_{\tilde{\rho}}(P)$.
3) $\lim_{\rho \to 0} R_{\rho}(P, \Delta) = R(P, \Delta)$.
4) $\lim_{\rho \to \infty} R_{\rho}(P, \Delta) = \max_{Q \in \mathcal{P}(\mathcal{X})} R(Q, \Delta)$.

The fact that $R_{\rho}(P, \Delta)$ is a continuous function of $\Delta$ (Property 1) allows us to strengthen the converse statement in Theorem VII.1 as follows. Suppose that for every positive integer $n$ the encoder/decoder pair $(f_n, \varphi_n)$ is as in (95) and (96) with $R < R_{\rho}(P, \Delta)$ and satisfies (97) for some $\Delta_n$ such that $\limsup_{n \to \infty} \Delta_n \leq \Delta$. Then (100) holds. Indeed, continuity implies that $R < R_{\rho}(P, \Delta + \varepsilon)$ for a sufficiently small $\varepsilon > 0$, and $\limsup_{n \to \infty} \Delta_n \leq \Delta$ implies that $\Delta_n \leq \Delta + \varepsilon$ for all sufficiently large $n$. The claim thus follows from the converse part of Theorem VII.1 with $\Delta$ replaced by $\Delta + \varepsilon$.

Considering Property 2, Theorem I.2 particularized to IID sources can be recovered from Theorem VII.1 by taking $\hat{\mathcal{X}} = \mathcal{X}$ and the Hamming distortion function

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{otherwise.} \end{cases} \tag{101}$$

It was noted in [17] that $R_{\rho}(P, \Delta)$ can be expressed in closed form for binary sources and Hamming distortion:

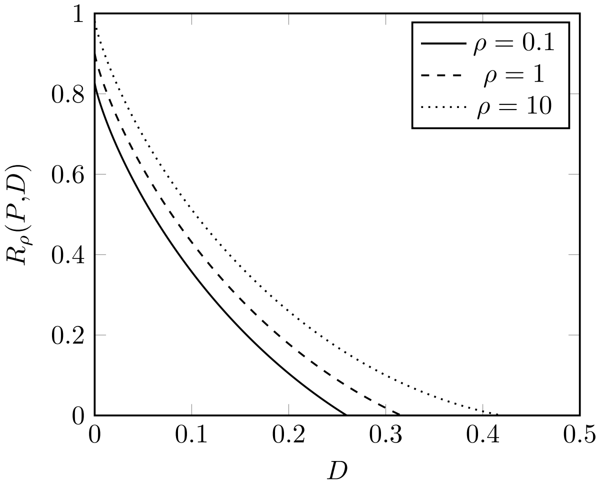

Fig. 1. $R_\rho(P,\Delta)$ in bits for an IID Bernoulli-(1/4) source and Hamming distortion.

*Proposition VII.2: If $\mathcal{X} = \hat{\mathcal{X}} = \{0,1\}$, $d$ is the Hamming distortion function* (101)*, and $P(0) = 1 - P(1) = p$, then*

$$R_\rho(P,\Delta) = \begin{cases} H_{\tilde{\rho}}(p) - h(\Delta) & \text{if } 0 \le \Delta < h^{-1}\big(H_{\tilde{\rho}}(p)\big), \\ 0 & \text{if } \Delta \ge h^{-1}\big(H_{\tilde{\rho}}(p)\big), \end{cases}$$

*where $h^{-1}(\cdot)$ denotes the inverse of the binary entropy function $h(\cdot)$ on the interval* $[0,1/2]$ *and, with slight abuse of notation, $H_{\tilde{\rho}}(p) = H_{\tilde{\rho}}(P)$.*

For a proof of Proposition VII.2 see [17, Thereom 3] and subsequent remarks. A plot of $R_\rho(P,\Delta)$ for $p = 1/4$ and different values of $\rho$ is shown in Figure 1.

