Variable-length Lossy Compression and Channel Coding: Non-asymptotic Converses via Cumulant Generating Functions

Thomas A. Courtade
EECS Department, University of California, Berkeley
Email: courtade@eecs.berkeley.edu

Sergio Verdú
Department of Electrical Engineering, Princeton University
Email: verdu@princeton.edu

***Abstract*—This paper gives non-asymptotic converse bounds on the cumulant generating function of the encoded lengths in variable-rate lossy compression and in variable-to-fixed channel coding. The results are given in terms of the Rényi mutual information and the d-tilted Rényi entropy. We also illustrate the application of the non-asymptotic bounds to obtain strong converses.**

I. Introduction

There have been several recent results on non-asymptotic bounds for the three major Shannon theoretic paradigms: fixed-length data transmission (e.g. [1]), fixed-length lossy compression (e.g. [2]), and variable-length lossless compression [3]. Those works give their bounds, not in terms of average quantities such as entropy and mutual information, but in terms of information spectra (i.e. the cumulative distribution functions of random variables such as information, information density and d-tilted information). The evaluation of those upper and lower bounds yields tight results for blocklengths as low as 100, and they have also been used to prove not only the conventional asymptotic fundamental limits but also to perform central-limit analyses based on the Berry-Esseén bound. In this paper, we take a different approach by studying the cumulant generating function (log moment generating function) of the random encoded lengths. The new non-asymptotic converse bounds are in terms of average quantities such as the Rényi mutual information, and are tight enough to recover the strong converse in the asymptotic regime and the large-deviation analysis above channel capacity. Together with achievability results (outside the scope of this paper) they may lead to alternative simple proofs of the dispersion analysis, as the authors have shown for lossless source coding in [4].

The normalized cumulant generating function was proposed by Campbell [5] as a proxy for average length in the design of prefix codes. Csiszár [6] also used this approach in the large deviations analysis of fixed-length channel coding and hypothesis testing. Arikan and Merhav [7] found the asymptotic moments of the number of guesses required to obtain an approximation with given distortion.

In this paper we deal with (a) the variable-length lossy compression paradigm in which the major available non-asymptotic converse is due to Kontoyiannis [8] (see also [9]), and (b) the variable-to-fixed channel coding paradigm introduced by Verdú and Shamai [10] for which they gave asymptotic results showing important gains over the conventional fixed-length setting for non-ergodic channels.

This work was supported by the NSF Center for Science of Information under grant agreement CCF-0939370.

II. Information measures

For $\alpha \geq 0$, $\alpha \neq 1$, and a discrete probability measure P_X, the *Rényi entropy* of order α is defined as[1]

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \mathbb{E}\left[\exp\{(1-\alpha)\, \imath_X(X)\}\right], \qquad (1)$$

where $\imath_X(x) = \log \frac{1}{P_X(x)}$ is the *information* with respect to P_X. Given $\alpha \geq 0$, $\alpha \neq 1$, and probability measures $P \ll Q$ defined on the same alphabet, the *Rényi divergence* of order α between distributions P and Q is defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \mathbb{E}\left[\exp\left\{\alpha\, \imath_{P\|Q}(Y)\right\}\right] \qquad (2)$$

$$= \frac{1}{\alpha-1} \log \mathbb{E}\left[\exp\left\{(\alpha-1)\, \imath_{Q\|P}(X)\right\}\right], \qquad (3)$$

where X and Y are distributed according to P and Q, respectively, and $\imath_{P\|Q}(x) = \log \frac{dP}{dQ}(x)$ is the *relative information*. For a conditional probability measure $P_{Y|X} \colon \mathcal{X} \mapsto \mathcal{Y}$, let $P_X \to P_{Y|X} \to P_Y$, with $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Following Sibson [11], we define the order-α *Rényi mutual information*

$$I_\alpha(X;Y) = \inf_{Q_Y} D_\alpha(P_{Y|X}P_X \| Q_Y P_X) \qquad (4)$$

$$= D_\alpha(P_{Y|X} \| P_{Y_\alpha} | P_X), \qquad (5)$$

where P_{Y_α} is implicitly defined by

$$\imath_{Y_\alpha\|Y}(y) = \frac{1}{\alpha} \log \mathbb{E}\left[\exp\left\{\alpha\, \imath_{X;Y}(X;y) - \kappa_\alpha\right\}\right]. \qquad (6)$$

In (6), the average is with respect to unconditional X, κ_α is chosen so that P_{Y_α} is a probability measure, and $\imath_{X;Y}(x;y) = \imath_{P_{Y|X=x}\|P_Y}(y)$ is the *information density*. We remark that the term $\sup_{P_X} I_\alpha(X;Y)$ has been called the *channel capacity of order* α by Csiszár [6].

