# Non-Asymptotic Bounds of Cumulant Generating Function of Codeword Lengths in Variable-Length Lossy Compression

Shota Saito, *Member, IEEE,* and Toshiyasu Matsushima, *Member, IEEE*

***Abstract*—This paper investigates the problem of variable-length source coding with the criteria of the normalized cumulant generating function of codeword lengths and the excess distortion probability. We analyze the non-asymptotic fundamental limit of the normalized cumulant generating function of codeword lengths under the constraint that the excess distortion probability is allowed up to $\epsilon \in [0,1)$. Our non-asymptotic achievability and converse bounds are characterized by the quantity related to the Rényi entropy.**

***Index Terms*—Cumulant generating function of codeword lengths, excess distortion probability, Rényi entropy, Shannon theory, variable-length lossy source coding**

## I. Introduction

### A. Research Motivation

THE problem of variable-length source coding is one of the important research topics in Shannon theory. A variable-length *lossless* source code without prefix constraints is an injective mapping $f : \mathcal{X} \to \{0,1\}^\star$, where $\mathcal{X}$ denotes a source alphabet and $\{0,1\}^\star$ is the set of all finite-length binary strings and the empty string $\lambda$, i.e.,

$$\{0,1\}^\star := \{\lambda, 0, 1, 00, 01, 10, 11, 000, \ldots\}. \tag{1}$$

For a discrete set $\mathcal{X}$ and a probability mass function $P_X$ on $\mathcal{X}$, a $P_X$-optimal source code is a variable-length lossless source code without prefix constraints that satisfies the following properties (see, e.g., [5], [9], [18]):

i) For all $x, x' \in \mathcal{X}$, $P_X(x) \geq P_X(x') \iff \ell(f(x)) \leq \ell(f(x'))$, where $\ell(f(x))$ denotes the length of $f(x)$;
ii) $\ell(f(x')) = l \in \{0,1,2,\ldots\}$ for an $x' \in \mathcal{X} \Longrightarrow |\{x \in \mathcal{X} : \ell(f(x)) < l\}| = 2^l - 1$, where $|\cdot|$ denotes the cardinality of a set.

Although a $P_X$-optimal source code is not unique, it is convenient to consider the following optimal source code $f^* : \mathcal{X} \to \{0,1\}^\star$: let $x_i \in \mathcal{X}$ be a source symbol that has the $i$-th largest probability (i.e., $P_X(x_1) \geq P_X(x_2) \geq P_X(x_3) \geq \cdots$), then $f^*$ maps source symbols $x_1, x_2, x_3, \ldots$ to the elements of $\{0,1\}^\star$ in the lexicographic order, i.e., $f^*(x_1) = \lambda, f^*(x_2) = 0, f^*(x_3) = 1, f^*(x_4) = 00, f^*(x_5) = 01, f^*(x_6) = 10, f^*(x_7) = 11, f^*(x_8) = 000, \ldots$.

For $f^*$, the normalized cumulant generating function of codeword lengths $(1/\rho)\log_2 \mathbb{E}[2^{\rho \ell(f^*(X))}]$ ($\rho > 0$ is a free parameter) was investigated by Courtade and Verdú [5], which was later improved by Sason and Verdú [18]. The normalized cumulant generating function of codeword lengths was first proposed by Campbell [4] and this is a generalized notion of the mean codeword length and the maximum codeword length. In [5] and [18], it was shown that *non-asymptotic achievability and converse bounds* of $(1/\rho)\log_2 \mathbb{E}[2^{\rho \ell(f^*(X))}]$ are characterized by $H_{\frac{1}{1+\rho}}(X)$—the Rényi entropy of order $1/(1+\rho)$ [16], i.e.,

$$H_{\frac{1}{1+\rho}}(X) := \frac{1}{1-(1/(1+\rho))} \log_2 \sum_{x \in \mathcal{X}} [P_X(x)]^{\frac{1}{1+\rho}}. \tag{2}$$

On the other hand, regarding variable-length *lossy* source coding, the problem is more complicated because we have to consider a distortion $d : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$ between a source symbol $x \in \mathcal{X}$ and a reproduction symbol $y \in \mathcal{Y}$, where $\mathcal{Y}$ denotes a reproduction alphabet. For this problem, is it possible to derive *non-asymptotic achievability and converse bounds* of the cumulant generating function of codeword lengths by using a *quantity related to the Rényi entropy*?

### B. Contributions

This paper gives a positive answer to this question. The following quantity related to the Rényi entropy plays an important role in characterizing non-asymptotic achievability and converse bounds.

*Definition 1:* Given $D \geq 0$, $\epsilon \in [0,1)$, and $\alpha > 0, \alpha \neq 1$, $\mathbb{H}_\alpha(X, D, \epsilon)$ is defined as

$$\mathbb{H}_\alpha(X, D, \epsilon) := \inf_{\substack{P_{Y|X}: \\ \mathbb{P}[d(X,Y) > D] \leq \epsilon}} H_\alpha(Y), \tag{3}$$

where $P_{Y|X}$ denotes a conditional probability distribution of $Y$ given $X$.