We now prove the direct part of Theorem VII.1. Fix $\Delta \ge 0$ and select an arbitrary $\delta > 0$. According to the Type Covering Lemma [15, Lemma 9.1], there is a positive integer $n(\delta)$ such that for all $n \ge n(\delta)$ and every type $Q \in \mathcal{P}_n(\mathcal{X})$ we can find a set $B_Q^{(n)} \subset \hat{\mathcal{X}}^n$ of cardinality at most $2^{n(R(Q,\Delta)+\delta)}$ that covers $T_Q^{(n)}$ in the sense that for every $x^n \in T_Q^{(n)}$ there is at least one $\hat{x}^n \in B_Q^{(n)}$ with $d(x^n,\hat{x}^n) \le \Delta$. We henceforth assume that $n \ge n(\delta)$. For each type $Q \in \mathcal{P}_n(\mathcal{X})$ we partition $B_Q^{(n)}$ into $2^{n(R-\delta_n)}$ subsets of cardinality at most

$$\big\lceil 2^{n(R(Q,\Delta)+\delta-R+\delta_n)}\big\rceil, \tag{102}$$

where $\delta_n = n^{-1}|\mathcal{X}|\log(n+1)$. Since the total number of types is less than $(n+1)^{|\mathcal{X}|}$, we can enumerate all the subsets of all the different $B_Q^{(n)}$'s with the numbers $1,\ldots,2^{nR}$. Let $\varphi_n\colon \{1,\ldots,2^{nR}\} \to 2^{\hat{\mathcal{X}}^n}$ be the mapping that maps the index to the corresponding subset. (If there are less than $2^{nR}$ subsets in our construction, then we map the remaining indices to, say, the empty set.) We then construct $f_n\colon \mathcal{X}^n \to \{1,\ldots,2^{nR}\}$ by mapping each $x^n \in \mathcal{X}^n$ of type $Q$ to an index of a subset of $B_Q^{(n)}$ that contains an $\hat{x}^n$ with $d(x^n,\hat{x}^n) \le \Delta$. Note that the encoder/decoder pair thus constructed satisfies (97), and

$$\begin{aligned} &\mathrm{E}\big[|\varphi_n(f_n(X^n))|^\rho\big] \\ &\quad= \sum_{x^n\in\mathcal{X}^n} P^n(x^n)|\varphi_n(f_n(x^n))|^\rho && (103) \\ &\quad\le \sum_{Q\in\mathcal{P}_n(\mathcal{X})} \sum_{x^n\in T_Q^{(n)}} P^n(x^n)\big\lceil 2^{n(R(Q,\Delta)+\delta-R+\delta_n)}\big\rceil^\rho && (104) \\ &\quad< 1 + 2^\rho \sum_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{n\rho(R(Q,\Delta)+\delta-R+\delta_n)} \sum_{x^n\in T_Q^{(n)}} P^n(x^n) && (105) \\ &\quad< 1 + 2^\rho \sum_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{-n\rho(R+\rho^{-1}D(Q\|P)-R(Q,\Delta)-\delta-\delta_n)} && (106) \\ &\quad\le 1 + 2^{-n\rho(R-R_\rho(P,\Delta)-\delta-\delta_n')}, && (107) \end{aligned}$$

where

$$\delta_n' = \frac{1+(1+\rho^{-1})|\mathcal{X}|\log(n+1)}{n}. \tag{108}$$

Here (104) follows from the construction of $f_n$ and $\varphi_n$; (105) follows from (26); (106) follows because the probability of an IID $P$ source emitting a sequence of type $Q$ is at most $2^{-nD(Q\|P)}$; and (107) follows from the definition of $R_\rho(P,\Delta)$ in (98) and the fact that $|\mathcal{P}_n(\mathcal{X})| < (n+1)^{|\mathcal{X}|}$. The proof of the direct part is completed by noting that if $R > R_\rho(P,\Delta)$, then for sufficiently small $\delta > 0$ the RHS of (107) tends to one as $n$ tends to infinity.

To prove the converse, we fix for each $n \in \mathbb{N}$ an encoder/decoder pair $(f_n,\varphi_n)$ as in (95) and (96) satisfying (97). We may assume that

$$\varphi_n(m) \cap \varphi_n(m') = \emptyset \quad \text{whenever } m \ne m'. \tag{109}$$

Indeed, if $m \ne m'$ and $\hat{x}^n \in \varphi_n(m) \cap \varphi_n(m')$, then we can delete $\hat{x}^n$ from the larger of the two subsets, say $\varphi_n(m)$, and map to $m'$ all the source sequences $x^n$ that were mapped to $m$ by $f_n$ and satisfy $d(x^n,\hat{x}^n) \le \Delta$. This could only reduce the $\rho$-th moment of $|\varphi_n(f_n(X^n))|$ while preserving the property (97).