By taking limits as $\alpha \to 1$ and applying l'Hôpital's rule, it is customary to identify $H_1(X) = H(X)$ and $D_1(P\|Q) = D(P\|Q)$, where $H(X)$ and $D(P\|Q)$ are the Shannon entropy and relative entropy, respectively.

[1] log denotes $\log_2$, and $\exp\{\cdot\}$ denotes $2^{(\cdot)}$.

For sets $\mathcal{S}, \mathcal{Z}$, a given probability measure P_S on $\mathcal{S}$, and a distortion measure $\mathsf{d}\colon \mathcal{S} \times \mathcal{Z} \to [0, +\infty)$, denote

$$\mathbb{R}_S(d) = \inf_{\substack{P_{Z|S}:\\ \mathbb{E}[\mathsf{d}(S,Z)] \le d}} I(S;Z) \tag{7}$$

which is equal to the rate-distortion function for a memoryless source with distribution P_S and separable distortion measure with per-letter distortion function d. We will assume throughout that the infimum in (7) is achieved by some $P^\star_{Z|S}$ satisfying $\mathbb{E}[\mathsf{d}(S, Z^\star)] = d$ (e.g., this is true when $\max\{|\mathcal{S}|, |\mathcal{Z}|\} < \infty$). To avoid degeneracy, we impose the mild condition that $\mathbb{R}_S(d)$ is finite for some $d < \infty$, and define $d_{\min} = \inf\{d\colon \mathbb{R}_S(d) < \infty\}$. For $d > d_{\min}$, and $s \in \mathcal{S}$, the d-*tilted information* is defined as [2]

$$\jmath_S(s, d) = \log \frac{1}{\mathbb{E}\left[\exp\{\lambda^\star d - \lambda^\star \mathsf{d}(s, Z^\star)\}\right]}, \tag{8}$$

where the expectation is with respect to the unconditional distribution on $\mathcal{Z}$ induced by $P_S \to P^\star_{Z|S} \to P^\star_Z$, and $\lambda^\star = -\mathbb{R}'_S(d)$. Analogous to (1), but exchanging information for d-tilted information, we define the d-*tilted Rényi entropy* of order α as

$$H_\alpha(S, d) = \frac{1}{1-\alpha} \log \mathbb{E}\left[\exp\{(1-\alpha)\, \jmath_S(S, d)\}\right], \tag{9}$$

where $S \sim P_S$. Applying l'Hôpital's rule and the fact that $\mathbb{E}[\jmath_S(S, d)] = \mathbb{R}_S(d)$ (see [2]), we find that $\lim_{\alpha \to 1} H_\alpha(S, d) = \mathbb{R}_S(d)$, yielding the lossy counterpart to $H_1(X) = H(X)$.

III. Preview of Results

We contrast the bounds in this paper to the converse bounds for variable-length lossless source coding. For any prefix code $\mathsf{f}\colon \mathcal{X} \to \{0,1\}^+$ (the set of all nonempty binary strings) and nonzero $t > -1$, Campbell [5] showed

$$\frac{1}{t} \log \mathbb{E}\left[2^{t\ell(\mathsf{f}(X))}\right] \ge H_{\frac{1}{1+t}}(X), \tag{10}$$

while lifting the prefix restriction (extraneous outside the domain of symbol-by-symbol encoding) results in [4]

$$\frac{1}{t} \log \mathbb{E}\left[2^{t\ell(\mathsf{f}(X))}\right] \ge H_{\frac{1}{1+t}}(X) - \log\log(1 + |\mathcal{X}|). \tag{11}$$

Our non-asymptotic converses are:

- In variable-length lossy source coding with source alphabet $\mathcal{S}$ and reproduction alphabet $\mathcal{Z}$, we show that any code f operating at distortion-level d must satisfy

$$\frac{1}{t} \log \mathbb{E}\left[2^{t\ell(\mathsf{f}(S))}\right] \ge H_{\frac{1}{1+t}}(S, d) - \log\log(1 + \min\{|\mathcal{S}|, |\mathcal{Z}|\}) \tag{12}$$

  for nonzero $t > -1$.
- Any variable-to-fixed channel code for $P_{Y|X}$ satisfies

$$\frac{1}{t} \log \mathbb{E}\left[2^{t\ell(S, \mathsf{f}(Y))}\right] \le \sup_{P_X} I_{\frac{1}{1-t}}(X;Y) + \log\left(\frac{1 + \log_e |\mathcal{S}|}{(2 - 2^t)^{1/t}}\right) \tag{13}$$

  for nonzero $t < 1$, where $\ell(S, \mathsf{f}(Y))$ measures the length of agreement between the source sequence S and the decoder output $\mathsf{f}(Y)$.

Though a formal presentation of these results must wait, note the beautiful parallelism among the three inequalities and the emergence of the fundamental information quantities $\imath_X(x)$, $\jmath_S(s, d)$, and $\imath_{X;Y}(x; y)$ in their respective settings through the Rényi entropy, d-tilted Rényi entropy, and Rényi mutual information. These non-asymptotic bounds yield simple proofs of strong converses and can simplify derivation of dispersion results (central limit regime) and the reliability function (large deviations) bounds, as shown in [4].

IV. Variable-Length Lossy Source Coding

Let $\mathcal{S}$ and $\mathcal{Z}$ be finite sets. Let $\{0,1\}^*$ be the set of all (possibly empty) binary strings. A *variable-length rate-distortion code* operating at distortion d is defined by a pair of functions

$$\mathsf{f}\colon \mathcal{S} \to \{0,1\}^* \qquad\qquad \mathsf{c}\colon \{0,1\}^* \to \mathcal{Z} \tag{14}$$

which satisfy $\mathsf{d}(s, \mathsf{c}(\mathsf{f}(s))) \le d$, for all $s \in \mathcal{S}$.

Theorem 1. *For $d > d_{\min}$ and nonzero $t > -1$, any rate-distortion code operating at distortion d satisfies* (12).

Proof: The proof follows immediately from (9) and Lemmas 3-4, which are given in the appendix. ■

Letting $t \to 0$ in Theorem 1 and applying l'Hôpital's rule, we recover the (weak) converse to the rate-distortion theorem: $\mathbb{E}[\ell(\mathsf{f}(S))] \ge \mathbb{R}_S(d) - \log\log(1 + \min\{|\mathcal{S}|, |\mathcal{Z}|\})$. In addition, Theorem 1 also allows us to deduce a non-asymptotic converse bound for fixed-length lossy compression with little effort.

Corollary 1. *Consider a fixed-blocklength code $\mathsf{f}\colon \mathcal{S} \to \{0,1\}^m$ and decoder $\mathsf{c}\colon \{0,1\}^m \to \mathcal{Z}$. For any $t \in (-1, 0)$,*

$$\begin{aligned}&\frac{1}{t} \log \frac{1}{\mathbb{P}\left[\mathsf{d}(S, \mathsf{c}(\mathsf{f}(S))) \le d\right]}\\ &\le H_{\frac{1}{1+t}}(S, d) - m - \log\log(1 + \min\{|\mathcal{S}|, |\mathcal{Z}|\}).\end{aligned} \tag{15}$$

Proof: From $\{\mathsf{f}, \mathsf{c}\}$, we construct a variable-length source code $\{\mathsf{f}', \mathsf{c}'\}$ which operates at distortion d as follows: If $\mathsf{d}(s, \mathsf{c}(\mathsf{f}(s))) \le d$, let $\mathsf{f}'(s) = \mathsf{f}(s)$ and $\mathsf{c}'(\mathsf{f}'(s)) = \mathsf{c}(\mathsf{f}(s))$. On the other hand, if $\mathsf{d}(s, \mathsf{c}(\mathsf{f}(s))) > d$, put $\mathsf{f}'(s) = s$ (using at most $\lceil \log |\mathcal{S}| \rceil$ bits), and $\mathsf{c}'(s) = \min_z \mathsf{d}(s, z)$. Then, a Chernoff bound

$$\mathbb{P}\left[\mathsf{d}(S, \mathsf{c}(\mathsf{f}(S))) \le d\right] \le \mathbb{E}\left[\mathbf{1}\{\ell(\mathsf{f}'(S)) \le m\}\right] \tag{16}$$

$$\le \mathbb{E}\left[\exp\{t(\ell(\mathsf{f}'(S)) - m)\}\right] \tag{17}$$

followed by Theorem 1 on $t \in (-1, 0)$ yields (15). ■

We remark that Theorem 1 also serves as a non-asymptotic converse for the lossy guesswork problem considered by Arikan and Merhav in [7].