The properties of $\mathbb{H}_\alpha(X, D, \epsilon)$ are summarized in the following proposition (the proof is in Appendix A):

*Proposition 1:*

a) $\mathbb{H}_\alpha(X, D, \epsilon) \geq 0$.
b) $\mathbb{H}_\alpha(X, D, \epsilon)$ is a concave function of $P_X$.
c) $\mathbb{H}_\alpha(X, D, \epsilon)$ is decreasing in $\alpha$, i.e., $\mathbb{H}_{\alpha_2}(X, D, \epsilon) \leq \mathbb{H}_{\alpha_1}(X, D, \epsilon)$ for $0 < \alpha_1 \leq \alpha_2$.

Shota Saito is with the Faculty of Informatics, Gunma University, 4-2, Maebashi, Gunma 371-8510, JAPAN (e-mail: shota.s@gunma-u.ac.jp)

Toshiyasu Matsushima is with the Department of Applied Mathematics, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, JAPAN (e-mail: toshimat@waseda.jp).

Manuscript received April 19, 2005; revised August 26, 2015. This work was supported in part by JSPS KAKENHI Grant Numbers JP17K06446, JP18K11585, JP19K04914, and JP19K14989. This paper was presented in part at the 2018 International Symposium on Information Theory (ISIT2018) [17].

d) $\mathbb{H}_\alpha(X, D, \epsilon)$ is decreasing in $\epsilon$, i.e., $\mathbb{H}_\alpha(X, D, \epsilon_2) \leq \mathbb{H}_\alpha(X, D, \epsilon_1)$ for $0 \leq \epsilon_1 \leq \epsilon_2 < 1$.

e) $\mathbb{H}_\alpha(X, D, \epsilon)$ is decreasing in $D$, i.e., $\mathbb{H}_\alpha(X, D_2, \epsilon) \leq \mathbb{H}_\alpha(X, D_1, \epsilon)$ for $0 \leq D_1 \leq D_2$.

f) $\mathbb{H}_\alpha(X, 0, 0) = H_\alpha(X)$ if $\mathcal{X} = \mathcal{Y}$, $d(x, y) = 0$ for $x = y$, and $d(x, y) > 0$ for $x \neq y$.

By using $\mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon)$, we characterize the *non-asymptotic achievability and converse bounds* of the normalized cumulant generating function of codeword lengths under the constraint that $\mathbb{P}[d(X, Y) > D] \leq \epsilon$ in variable-length *lossy* source coding. In variable-length *lossless* source coding, on the other hand, the non-asymptotic fundamental limit of the normalized cumulant generating function of codeword lengths is characterized by the Rényi entropy $H_{\frac{1}{1+\rho}}(X)$ as we have described in Section I-A. From Item f) above, when we consider lossless setup, we see that the leading term in our upper and lower bounds $\mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon)$ reduces to the leading term $H_{\frac{1}{1+\rho}}(X)$ in the previous studies [4], [5]. Therefore, $\mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon)$ can be seen as the "lossy source coding version" of the Rényi entropy $H_{\frac{1}{1+\rho}}(X)$.

### C. Related Works

Regarding the problem of variable-length source coding under the criterion of the normalized cumulant generating function of codeword lengths, previous works are summarized as follows:

1) For the problem of variable-length lossless source coding, Campbell [4] and Sason [19] investigated the non-asymptotic fundamental limit for prefix codes. Courtade and Verdú [5] studied it for non-prefix codes, which was improved by Sason and Verdú [18]. All of these works [4], [5], [18], [19] characterized the fundamental limit by using the Rényi entropy. As we shall show in Remark 1, our result coincides with the result in [5] when we consider a special case.
2) For the problem of variable-length source coding allowing errors, Kuzuoka [14] characterized the non-asymptotic fundamental limit for prefix codes by using the smooth Rényi entropy.
3) For the problem of variable-length lossy source coding, Courtade and Verdú [6] derived the non-asymptotic converse bound for non-prefix codes by using the $d$-tilted Rényi entropy. It should be noted that our study extends the previous work [6] in the sense that i) we evaluate the non-prefix code whose excess distortion probability may be positive and ii) we derive both achievability and converse bounds (see Remark 3).

The problems of variable-length source coding are related to the problems of guessing (e.g., [1], [2], [14], [18],[19]). In the problem of guessing subject to distortion studied by Arikan and Merhav [2], they derived a single-letter characterization of the smallest achievable exponential growth rate of the $\rho$th moment of the number of guesses for a memoryless source. In Remark 6, we shall describe the relationship between the result in [2] and our result.

### D. Organization of the Paper

The rest of the paper is organized as follows. In Section II, we explain the problem formulation and previous study; Section II-A describes the setup and Section II-B reviews the previous work by Courtade and Verdú [6]. In Section III, we show our main results—the non-asymptotic achievability and converse bounds of the normalized cumulant generating function of codeword lengths in variable-length lossy source coding—and provide the proofs.

## II. Problem Formulation and Previous Study

### A. Setup

We consider the variable length lossy source code *without* the prefix condition[1], that is, code defined by triplets[2] $(\widehat{\mathcal{Y}}, \phi, \psi)$ where

- $\widehat{\mathcal{Y}}$ is a subset of $\mathcal{Y}$ called the codebook,
- $\phi : \mathcal{X} \to \widehat{\mathcal{Y}}$ is a (possibly stochastic) encoder,
- $\psi : \widehat{\mathcal{Y}} \to \{0, 1\}^\star := \{\lambda, 0, 1, 00, 01, 10, 11, 000, \ldots\}$ is a one-to-one representation of the elements of $\widehat{\mathcal{Y}}$ by finite-length binary strings and the empty string $\lambda$.