Define the set

$$\mathcal{Z}_n = \bigcup_{m=1}^{2^{nR}} \varphi_n(m). \tag{110}$$

The assumption (109) implies that the union on the RHS of (110) is disjoint. Consequently, we may define $\mu_n(\hat{x}^n)$ for every $\hat{x}^n \in \mathcal{Z}_n$ as the unique element of $\{1,\ldots,2^{nR}\}$ for which $\hat{x}^n \in \varphi_n(\mu_n(\hat{x}^n))$. Moreover, (97) guarantees the existence of a mapping $g_n\colon \mathcal{X}^n \to \mathcal{Z}_n$ (not necessarily unique) such that, for all $x^n \in \mathcal{X}^n$,

$$g_n(x^n) \in \varphi_n\big(f_n(x^n)\big) \quad \text{and} \quad d\big(x^n, g_n(x^n)\big) \le \Delta. \tag{111}$$

We also define the PMF on $\mathcal{Z}_n$,

$$\tilde{P}_n(\hat{x}^n) = P^n\big(g_n^{-1}(\hat{x}^n)\big), \quad \hat{x}^n \in \mathcal{Z}_n, \tag{112}$$

where

$$g_n^{-1}(\hat{x}^n) = \{x^n \in \mathcal{X}^n : g_n(x^n) = \hat{x}^n\}. \tag{113}$$

With these definitions of $\mu_n$, $g_n$, and $\tilde{P}_n$, we have

$$\begin{aligned}
&\mathrm{E}\big[|\varphi_n(f_n(X^n))|^\rho\big] \\
&= \sum_{x^n\in\mathcal{X}^n} P^n(x^n)\big|\varphi_n\big(f_n(x^n)\big)\big|^\rho && (114)\\
&= \sum_{\hat{x}^n\in\mathcal{Z}_n} P^n\big(g_n^{-1}(\hat{x}^n)\big)\big|\varphi_n\big(\mu_n(\hat{x}^n)\big)\big|^\rho && (115)\\
&= \sum_{\hat{x}^n\in\mathcal{Z}_n} \tilde{P}_n(\hat{x}^n)\big|\varphi_n\big(\mu_n(\hat{x}^n)\big)\big|^\rho && (116)\\
&\geq 2^{\rho(H_{\tilde{\rho}}(\tilde{P}_n)-nR)}, && (117)
\end{aligned}$$

where the inequality (117) follows from (6) (with $\mathcal{Z}_n$, $\tilde{P}_n$, and $\mu_n$ taking the roles of $\mathcal{X}$, $P$ and $f$) by noting that $\varphi_n = \mu_n^{-1}$. In view of (114)–(117) the converse is proved once we show that

$$H_{\tilde{\rho}}(\tilde{P}_n) \geq nR_\rho(P,\Delta). \tag{118}$$

To prove (118), note that on account of (72) we have for every PMF $Q$ on $\mathcal{Z}_n$

$$H_{\tilde{\rho}}(\tilde{P}_n) \geq H(Q) - \rho^{-1}D(Q||\tilde{P}_n). \tag{119}$$

The PMF $\tilde{P}_n$ can be written as

$$\tilde{P}_n = P^n W_n, \tag{120}$$

where $W_n$ is the deterministic channel from $\mathcal{X}^n$ to $\hat{\mathcal{X}}^n$ induced by $g_n$:

$$W_n(\hat{x}^n|x^n) = \begin{cases} 1 & \text{if } \hat{x}^n = g_n(x^n), \\ 0 & \text{otherwise.}\end{cases} \tag{121}$$

Let $Q_\star$ be a PMF on $\mathcal{X}$ that achieves the maximum in the definition of $R_\rho(P,\Delta)$, i.e.,

$$R_\rho(P,\Delta) = R(Q_\star,\Delta) - \rho^{-1}D(Q_\star||P). \tag{122}$$

Substituting $Q_\star^n W_n$ for $Q$ in (119) and using (120),

$$\begin{aligned}
H_{\tilde{\rho}}(\tilde{P}_n) &\geq H(Q_\star^n W_n) - \rho^{-1}D(Q_\star^n W_n||P^n W_n) && (123)\\
&\geq H(Q_\star^n W_n) - \rho^{-1}D(Q_\star^n||P^n) && (124)\\
&= H(Q_\star^n W_n) - n\rho^{-1}D(Q_\star||P), && (125)
\end{aligned}$$

where (124) follows from the Data Processing Inequality [15, Lemma 3.11]. Let the source $\{\tilde{X}_i\}_{i=1}^\infty$ be IID $Q_\star$ and set $\hat{X}^n = g_n(\tilde{X}^n)$. Then

$$\begin{aligned}
H(Q_\star^n W_n) &= H(\hat{X}^n) && (126)\\
&= I(\tilde{X}^n \wedge \hat{X}^n). && (127)
\end{aligned}$$