An easy consequence of Corollary 1 is the following strong converse for memoryless sources:

Corollary 2. *For finite sets $\mathcal{S}, \mathcal{Z}$, let $P_{S^n} = P_S \times \cdots \times P_S$ be a product measure on $\mathcal{S}^n$, and let $\{\mathsf{f}_n, \mathsf{c}_n\}$ be blocklength-nR*

rate-distortion codes operating under the separable distortion measure $\mathsf{d}_n(s^n, z^n) = \frac{1}{n}\sum_{i=1}^n \mathsf{d}(s_i, z_i)$. *We have*

$$\frac{1}{n}\log \mathbb{P}\left[\mathsf{d_n}(S^n, \mathsf{c}_n(\mathsf{f}_n(S^n))) \leq d\right] \leq \inf_{t\in(-1,0)} t\left(H_{\frac{1}{1+t}}(S,d) - R - \frac{\log n}{n} + O\left(\frac{1}{n}\right)\right) \quad (18)$$

Proof: $\jmath_{S^n}(S^n, d) = \sum_{i=1}^n \jmath_S(S_i, d)$ implies $H_{\frac{1}{1+t}}(S^n, d) = nH_{\frac{1}{1+t}}(S, d)$. Thus, we apply Corollary 1 using independence of the S_i's to get the desired result. ■

To see that (18) is indeed a strong converse, recall that $\lim_{\alpha\to 1} H_\alpha(S, d) = \mathbb{R}_S(d)$. Therefore, if $R < \mathbb{R}_S(d)$, we can take t sufficiently close to 0 to infer the existence of a constant $c > 0$ such that the probability of faithful reproduction is exponentially small $\mathbb{P}[\mathsf{d_n}(S^n, \mathsf{c}_n(\mathsf{f}_n(S^n))) \leq nd] < 2^{-nc}$ for all n sufficiently large. Thus, taking the infimum of (18) over $t \in (0, 1)$ can only make the exponential decay more steep.

V. Variable-to-Fixed Channel Coding

A channel is defined by a space of inputs $\mathcal{X}$ and outputs $\mathcal{Y}$, and a conditional probability measure $P_{Y|X}\colon \mathcal{X} \mapsto \mathcal{Y}$. Let the message space $\mathcal{S} = \{0,1\}^m$ be equipped with the equiprobable probability measure $P_S(s) = 2^{-m}$ for all $s \in \mathcal{S}$. Corresponding to the notion of variable-rate channel capacity [10], a variable-to-fixed channel code for $(\mathcal{S}, P_{Y|X})$ is defined by encoding/decoding maps $\mathsf{c}\colon \mathcal{S} \to \mathcal{X}$ and $\mathsf{f}\colon \mathcal{Y} \to \mathcal{S}$.

Let $\mathsf{f}_i(y)$ denote the i-th coordinate of $\mathsf{f}(y)$. That is, if $\mathsf{f}(y) = (s_1, s_2, \ldots, s_m) \in \mathcal{S}$, then $\mathsf{f}_i(y) = s_i$. The quantity of interest in our setting is given by $\ell(s, \mathsf{f}(y)) = \min\{i : \mathsf{f}_i(y) \neq s_i\} - 1$ (i.e., the length of the initial segment on which s and $\mathsf{f}(y)$ agree). As we did for lossy source coding, we will bound the exponential moments of the random variable $\ell(S, \mathsf{f}(Y))$, where S, Y are dependent via $P_S \xrightarrow{\mathsf{c}} P_X \to P_{Y|X} \to P_Y$.