The performance criteria we consider in this paper are the excess distortion probability and the normalized cumulant generating function of codeword lengths. Although the average distortion $\mathbb{E}[d(X, \phi(X))]$ is a popular criterion, the excess distortion probability is also used in previous studies such as [7], [8], [10], [11], [12], [13]. The motivation for using the excess distortion probability is elaborated in [11, Section 1.8.2]. The motivation for using the cumulant generating function of codeword lengths is described in Remark 1.

*Definition 2:* Given $D \geq 0$, the excess distortion probability is defined as

$$\mathbb{P}[d(X, \phi(X)) > D]. \tag{4}$$

*Definition 3:* Let $\ell(\psi(\phi(x)))$ denote a length of $\psi(\phi(x))$. Then, given $\rho > 0$, the normalized cumulant generating function of codeword lengths is defined as

$$\frac{1}{\rho} \log_2 \mathbb{E}[2^{\rho \ell(\psi(\phi(X)))}]. \tag{5}$$

*Remark 1:* The l'Hôspital theorem yields

$$\lim_{\rho \downarrow 0} \frac{1}{\rho} \log_2 \mathbb{E}[2^{\rho \ell(\psi(\phi(X)))}] = \mathbb{E}[\ell(\psi(\phi(X)))], \tag{6}$$

$$\lim_{\rho \to \infty} \frac{1}{\rho} \log_2 \mathbb{E}[2^{\rho \ell(\psi(\phi(X)))}] = \max_{x \in \mathcal{X}} \ell(\psi(\phi(x))). \tag{7}$$

Thus, the normalized cumulant generating function of codeword lengths contains the mean codeword length and the maximum codeword length as its special cases. As mentioned in [18, Section I] and [19, Section 5.2.1], by using the normalized cumulant generating function of codeword lengths as a design criterion, we can control the contribution of the longer codewords via a free parameter $\rho$; if we increase the value of $\rho$, we impose a more severe penalty for longer codewords. This

[1] Variable-length source coding without the prefix condition is discussed by various authors [5], [6], [9], [12], and [18]. In particular, the motivation for considering such code is elaborated in [9, Section I].

[2] This definition is based on [7] and [8].

is one of the motivations for using the cumulant generating function of codeword lengths.

Using these criteria, we define a $(D, R, \epsilon, \rho)$ code and the fundamental limit $R_X^*(D, \epsilon, \rho)$:

*Definition 4:* Given $D, R \geq 0$, $\epsilon \in [0, 1)$, and $\rho > 0$, a code $(\widehat{\mathcal{Y}}, \phi, \psi)$ satisfying

$$\mathbb{P}[d(X, \phi(X)) > D] \leq \epsilon, \tag{8}$$

$$\frac{1}{\rho} \log_2 \mathbb{E}[2^{\rho \ell(\psi(\phi(X)))}] \leq R \tag{9}$$

is called a $(D, R, \epsilon, \rho)$ code. The fundamental limit is

$$R_X^*(D, \epsilon, \rho) := \inf\{R : \exists \text{ a } (D, R, \epsilon, \rho) \text{ code}\}. \tag{10}$$

*Remark 2:* For a given $D \geq 0$ and $\epsilon \in [0, 1)$, suppose that

$$\mathbb{P}\left[\min_{y \in \mathcal{Y}} d(X, y) > D\right] > \epsilon. \tag{11}$$

Then, there are no codes whose excess distortion probability is less than or equal to $\epsilon$. Conversely, if such codes do not exist for given $D$ and $\epsilon$, (11) holds. In this case, we define $R_X^*(D, \epsilon, \rho) = +\infty$. Furthermore, if (11) holds, we also define $\mathbb{H}_\alpha(X, D, \epsilon) = +\infty$ because there is no conditional probability distribution $P_{Y|X}$ satisfying $\mathbb{P}[d(X, Y) > D] \leq \epsilon$.

### B. Previous Study

For $D \geq 0$, let $R(D, P_X)$ be the rate-distortion function, i.e.,

$$R(D, P_X) := \inf_{\substack{P_{Y|X}: \\ \mathbb{E}[d(X,Y)] \leq D}} I(X; Y), \tag{12}$$

where $I(X; Y)$ denotes the mutual information between random variables $X$ and $Y$. We assume that the infimum in the rate-distortion function $R(D, P_X)$ is achieved by $P_{Y^\dagger|X}$. Further, let $Y^\dagger$ be a random variable taking values in $\mathcal{Y}$ and whose distribution $P_{Y^\dagger}$ is the marginal of $P_{Y^\dagger|X} P_X$. Then, the $d$-tilted information of $x \in \mathcal{X}$ is defined as[3]

$$\jmath_X(x, D) := -\log_2 \mathbb{E}\left[2^{\lambda^\dagger \{D - d(x, Y^\dagger)\}}\right], \tag{13}$$

where the expectation is with respect to $P_{Y^\dagger}$ and $\lambda^\dagger := -\frac{\partial}{\partial D} R(D, P_X)$. Further, the $d$-tilted Rényi entropy of order $\alpha$ ($\alpha > 0, \alpha \neq 1$) is defined as [6]

$$\mathbb{R}_\alpha(X, D) := \frac{1}{1 - \alpha} \log_2 \mathbb{E}[2^{(1-\alpha)\jmath_X(X, D)}]. \tag{14}$$