By (111), we have

$$\mathrm{E}[d(\tilde{X}^n,\hat{X}^n)] \leq \Delta, \tag{128}$$

so applying [14, Th. 9.2.1] (which is the main ingredient in the classical rate-distortion converse) to the pair $(\tilde{X}^n,\hat{X}^n)$ gives

$$\begin{aligned}
I(\tilde{X}^n \wedge \hat{X}^n) &\geq nR\big(Q_\star, \mathrm{E}[d(\tilde{X}^n,\hat{X}^n)]\big) && (129)\\
&\geq nR(Q_\star,\Delta), && (130)
\end{aligned}$$

where (130) follows from (128) by the monotonicity of the rate-distortion function. Combining (129)–(130), (126)–(127), (123)–(125), and (122) establishes (118). □

## VIII. Tasks With Costs

We have so far assumed that every task requires an equal amount of effort. In this section, we discuss an extension where a nonnegative, finite cost $c(x)$ is associated with each task $x \in \mathcal{X}$. For the sake of simplicity, we limit ourselves to IID sources and $\rho = 1$.

For an $n$-tuple of tasks $x^n \in \mathcal{X}^n$, we denote by $c(x^n)$ the average cost per task:

$$c(x^n) = \frac{1}{n}\sum_{i=1}^n c(x_i). \tag{131}$$

We still assume that $n$-tuples of tasks are describe using $nR$ bits by an encoder of the form $f\colon \mathcal{X}^n \to \{1,\ldots,2^{nR}\}$, and that if $x^n$ is assigned, then all $n$-tuples in the set $f^{-1}(f(x^n))$ are performed. Thus, if $x^n$ is assigned, then the average cost per assigned task is

$$c(f,x^n) \triangleq \sum_{\tilde{x}^n\in f^{-1}(f(x^n))} c(\tilde{x}^n). \tag{132}$$

The following result extends Theorem I.2 to this setting (for IID tasks and $\rho = 1$). We focus on the case $\mathrm{E}[c(X_1)] > 0$ because otherwise we can achieve

$$\mathrm{E}\big[c(f,X^n)\big] = 0 \tag{133}$$

using only one bit by setting $f(x^n) = 1$ if $c(x^n) = 0$ and $f(x^n) = 2$ otherwise.

*Theorem VIII.1: Let $\{X_i\}_{i=1}^\infty$ be IID with finite alphabet $\mathcal{X}$ and* $\mathrm{E}[c(X_1)] > 0$.

1) *If $R > H_{1/2}(X_1)$, then there exist encoders $f_n\colon \mathcal{X}^n \to \{1,\ldots,2^{nR}\}$ such that*

$$\lim_{n\to\infty}\mathrm{E}\big[c(f_n,X^n)\big] = \mathrm{E}[c(X_1)]. \tag{134}$$

2) *If $R < H_{1/2}(X_1)$, then for any choice of encoders $f_n\colon \mathcal{X}^n \to \{1,\ldots,2^{nR}\}$,*

$$\lim_{n\to\infty}\mathrm{E}\big[c(f_n,X^n)\big] = \infty. \tag{135}$$

*Proof of Theorem VIII.1:* We begin with the case $R > H_{1/2}(X_1)$, i.e., the direct part. Let us denote by $c_{\max}$ the largest cost of any single task in $\mathcal{X}$

$$c_{\max} = \max_{x\in\mathcal{X}} c(x). \tag{136}$$

Select a sequence $f_n\colon \mathcal{X}^n \to \{1,\ldots,2^{nR}\}$ as in the direct part of Theorem I.2 and observe that

$$\begin{aligned}
&\mathrm{E}\big[c(f_n,X^n)\big] \\
&= \sum_{x^n\in\mathcal{X}^n} P^n(x^n)c(f_n,x^n) && (137)\\
&= \sum_{x^n\in\mathcal{X}^n} P^n(x^n)\Big(c(x^n) + \sum_{\tilde{x}^n\in f_n^{-1}(f_n(x^n))\setminus\{x^n\}} c(\tilde{x}^n)\Big) && (138)\\
&= \mathrm{E}\big[c(X_1)\big] + \sum_{x^n\in\mathcal{X}^n} P^n(x^n)\sum_{\tilde{x}^n\in f_n^{-1}(f_n(x^n))\setminus\{x^n\}} c(\tilde{x}^n) && (139)\\
&\leq \mathrm{E}\big[c(X_1)\big] + c_{\max}\sum_{x^n\in\mathcal{X}^n} P^n(x^n)|f_n^{-1}(f_n(x^n)) \setminus \{x^n\}| && (140)\\
&= \mathrm{E}\big[c(X_1)\big] + c_{\max}\big(\mathrm{E}\big[|f_n^{-1}(f_n(X^n))|\big] - 1\big), && (141)
\end{aligned}$$

and the second term on the RHS of (141) tends to zero as $n \to \infty$ by Theorem I.2.