Theorem 2. *If* $\{\mathsf{c}, \mathsf{f}\}$ *is a variable-to-fixed channel code for* $(\mathcal{S}, P_{Y|X})$ *and* $t < 1$ *is nonzero, then* (13) *is satisfied.*

We remark that the term $-t^{-1}\log(2 - 2^t)$ is essential. Indeed, consider the degenerate channel $P_{Y|X} = P_Y$, in which case the divergence term is zero and $\ell(S, \mathsf{f}(Y))$ behaves like a geometric random variable with success probability $1/2$. Since the moment generating function of a geometric random variable $G \sim \text{Geometric}(1/2)$ is given by $\mathbb{E}[2^{tG}] = \frac{1}{2-2^t}$, the term $-t^{-1}\log(2 - 2^t)$ is necessary.

Letting $t \to 0$ in Theorem 2 and applying l'Hôpital's rule, we discover the (weak) converse

$$\mathbb{E}\Big[\ell(S, \mathsf{f}(Y))\Big] \leq \sup_{P_X} I(X;Y) + \log(1 + \log_e|\mathcal{S}|) \quad (19)$$

for variable-to-fixed channel coding which does not depend on the existence of a strong converse for the channel $P_{Y|X}$ as required in [10]. A direct application of Theorem 2 and a Chernoff bound yields

Corollary 3. *Consider any code* $\{\mathsf{c}, \mathsf{f}\}$ *for* $(\mathcal{S}, P_{Y|X})$. *For any* $t \in (0, 1)$

$$\frac{1}{t}\log\frac{1}{\mathbb{P}[S = \mathsf{f}(Y)]} \geq \log|\mathcal{S}| - \sup_{P_X} I_{\frac{1}{1-t}}(X;Y) - \log\left(\frac{1 + \log_e|\mathcal{S}|}{(2-2^t)^{1/t}}\right). \quad (20)$$

Specializing to the setting of a memoryless channel, we obtain the following strong converse by applying an identity of Gallager [12] and Arimoto [13] to Corollary 3.

Corollary 4. *Let* $\{\mathsf{c}_n, \mathsf{f}_n\}$ *be a sequence of codes for* $(\{0,1\}^{nR}, P_{Y^n|X^n})$, *respectively, where* $P_{Y^n|X^n=x^n}(y^n) = \prod_{i=1}^n P_{Y|X=x_i}(y_i)$, *and* $S \in \{0,1\}^{nR}$. *For any* $t \in (0, 1)$,

$$\frac{1}{tn}\log\frac{1}{\mathbb{P}[S = \mathsf{f}_n(Y^n)]} \geq R - \sup_{P_X} I_{\frac{1}{1-t}}(X;Y) + O\left(\frac{\log n}{n}\right). \quad (21)$$

To first order, the exponent in (21) is that given in [13], [14], known to be tight at rates above capacity [15].

Proof of Theorem 2: Although the approach applies to non-discrete alphabets, for simplicity, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are both finite. We first treat the case of $t \in (0, 1)$. To simplify notation, for each $y \in \mathcal{Y}$, let $W_{X|Y=y}$ be an unnormalized measure on $\mathcal{X}$ defined by

$$W_{X|Y=y}(x) = (2 - 2^t)P_X(x)\mathbb{E}\left[\exp\{t\,\ell(S, \mathsf{f}(y))\}|X = x\right]. \quad (22)$$

With this definition, we have the identity

$$\frac{1}{t}\log\mathbb{E}\left[\exp\{t\ell(S, \mathsf{f}(Y)\}\right] = \frac{1}{t}\log\left(\sum_{x,y} P_{Y|X}(y)W_{X|Y=y}(x)\right) + \frac{1}{t}\log\left(\frac{1}{2-2^t}\right). \quad (23)$$

For a given $y \in \mathcal{Y}$, let $x_{[1,y]} \succ_y x_{[2,y]} \succ_y \cdots \succ_y x_{[|\mathcal{X}|,y]}$ be an arbitrary strict order on $\mathcal{X}$ so that

$$P_{Y|X=x_{[1,y]}}(y) \geq \cdots \geq P_{Y|X=x_{[|\mathcal{X}|,y]}}(y). \quad (24)$$

Fix y and focus on the term

$$\mathbb{E}[W_{X|Y}(X|y)] = \sum_{j=1}^{|\mathcal{X}|} P_{Y|X}(y|x_{[j,y]})W_{X|Y}(x_{[j,y]}|y). \quad (25)$$