Courtade and Verdú [6] considered the same setup as in Section II-A with the restriction that a code satisfies

$$\mathbb{P}[d(X, \phi(X) > D] = 0, \tag{15}$$

i.e., $\epsilon = 0$ in (8). They showed the converse bound on $R^*(D, 0, \rho)$ by using the $d$-tilted Rényi entropy:

*Theorem 1 ([6]):* Let $D \geq 0$ be such that $R(D, P_X) < +\infty$. Then, for any $\rho > 0$,

$$R_X^*(D, 0, \rho) \geq \mathbb{R}_{\frac{1}{1+\rho}}(X, D) - \log_2 \log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}). \tag{16}$$

[3]Kostina and Verdú [10] named this quantity the $d$-tilted information. This quantity was used in earlier work by, e.g., Kontoyiannis (see [7], [8]). Kostina [13] looked into the behavior of the $d$-tilted information.

*Remark 3:* As we have described above, the previous study [6] investigated the case where the excess distortion probability is zero and showed the converse bound. On the other hand, our study deals with positive excess distortion probability and investigates both achievability and converse bounds.

## III. Main Results and Proofs

The next proposition shows the achievability result on $R$ of a $(D, R, \epsilon, \rho)$ code.

*Proposition 2:* Let $D \geq 0$ and $\epsilon \in [0, 1)$ be such that $\mathbb{H}_\alpha(X, D, \epsilon) < +\infty$ for $\alpha > 0, \alpha \neq 1$. Then, for any $\rho > 0$, there exists a $(D, R, \epsilon, \rho)$ code such that

$$R = \mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon). \tag{17}$$

*Proof:* Let $P_{Y^\star|X}$ be the conditional probability distribution that achieves the infimum of $\mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon)$. We denote by $P_{Y^\star}$ the marginal of $P_{Y^\star|X} P_X$, i.e., $P_{Y^\star}$ is the probability distribution on $\mathcal{Y}$ defined by

$$P_{Y^\star}(y) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y^\star|X}(y|x) \tag{18}$$

for $y \in \mathcal{Y}$. Let $\widehat{\mathcal{Y}}$ be the support of $P_{Y^\star}$.

Then, we define the stochastic encoder $\phi : \mathcal{X} \to \widehat{\mathcal{Y}}$ as $\phi(x) = y$ with probability $P_{Y^\star|X}(y|x)$. The function $\psi : \widehat{\mathcal{Y}} \to \{0, 1\}^\star$ is defined as follows: first, we index the elements in $\widehat{\mathcal{Y}}$ as $y_1, y_2, y_3, \ldots$ such that

$$P_{Y^\star}(y_1) \geq P_{Y^\star}(y_2) \geq P_{Y^\star}(y_3) \geq \cdots. \tag{19}$$

Then, $\psi$ maps symbols $y_1, y_2, y_3, \ldots \in \widehat{\mathcal{Y}}$ to the elements of $\{0, 1\}^\star$ in the lexicographic order, i.e.,

$$\begin{aligned}
\psi(y_1) &= \lambda, \\
\psi(y_2) &= 0, \\
\psi(y_3) &= 1, \\
\psi(y_4) &= 00, \\
\psi(y_5) &= 01, \\
\psi(y_6) &= 10, \\
\psi(y_7) &= 11, \\
\psi(y_8) &= 000, \cdots
\end{aligned} \tag{20}$$

Regarding this code, we evaluate the excess distortion probability and the normalized cumulant generating function of codeword lengths. First, the excess distortion probability is evaluated as

$$\begin{aligned}
&\mathbb{P}[d(X, \phi(X)) > D] \\
&\quad = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y^\star|X}(y|x) \mathbf{1}\{d(x, y) > D\} \qquad (21) \\
&\quad \leq \epsilon, \qquad (22)
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and the final inequality follows from the definition of $P_{Y^\star|X}$.

Next, the normalized cumulant generating function of codeword lengths is calculated as

$$\begin{aligned}
&\frac{1}{\rho}\log_2 \mathbb{E}[2^{\rho\ell(\psi(\phi(X)))}] \\
&= \frac{1}{\rho}\log_2 \sum_{x\in\mathcal{X}}\sum_{i=1}^{|\widehat{\mathcal{Y}}|} P_X(x)P_{Y^\star|X}(y_i|x)2^{\rho\ell(\psi(y_i))} && (23)\\
&= \frac{1}{\rho}\log_2 \sum_{i=1}^{|\widehat{\mathcal{Y}}|}\left\{\sum_{x\in\mathcal{X}} P_X(x)P_{Y^\star|X}(y_i|x)\right\}2^{\rho\ell(\psi(y_i))} && (24)\\
&= \frac{1}{\rho}\log_2 \sum_{i=1}^{|\widehat{\mathcal{Y}}|} P_{Y^\star}(y_i)2^{\rho\ell(\psi(y_i))} && (25)\\
&\overset{(a)}{\le} \frac{1}{\rho}\log_2 \sum_{i=1}^{|\widehat{\mathcal{Y}}|} P_{Y^\star}(y_i)\left[\sum_{j=1}^{|\widehat{\mathcal{Y}}|}\left(\frac{P_{Y^\star}(y_j)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}}\right]^{\rho} && (26)\\
&= \frac{1}{\rho}\log_2 \left(\sum_{k=1}^{|\widehat{\mathcal{Y}}|}[P_{Y^\star}(y_k)]^{\frac{1}{1+\rho}}\right)^{1+\rho} && (27)\\
&\overset{(b)}{=} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon), && (28)
\end{aligned}$$

where $(a)$ follows from the next lemma (the proof of Lemma 1 is in Appendix B) and $(b)$ is due to the definition of $P_{Y^\star|X}$.