We now turn to the case $R < H_{1/2}(X_1)$, i.e., the converse part. If the minimum cost of any single task $c_{\min}$ is positive, then (135) follows from the converse part of Theorem I.2 by replacing in (140) $c_{\max}$ with $c_{\min}$ and "$\leq$" with "$\geq$". If at least one task has zero cost (i.e., $c_{\min} = 0$), then we need a different proof.

The assumption $\mathrm{E}[c(X_1)] > 0$ implies that there is some $x^\star \in \mathcal{X}$ with $P(x^\star)c(x^\star) > 0$. Using Hölder's Inequality as in (34) with $p = q = 2$, $a(x) = \sqrt{P^n(x^n)c(f_n, x^n)}$, and $b(x) = \sqrt{c(x^n)/c(f_n, x^n)}$ gives

$$\begin{aligned}
&\sum_{x^n \in \mathcal{X}^n} P^n(x^n)c(f_n, x^n) \\
&\quad \geq \sum_{x^n : c(x^n) > 0} P^n(x^n)c(f_n, x^n) && (142) \\
&\quad \geq \frac{\left(\sum_{x^n : c(x^n) > 0} \sqrt{c(x^n)P^n(x^n)}\right)^2}{\sum_{x^n : c(x^n) > 0} \frac{c(x^n)}{c(f_n, x^n)}}. && (143)
\end{aligned}$$

To bound the denominator on the RHS of (143), observe that

$$\begin{aligned}
&\sum_{x^n : c(x^n) > 0} \frac{c(x^n)}{c(f_n, x^n)} \\
&= \sum_{m=1}^{2^{nR}} \sum_{x^n \in f_n^{-1}(m), c(x^n) > 0} \frac{c(x^n)}{\sum_{\tilde{x}^n \in f_n^{-1}(m)} c(\tilde{x}^n)} && (144) \\
&\leq 2^{nR}, && (145)
\end{aligned}$$

where the inequality follows because for some $m$ the set $\{x^n \in f_n^{-1}(m) : c(x^n) > 0\}$ may be empty. Combining (145) and (143) gives

$$\begin{aligned}
&\sum_{x^n \in \mathcal{X}^n} P^n(x^n)c(f_n, x^n) \\
&\quad \geq 2^{-nR}\Bigg(\sum_{x^n : c(x^n) > 0} \sqrt{c(x^n)P^n(x^n)}\Bigg)^2. && (146)
\end{aligned}$$

We can bound the sum on the RHS of (146) as follows.

$$\begin{aligned}
&\sum_{x^n : c(x^n) > 0} \sqrt{c(x^n)P^n(x^n)} \\
&\geq \sqrt{\frac{c(x^\star)}{n}} \sum_{Q \in \mathcal{P}_n(\mathcal{X}), Q(x^\star) > 0} \sum_{x^n \in T_Q} \sqrt{P^n(x^n)} && (147) \\
&\geq \sqrt{\frac{c(x^\star)}{n}} \max_{\substack{Q \in \mathcal{P}_n(\mathcal{X}) \\ Q(x^\star) > 0}} 2^{n(H(Q) - \delta_n)} 2^{-\frac{n}{2}(D(Q||P) + H(Q))} && (148) \\
&= 2^{\frac{n}{2}(\max_{Q \in \mathcal{P}_n(\mathcal{X}), Q(x^\star) > 0} H(Q) - D(Q||P) - \delta'_n)} && (149) \\
&= 2^{\frac{n}{2}(H_{1/2}(X_1) - \varepsilon_n - \delta'_n)}, && (150)
\end{aligned}$$

where $\delta_n = n^{-1}|\mathcal{X}| \log(n + 1)$, where $\delta'_n = 2\delta_n + n^{-1} \log(n/c(x^\star))$, and where $\varepsilon_n \to 0$ as $n \to \infty$. Here, (147) follows because if $x^n \in T_Q$ and $Q(x^\star) > 0$, then $x_i = x^\star$ for at least one $i$ and hence $c(x^n) \geq c(x^\star)/n > 0$; (148) follows because $P^n(x^n) = 2^{-n(D(Q||P)+H(Q))}$ when $x^n \in T_Q$, and because $|T_Q| \geq 2^{n(H(Q)-\delta_n)}$; (150) follows from (72) because the set of rational PMFs $Q$ with $Q(x^\star) > 0$ is dense in the set of all PMFs on $\mathcal{X}$, and $H(Q) - D(Q||P)$ is continuous in $Q$ (provided that $Q(x) = 0$ whenever $P(x) = 0$, which is certainly satisfied by the maximizing $Q$ in (72)). Combining (150) and (146) completes the proof of the converse. □

## Appendix A
## Proof of Proposition III.2

Since the labels do not matter, we may assume for convenience of notation that $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ and

$$\lambda(1) \leq \lambda(2) \leq \cdots \leq \lambda(|\mathcal{X}|). \tag{151}$$

We construct a partition of $\mathcal{X}$ as follows. The first subset is

$$\mathcal{L}_0 = \{x \in \mathcal{X} : \lambda(x) \geq |\mathcal{X}|\}. \tag{152}$$

If $\mathcal{X} = \mathcal{L}_0$, then the construction is complete and (39) and (40) are clearly satisfied. Otherwise we follow the steps below to construct additional subsets $\mathcal{L}_1, \ldots, \mathcal{L}_M$. (Note that if $\mathcal{L}_0 \neq \mathcal{X}$, then $\mathcal{X} \setminus \mathcal{L}_0 = \{1, \ldots, |\mathcal{X}| - |\mathcal{L}_0|\}$.)

*Step 1:* If

$$|\mathcal{X} \setminus \mathcal{L}_0| \leq \lambda(1), \tag{153}$$

then we complete the construction by setting $\mathcal{L}_1 = \mathcal{X} \setminus \mathcal{L}_0$ and $M = 1$. Otherwise we set

$$\mathcal{L}_1 = \big\{1, \ldots, \lambda(1)\big\} \tag{154}$$

and go to Step 2.

*Step $m \geq 2$:* If

$$\left|\mathcal{X} \setminus \bigcup_{i=0}^{m-1} \mathcal{L}_i\right| \leq \lambda(|\mathcal{L}_1| + \ldots + |\mathcal{L}_{m-1}| + 1), \tag{155}$$

then we complete the construction by setting $\mathcal{L}_m = \mathcal{X} \setminus \bigcup_{i=0}^{m-1} \mathcal{L}_i$ and $M = m$. Otherwise we let $\mathcal{L}_m$ contain the $\lambda(|\mathcal{L}_1| + \ldots + |\mathcal{L}_{m-1}| + 1)$ smallest elements of $\mathcal{X} \setminus \bigcup_{i=0}^{m-1} \mathcal{L}_i$, i.e., we set

$$\begin{aligned}
\mathcal{L}_m = \big\{&|\mathcal{L}_1| + \ldots + |\mathcal{L}_{m-1}| + 1, \ldots, |\mathcal{L}_1| + \ldots \\
&+ |\mathcal{L}_{m-1}| + \lambda(|\mathcal{L}_1| + \ldots + |\mathcal{L}_{m-1}| + 1)\big\}
\end{aligned} \tag{156}$$

and go to Step $m + 1$.

We next verify that (40) is satisfied and that the total number of subsets $M + 1$ does not exceed (39). Clearly, $L(x) \leq |\mathcal{X}|$ for every $x \in \mathcal{X}$, so to prove (40) we check that $L(x) \leq \lambda(x)$ for every $x \in \mathcal{X}$. From (152) it is clear that $L(x) \leq \lambda(x)$ for all $x \in \mathcal{L}_0$. Let $k(x)$ denote the smallest element in the subset containing $x$. Then $L(x) \leq \lambda(k(x))$ for all $x \in \bigcup_{m=1}^{M} \mathcal{L}_m$ by construction (the inequality can be strict only if $x \in \mathcal{L}_M$), and since $k(x) \leq x$, we have $\lambda(k(x)) \leq \lambda(x)$ by the assumption (151), and hence $L(x) \leq \lambda(x)$ for all $x \in \mathcal{X}$.