*Lemma 1:* For any $\rho > 0$ and $i \in \{1, 2, \ldots, |\widehat{\mathcal{Y}}|\}$, we have

$$2^{\rho\ell(\psi(y_i))} \le i^\rho \le \left[\sum_{j=1}^{|\widehat{\mathcal{Y}}|}\left(\frac{P_{Y^\star}(y_j)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}}\right]^{\rho}. \quad (29)$$

Putting together the pieces, we complete the proof of Proposition 2. ∎

The next proposition shows the converse bound on $R$ of a $(D, R, \epsilon, \rho)$ code.

*Proposition 3:* Let $D \ge 0$ and $\epsilon \in [0, 1)$ be such that $\mathbb{H}_\alpha(X, D, \epsilon) < +\infty$ for $\alpha > 0, \alpha \ne 1$. Then, for any $\rho > 0$, every $(D, R, \epsilon, \rho)$ code satisfies

$$R \ge \mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon) - \log_2\log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}). \quad (30)$$

*Proof:* Fix a $(D, R, \epsilon, \rho)$ code $(\widehat{\mathcal{Y}}, \phi, \psi)$ arbitrarily and let $Y := \phi(X)$. Then, the definition of a $(D, R, \epsilon, \rho)$ code gives

$$\mathbb{P}[d(X, Y) > D] \le \epsilon, \quad (31)$$

$$\frac{1}{\rho}\log_2 \mathbb{E}[2^{\rho\ell(\psi(Y))}] \le R, \quad (32)$$

and the assumption that $\psi$ is a one-to-one mapping yields the next inequality (see [5]):

$$\sum_{y\in\widehat{\mathcal{Y}}} 2^{-\ell(\psi(y))} \le \log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}). \quad (33)$$

The key lemma in the proof of the converse result is the following (the proof of Lemma 2 is in Appendix C):

*Lemma 2:* For any $\rho > 0$, we have

$$\frac{1}{\rho}\log_2 \mathbb{E}[2^{\rho\ell(\psi(Y))}] \ge H_{\frac{1}{1+\rho}}(Y) - \log_2 \sum_{y\in\widehat{\mathcal{Y}}} 2^{-\ell(\psi(y))}. \quad (34)$$

Combination of (32), (33), and (34) yields

$$R \ge H_{\frac{1}{1+\rho}}(Y) - \log_2\log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}). \quad (35)$$

Finally, from (3), (31), and (35), we have (30). ∎

*Remark 4:* If both $\mathcal{X}$ and $\mathcal{Y}$ are infinite set, the bound (30) is useless. This problem is inherited from the analysis of [5] and [6]. In [5], the bound of the normalized cumulant generating function of codeword lengths was applied to the analysis of the asymptotic normality of codeword lengths of $P_X$-optimal lossless source code[4] for infinite alphabet $\mathcal{X}$ via a procedure that made the term $\log_2\log_2(1 + |\mathcal{X}|)$ negligible. There is a possibility that a similar application of our result can be considered, but this is an important open question.

Combining Propositions 2 and 3, we can immediately obtain the following result on $R^*_X(D, \epsilon, \rho)$.

*Theorem 2:* Let $D \ge 0$ and $\epsilon \in [0, 1)$ be such that $\mathbb{H}_\alpha(X, D, \epsilon) < +\infty$ for $\alpha > 0, \alpha \ne 1$. Then, for any $\rho > 0$, we have

$$\begin{aligned}
&\mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon) - \log_2\log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}) \\
&\le R^*_X(D, \epsilon, \rho) \le \mathbb{H}_{\frac{1}{1+\rho}}(X, D, \epsilon). && (36)
\end{aligned}$$

We give two examples regarding the general formula (36) in the following.

*Example 1:* Let us consider the special case where $\epsilon = 0$, $D = 0$, $\mathcal{X} = \mathcal{Y}$, $d(x, y) = 0$ for $x = y$, and $d(x, y) = 1$ for $x \ne y$. In this case, from Item f) of Proposition 1, the bound (36) reduces to

$$H_{\frac{1}{1+\rho}}(X) - \log_2\log_2(1 + |\mathcal{X}|) \le R^*_X(0, 0, \rho) \le H_{\frac{1}{1+\rho}}(X), \quad (37)$$

which coincides with the bound in [5].

*Example 2:* Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets. In this example, we consider the special case $\rho \to 0$ and drive a single-letter characterization of the fundamental limit $R^*_{X^n}(D, \epsilon, 0) := \lim_{\rho\to 0} R^*_{X^n}(D, \epsilon, \rho)$ under a stationary memoryless source. As in [12, Section III-E], we assume the following conditions:

1) The source $\{X_i\}$ is stationary and memoryless.
2) The distortion measure is separable, i.e., $d(x^n, y^n) := \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$ for $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.
3) $D$ satisfies $D \in (D_{\min}, D_{\max})$, where $D_{\min} := \inf\{D : R(D, P_X) < \infty\}$ and $D_{\max} := \min_{y\in\mathcal{Y}} \mathbb{E}[d(X, y)]$.
4) $\mathbb{E}[d^{12}(X, Y^\dagger)] < \infty$, where the expectation is with respect to $P_X \times P_{Y^\dagger}$ (regarding $Y^\dagger$ and $P_{Y^\dagger}$, see Section II-B).