It remains to check that $M + 1$ does not exceed (39). This is clearly true when $M = 1$, so we assume that $M \geq 2$. Fix an arbitrary $\alpha > 1$ and let $\mathcal{M}$ be the set of indices $m \in \{1, \ldots, M - 1\}$ such that there is an $x \in \mathcal{L}_m$ with $\lambda(x) > \alpha\lambda(k(x))$. We next show that

$$|\mathcal{M}| < \log_\alpha |\mathcal{X}|. \tag{157}$$

To this end, enumerate the indices in $\mathcal{M}$ as $m_1 < m_2 < \cdots < m_{|\mathcal{M}|}$. For each $i \in \{1, \ldots, |\mathcal{M}|\}$ select some $x_i \in \mathcal{L}_{m_i}$ for which $\lambda(x_i) > \alpha\lambda(k(x_i))$. Then

$$\begin{aligned} \lambda(x_1) &> \alpha\lambda(k(x_1)) && (158)\\ &\geq \alpha. && (159) \end{aligned}$$

Note that if $1 \leq m < m'$ and $x \in \mathcal{L}_m$ and $x' \in \mathcal{L}_{m'}$, then $x < x'$. Thus, $x_1 < k(x_2)$ because $x_1 \in \mathcal{L}_{m_1}$, $k(x_2) \in \mathcal{L}_{m_2}$, and $m_1 < m_2$. Consequently, $\lambda(x_1) \leq \lambda(k(x_2))$ and hence

$$\begin{aligned} \lambda(x_2) &> \alpha\lambda(k(x_2)) && (160)\\ &\geq \alpha\lambda(x_1) && (161)\\ &> \alpha^2. && (162) \end{aligned}$$

Iterating this argument shows that

$$\lambda(x_{|\mathcal{M}|}) > \alpha^{|\mathcal{M}|}. \qquad (163)$$

And since $\lambda(x) < |\mathcal{X}|$ for every $x \notin \mathcal{L}_0$ by (152), the desired inequality (157) follows from (163). Let $\mathcal{M}^{\mathrm{c}}$ denote the complement of $\mathcal{M}$ in $\{1, \ldots, M-1\}$. Using Proposition III.1 and the fact that $L(x) = \lambda(k(x)) \geq \lambda(x)/\alpha$ for all $x \in \bigcup_{m \in \mathcal{M}^{\mathrm{c}}} \mathcal{L}_m$,

$$\begin{aligned} M &= \sum_{x \in \bigcup_{m=1}^{M} \mathcal{L}_m} \frac{1}{L(x)} && (164)\\ &= 1 + |\mathcal{M}| + \sum_{x \in \bigcup_{m \in \mathcal{M}^{\mathrm{c}}} \mathcal{L}_m} \frac{1}{L(x)} && (165)\\ &\leq 1 + |\mathcal{M}| + \alpha \sum_{x \in \bigcup_{m \in \mathcal{M}^{\mathrm{c}}} \mathcal{L}_m} \frac{1}{\lambda(x)} && (166)\\ &< 1 + \log_\alpha |\mathcal{X}| + \alpha\mu, && (167) \end{aligned}$$

where (167) follows from (157) and the hypothesis of the proposition (38). Since $M + 1$ is an integer and $\alpha > 1$ is arbitrary, it follows from (164)–(167) that $M + 1$ is upper-bounded by (39). □

## Appendix B
## Proof of (91)

We first show that $H(V|Q) - \rho^{-1} D(Q \circ V || P_{X,Y}) \leq H_{\tilde{\rho}}(X_1|Y_1)$ for every $Q \in \mathcal{P}(\mathcal{Y})$ and $V \in \mathcal{P}(\mathcal{X}|\mathcal{Y})$. This is clearly true when $D(Q \circ V || P_{X,Y}) = \infty$, so we may assume that $P_{X,Y}(x,y) = 0$ implies $Q(y)V(x|y) = 0$, and hence that $P_Y(y) = 0$ implies $Q(y) = 0$. Now observe that

$$\begin{aligned} &H(V|Q) - \rho^{-1} D(Q \circ V || P_{X,Y})\\ &= \frac{1+\rho}{\rho} \sum_{y \in \mathcal{Y}} Q(y) \sum_{x \in \mathcal{X}} V(x|y) \log \frac{P_{X|Y}(x|y)^{\frac{1}{1+\rho}}}{V(x|y)}\\ &\quad - \frac{1}{\rho} \sum_{y \in \mathcal{Y}} Q(y) \log \frac{Q(y)}{P_Y(y)} && (168)\\ &\leq \frac{1+\rho}{\rho} \sum_{y \in \mathcal{Y}} Q(y) \log \sum_{x \in \mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\\ &\quad - \frac{1}{\rho} \sum_{y \in \mathcal{Y}} Q(y) \log \frac{Q(y)}{P_Y(y)} && (169)\\ &= \frac{1}{\rho} \sum_{y \in \mathcal{Y}} Q(y) \log \frac{P_Y(y) \big(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\big)^{1+\rho}}{Q(y)} && (170)\\ &\leq \frac{1}{\rho} \log \sum_{y \in \mathcal{Y}} P_Y(y) \bigg(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\bigg)^{1+\rho} && (171)\\ &= H_{\tilde{\rho}}(X_1|Y_1), && (172) \end{aligned}$$

where (169) and (171) follow from Jensen's Inequality. The proof is completed by noting that equality is attained in both inequalities by the choice

$$Q(y) = \frac{P_Y(y) \big(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\big)^{1+\rho}}{\sum_{y' \in \mathcal{Y}} P_Y(y') \big(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y')^{\frac{1}{1+\rho}}\big)^{1+\rho}}, \qquad (173)$$

and

$$V(x|y) = \frac{P_{X|Y}(x|y)^{\frac{1}{1+\rho}}}{\sum_{x' \in \mathcal{X}} P_{X|Y}(x'|y)^{\frac{1}{1+\rho}}}, \quad Q(y) > 0. \qquad (174)$$

(Note that $P_Y(y) > 0$ when $Q(y) > 0$ so the RHS of (174) makes sense. How we define $V(x|y)$ when $Q(y) = 0$ does not matter.) □

## Acknowledgment

We are grateful for valuable comments from the anonymous reviewers.

## References

[1] Z. Rached, F. Alajaji, and L. L. Campbell, "Rényi's divergence and entropy rates for finite alphabet markov sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1553–1561, May 2001.

[2] C. E. Pfister and W. G. Sullivan, "Rényi entropy, guesswork moments, and large deviations," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[3] Z. Rached, A. Fady, and L. L. Campbell, "Rényi's entropy rate for discrete markov sources," in *Proc. CISS*, vol. 99. 1999, pp. 17–19.

[4] P.-N. Chen and F. Alajaji, "Csiszar's cutoff rates for arbitrary discrete sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 330–338, Jan. 2001.

[5] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. Inf. Theory*, vol. 50, no. 3, pp. 525–526, Mar. 2004.

[6] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.

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

[8] A. Rényi, "On the foundations of information theory," *Rev. Int. Statist. Inst.*, vol. 33, no. 1, pp. 1–14, 1965.

[9] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.

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

[11] J. L. Massey, "Guessing and entropy," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 1994, p. 204.

[12] A. Bracher, E. Hof, and A. Lapidoth. (2014). *Distributed Storage for Data Security* [Online]. Available: http://arxiv.org/abs/1405.4981

[13] O. Shayevitz. A note on a characterization of Rényi measures and its relation to composite hypothesis testing [Online]. Available: http://arxiv.org/abs/1012.4401

[14] R. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Cambridge Univ. Press, 2011.

[16] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 269–287, Jan. 2007.

[17] E. Arikan and N. Merhav, “Guessing subject to distortion,” *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[18] S. Arimoto, “Information measures and capacity of order $\alpha$ for discrete memoryless channels,” in *Topics in Information Theory*, vol. 17, I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland, 1977, pp. 41–52.

**Christoph Bunte** received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering in 2010, 2011, and 2014, all from ETH Zurich, Switzerland.

He is currently a research assistant in the Signal and Information Processing Laboratory at ETH Zurich.

His research interests are in the area of Shannon theory.

**Amos Lapidoth** (S'89–M'95–SM'00–F'04) received the B.A. degree in mathematics (summa cum laude, 1986), the B.Sc. degree in electrical engineering (summa cum laude, 1986), and the M.Sc. degree in electrical engineering (1990) all from the Technion—Israel Institute of Technology. He received the Ph.D. degree in electrical engineering from Stanford University in 1995.

In the years 1995–1999 he was an Assistant and Associate Professor at the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, and was the KDD Career Development Associate Professor in Communications and Technology. He is now Professor of Information Theory at ETH Zurich in Switzerland. He is the author of the book *A Foundation in Digital Communication* (Cambridge University Press, 2009). His research interests are in digital communications and information theory.

Dr. Lapidoth served in the years 2003–2004 and 2009 as Associate Editor for Shannon Theory for the IEEE TRANSACTIONS ON INFORMATION THEORY.