Under assumptions 1) – 4), we have

$$\begin{aligned}
R^*_{X^n}(D, \epsilon, 0) &= (1 - \epsilon)nR(D, P_X) \\
&\quad - \sqrt{\frac{nV(D, P_X)}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O(\log_2 n) && (38)
\end{aligned}$$

for any $\epsilon \in [0, 1)$, where $R(D, P_X)$ is the rate-distortion function (see (12)), $V(D, P_X)$ is the rate-dispersion function that is defined as the variance of the $d$-tilted information, i.e., $V(D, P_X) := \mathrm{Var}[\jmath_X(X, D)]$ (see, e.g., [10]), and $Q^{-1}(z)$ denotes the inverse function of $Q(z) = \int_z^\infty (1/\sqrt{2\pi})e^{-\frac{t^2}{2}}dt$ for $z \in \mathbb{R}$. The proof of (38) is in Appendix D.

[4]Regarding the $P_X$-optimal lossless source code, see Section I-A.

*Remark 5:* In the special case of $\epsilon = 0$, Theorem 2 reduces to

$$\mathbb{H}_{\frac{1}{1+\rho}}(X, D, 0) - \log_2 \log_2(1 + \min\{|\mathcal{X}|, |\mathcal{Y}|\}) \leq R_X^*(D, 0, \rho) \leq \mathbb{H}_{\frac{1}{1+\rho}}(X, D, 0). \tag{39}$$

On the other hand, as shown in Theorem 1, Courtade and Verdú [6] characterized the lower bound of $R_X^*(D, 0, \rho)$ by using the $d$-tilted Rényi entropy $\mathbb{R}_{\frac{1}{1+\rho}}(X, D)$. One question is the relationship between $\mathbb{H}_{\frac{1}{1+\rho}}(X, D, 0)$ and $\mathbb{R}_{\frac{1}{1+\rho}}(X, D)$, but we do not have the answer yet. To investigate the relationship is future work.

*Remark 6:* Let $\mathrm{KL}(\cdot||\cdot)$ denote the relative entropy. Then, for $\rho > 0$, the quantity $\mathbb{G}_X(D, \rho)$ is defined as

$$\mathbb{G}_X(D, \rho) := \sup_{Q_X} \left[\rho R(D, Q_X) - \mathrm{KL}(Q_X||P_X)\right], \tag{40}$$

where the supremum in (40) is taken over all probability distributions $Q_X$ on $\mathcal{X}$. For a stationary memoryless source and finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, Arikan and Merhav [2] proved that the smallest achievable exponential growth rate of the $\rho$th moment of the number of guesses is characterized by $\mathbb{G}_X(D, \rho)$ (see [2, Corollary 1]). From (57), (58), (64) and the fact that $\lim_{\rho\to 0} \frac{1}{\rho}\mathbb{G}_X(D, \rho) = R(D, P_X)$ (see [2, Proposition 1]), it holds that

$$\lim_{n\to\infty} \lim_{\rho\to 0} \frac{1}{n} \mathbb{H}_{\frac{1}{1+\rho}}(X^n, D, 0) = R(D, P_X) = \lim_{\rho\to 0} \frac{1}{\rho} \mathbb{G}_X(D, \rho) \tag{41}$$

for a stationary memoryless source and finite alphabets $\mathcal{X}$ and $\mathcal{Y}$.

## Appendix A
## Proof of Proposition 1

Property a) is clear from the non-negativity of the Rényi entropy. Property b) follows from the following reason: for $\alpha > 0, \alpha \neq 1$ and a fixed $P_{Y|X}$ satisfying $\mathbb{P}[d(X, Y) > D] \leq \epsilon$,

$$H_\alpha(Y) = \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} [P_Y(y)]^\alpha \tag{42}$$

$$= \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} \left[\sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_X(x)\right]^\alpha \tag{43}$$

is a concave function of $P_X$. Because the pointwise infimum of concave functions is also concave (see, e.g., [3]), we have Property b). Property c) follows from the fact that the Rényi entropy of order $\alpha$ is decreasing in $\alpha$, i.e., $H_{\alpha_2}(X) \leq H_{\alpha_1}(X)$ for $0 < \alpha_1 \leq \alpha_2$. Properties d), e), and f) are obvious from the definition of $\mathbb{H}_\alpha(X, D, \epsilon)$.

## Appendix B
## Proof of Lemma 1

First, we show the left inequality of (29). The construction of the code gives

$$\ell(\psi(y_i)) \leq \log_2 i \tag{44}$$

for any $i \in \{1, 2, \dots |\widehat{\mathcal{Y}}|\}$. This inequality yields

$$2^{\rho\ell(\psi(y_i))} \leq 2^{\rho \log_2 i} = i^\rho, \tag{45}$$

which is the left inequality of (29).

Next, we show the right inequality of (29). From (19), we have

$$\left(\frac{P_{Y^\star}(y_1)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} \geq 1, \left(\frac{P_{Y^\star}(y_2)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} \geq 1, \dots, \left(\frac{P_{Y^\star}(y_{i-1})}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} \geq 1, \quad \left(\frac{P_{Y^\star}(y_i)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} = 1 \tag{46}$$

for any $i \in \{1, 2, \dots, |\widehat{\mathcal{Y}}|\}$. Hence, for any $i \in \{1, 2, \dots, |\widehat{\mathcal{Y}}|\}$, we have

$$i = \underbrace{1 + 1 + \cdots + 1}_{i} \tag{47}$$

$$\overset{(a)}{\leq} \left(\frac{P_{Y^\star}(y_1)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} + \left(\frac{P_{Y^\star}(y_2)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} + \cdots + \left(\frac{P_{Y^\star}(y_i)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}} \tag{48}$$

$$\overset{(b)}{\leq} \sum_{j=1}^{|\widehat{\mathcal{Y}}|} \left(\frac{P_{Y^\star}(y_j)}{P_{Y^\star}(y_i)}\right)^{\frac{1}{1+\rho}}, \tag{49}$$

where $(a)$ follows from (46) and $(b)$ is due to

$$\frac{P_{Y^\star}(y_j)}{P_{Y^\star}(y_i)} \geq 0 \quad (\forall i, j \in \{1, 2, \dots, |\widehat{\mathcal{Y}}|\}). \tag{50}$$

The inequality (49) yields the right inequality of (29).

## Appendix C
## Proof of Lemma 2

For each $y \in \widehat{\mathcal{Y}}$, let

$$\alpha(y) := \left[2^{\ell(\psi(y))}\right]^{-\frac{\rho}{1+\rho}} \tag{51}$$

$$\beta(y) := [P_Y(y)]^{\frac{1}{1+\rho}} \left[2^{\ell(\psi(y))}\right]^{\frac{\rho}{1+\rho}}. \tag{52}$$

Then, Hölder's inequality gives

$$\sum_{y\in\widehat{\mathcal{Y}}} \alpha(y)\beta(y) \leq \left(\sum_{y\in\widehat{\mathcal{Y}}} [\alpha(y)]^{\frac{1+\rho}{\rho}}\right)^{\frac{\rho}{1+\rho}} \left(\sum_{y\in\widehat{\mathcal{Y}}} [\beta(y)]^{1+\rho}\right)^{\frac{1}{1+\rho}}. \tag{53}$$

Taking logarithm of both sides of (53) and substituting (51) and (52) for (53), we obtain

$$\frac{1+\rho}{\rho} \log_2 \sum_{y\in\widehat{\mathcal{Y}}} [P_Y(y)]^{\frac{1}{1+\rho}} \leq \log_2 \sum_{y\in\widehat{\mathcal{Y}}} 2^{-\ell(\psi(y))} + \frac{1}{\rho} \log_2 \mathbb{E}[2^{\rho\ell(\psi(Y))}]. \tag{54}$$

Furthermore, noticing that the left-hand side of (54) is $H_{\frac{1}{1+\rho}}(Y)$, we obtain the desired result (34).

## APPENDIX D
## PROOF OF (38)

First, two quantities are defined:

*Definition 5 ([12]):* Given $D \geq 0$ and $\epsilon \in [0,1)$, the quantity $R_{D,\epsilon}(X)$ is defined as

$$R_{D,\epsilon}(X) := \inf_{\substack{P_{Y|X}: \\ \mathbb{P}[d(X,Y)>D] \leq \epsilon}} I(X;Y). \tag{55}$$

*Definition 6 ([15]):* Given $D \geq 0$ and $\epsilon \in [0,1)$, the $(D,\epsilon)$-entropy $H_{D,\epsilon}(X)$ is defined as

$$H_{D,\epsilon}(X) := \min_{\substack{f:\mathcal{X}\to\mathcal{Y}: \\ \mathbb{P}[d(X,f(X))>D] \leq \epsilon}} H(f(X)), \tag{56}$$

where $H(\cdot)$ denotes the Shannon entropy.

We have the following inequalities between these quantities and $\lim_{\rho\to 0} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon)$:

$$\lim_{\rho\to 0} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon) \geq R_{D,\epsilon}(X), \tag{57}$$

$$\lim_{\rho\to 0} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon) \leq H_{D,\epsilon}(X). \tag{58}$$

Indeed, (57) is proved as follows: as in the proof of Proposition 2, let $P_{Y^\star|X}$ be the conditional probability distribution that achieves the infimum of $\mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon)$. Then,

$$\begin{aligned}
\lim_{\rho\to 0} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon) &= \lim_{\rho\to 0} H_{\frac{1}{1+\rho}}(Y^\star) && (59)\\
&\overset{(a)}{=} H(Y^\star) && (60)\\
&\geq \inf_{\substack{P_{Y|X}: \\ \mathbb{P}[d(X,Y)>D] \leq \epsilon}} H(Y) && (61)\\
&\geq \inf_{\substack{P_{Y|X}: \\ \mathbb{P}[d(X,Y)>D] \leq \epsilon}} I(X;Y) = R_{D,\epsilon}(X), && (62)
\end{aligned}$$

where (a) is due to the fact that the Rényi entropy $H_{\frac{1}{1+\rho}}(\cdot)$ approaches the Shannon entropy $H(\cdot)$ as $\frac{1}{1+\rho} \to 1$. The proof of (58) is as follows: let $\widetilde{f}: \mathcal{X} \to \mathcal{Y}$ be the function that achieves the minimum of $H_{D,\epsilon}(X)$ and let $\widetilde{Y} := \widetilde{f}(X)$. Then,

$$\begin{aligned}
\lim_{\rho\to 0} \mathbb{H}_{\frac{1}{1+\rho}}(X,D,\epsilon) &\leq \lim_{\rho\to 0} H_{\frac{1}{1+\rho}}(\widetilde{Y}) \\
&= H(\widetilde{Y}) = H_{D,\epsilon}(X). && (63)
\end{aligned}$$

Furthermore, under assumptions 1) – 4), Kostina et al. [12] showed the next asymptotic result on $H_{D,\epsilon}(X^n)$ and $R_{D,\epsilon}(X^n)$ for any $\epsilon \in [0,1)$ (see [12, Section III-E]):

$$\begin{aligned}
R_{D,\epsilon}(X^n) =& H_{D,\epsilon}(X^n) \\
=& (1-\epsilon) n R(D,P_X) \\
&- \sqrt{\frac{nV(D,P_X)}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O(\log_2 n).
\end{aligned} \tag{64}$$

Now, (38) follows immediately from the equations (36), (57), (58), (64) and the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are finite sets.


## ACKNOWLEDGMENT

The authors thank the associate editor and reviewers for their helpful comments. In particular, we are very grateful to an anonymous reviewer that pointed out the error in the achievability proof of the early version of this paper.


## REFERENCES


[1] 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.

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

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

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

[5] T. A. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, pp. 2494–2498, June–July 2014.

[6] T. A. Courtade and S. Verdú, "Variable-length lossy compression and channel coding: non-asymptotic converses via cumulant generating functions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, pp. 2499–2503, June–July 2014.

[7] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 136–152, Jan. 2000.

[8] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory,"*IEEE Trans. Inf. Theory*,vol. 48, no. 8, pp. 2276–2290, Aug. 2002.

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

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

[11] V. Kostina, "Lossy data compression: nonasymptotic fundamental limits," Ph.D. dissertation, Princeton University, Sep. 2013.

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

[13] V. Kostina, "Data compression with low distortion and finite blocklength," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4268–4285, July 2017.

[14] S. Kuzuoka, "On the conditional smooth Rényi entropy and its applications in guessing and source coding," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1674–1690, Mar. 2020.

[15] E. C. Posner, E. R. Rodemich, and H. Rumsey, "Epsilon entropy of stochastic processes," *Ann. Math. Statist.*, vol. 38, no. 4, pp. 1000–1020, Aug. 1967.

[16] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkley Symposium on Mathematics, Statistics and Probability*, pp. 547–561, 1960.

[17] S. Saito and T. Matsushima, "Cumulant generating function of codeword lengths in variable-length lossy compression allowing positive excess distortion probability," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, Colorado, USA, pp. 881–885, June 2018.

[18] I. Sason and S. Verdú, "Improved bounds on lossless source coding and guessing moments via Rényi measures," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4323–4346, June 2018.

[19] I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression," *Entropy*, vol. 20, no. 12, paper 896, pp. 1–25, Nov. 2018.

**Shota Saito** received the B.E. degree, the M.E. degree, and the Ph.D. degree in applied mathematics from Waseda University, Tokyo, Japan, in 2013, 2015, and 2018, respectively.

From 2018 to 2021, he was an Assistant Professor at the Department of Applied Mathematics, Waseda University, Tokyo, Japan. From 2021, he is an Associate Professor at the Faculty of Informatics, Gunma University, Gunma, Japan. His research interests include information theory and its applications for machine learning and information-theoretic security.

Dr. Saito is a recipient of IEEE IT Society Japan Chapter Young Researcher Best Paper Award and Student Paper Award in the 2016 International Symposium on Information Theory and Its Applications. He also received Waseda University Azusa Ono Memorial Award (Academic) in 2016 and Symposium on Information Theory and its Applications Young Researcher Paper Award in 2018.

Dr. Saito is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Electrical and Electronics Engineers (IEEE).

**Toshiyasu Matsushima** received the B.E. degree, the M.E. degree, and the Ph.D. degree in Industrial Engineering and Management from Waseda University, Tokyo, Japan, in 1978, 1980, and 1991, respectively.

From 1980 to 1986, he joined NEC Corporation, Kanagawa, Japan. From 1989 to 1993, he was a Lecturer at the Department of Management Information, Yokohama College of Commerce. From 1993, he was an Associate Professor, and from 1996 to 2007 a Professor at the Department of Industrial and Management System Engineering, Waseda University. Since 2007, he has been a Professor at the Department of Applied Mathematics, Waseda University, and since 2017, he has been a Director at the Center for Data Science, Waseda University. From 2001 to 2002, he was a visiting researcher at the Department of Electrical Engineering, University of Hawaii, USA. From 2011 to 2012, he was a visiting scalar at the Department of Statistics, University of California, Berkeley, USA. His research interests are information theory, statistics, learning theory, and their applications.

Dr. Matsushima is a fellow of the Institute of Electronics, Information and Communication Engineers (IEICE) and a member of the Institute of Electrical and Electronics Engineers (IEEE), the Japan Society for Quality Control, and the Japan Society for Artificial Intelligence.