In view of Lemma 6 from the appendix, we upper bound (25) by considering the submodular optimization problem with variable $Q_{X|Y=y} \in \mathbb{R}^{\mathcal{X}}$ and $0 < t < 1$:

$$\text{maximize:} \quad \sum_{j=1}^{|\mathcal{X}|} P_{Y|X=x_{[j,y]}}(y)Q_{X|Y=y}(x_{[j,y]}) \quad (26)$$

$$\text{subject to:} \quad Q_{X|Y=y}(\mathcal{E}) \leq P_X^{1-t}(\mathcal{E}) \quad \text{for all } \mathcal{E} \subseteq \mathcal{X}, \quad (27)$$

It is well known that the solution to this optimization problem is given by (see [16]):

$$W^\star_{X|Y=y}(x_j) = P_X^{1-t}(\mathcal{A}_j(y) \cup \{x_j\}) - P_X^{1-t}(\mathcal{A}_j(y)), \quad (28)$$

where $\mathcal{A}_j(y) = \{x_{[1,y]}, x_{[2,y]}, \ldots, x_{[j-1,y]}\}$. Note that, by definition, we maintain the property that $W^\star_{X|Y=y}(\mathcal{E}) \le P_X^{1-t}(\mathcal{E})$ for all $\mathcal{E} \subseteq \mathcal{X}$ and

$$\sum_y \sum_x P_{Y|X=x}(y) W^\star_{X|Y=y}(x) \tag{29}$$

$$= \sum_y \sum_{j=1}^{|\mathcal{X}|} P_{Y|X=x_j}(y) \left(P_X^{1-t}(\{\mathcal{A}_j(y), x_j\}) - P_X^{1-t}(\mathcal{A}_j(y)) \right).$$

Defining $N_y(j) = |\{s \in \mathcal{S} : \mathsf{c}(s) \prec_y \mathsf{c}(s_j)\}| + 1$ it is straightforward to express the right side of (29) as

$$\sum_y \sum_{k=1}^{2^m} P_{Y|X}(y|\mathsf{c}(s_k)) P_S^{1-t}(s_k) \Big(N_y^{1-t}(k) - (N_y(k)-1)^{1-t} \Big)$$

$$\le \sum_y \sum_{k=1}^{2^m} P_{Y|X=\mathsf{c}(s_k)}(y) P_S^{1-t}(s_k) N_y^{-t}(k) \tag{30}$$

$$\le \sum_y \left(\sum_{k=1}^{2^m} P^{\frac{1}{1-t}}_{Y|X=\mathsf{c}(s_k)}(y) P_S(s_k) \right)^{1-t} \left(\sum_{k=1}^{2^m} N_y^{-1}(k) \right)^t \tag{31}$$

$$= \sum_y \left(\sum_x P^{\frac{1}{1-t}}_{Y|X=x}(y) P_X(x) \right)^{1-t} \left(\sum_{k=1}^{2^m} k^{-1} \right)^t \tag{32}$$

$$\le \sum_y \left(\sum_x P^{\frac{1}{1-t}}_{Y|X=x}(y) P_X(x) \right)^{1-t} \Big(\log_e(2^m) + 1 \Big)^t, \tag{33}$$

where (30) holds since $z^{1-t} - (z-1)^{1-t} \le z^{-t}$ for $t > 0$ and $z \ge 1$, (31) is Hölder's inequality, and (32) follows since $\{N_y(k)\}_{k=1}^{2^m} = \{1, 2, \ldots, 2^m\}$ for each y.

Recalling (23) and taking logarithms of (33), we find

$$\frac{1}{t} \log \mathbb{E}\left[\exp(t\ell(S, \mathsf{f}(Y))\right]$$

$$\le \frac{1}{t} \log \left(\sum_y \left(\sum_x P^{\frac{1}{1-t}}_{Y|X=x}(y) P_X(x) \right)^{1-t} \right)$$

$$+ \log\left(1 + \log_e |\mathcal{S}|\right) + \frac{1}{t} \log \left(\frac{1}{2 - 2^t} \right) \tag{34}$$

$$= \inf_{Q_Y} D_{\frac{1}{1-t}} \left(P_{Y|X} P_X \middle\| Q_Y P_X \right) + \log \left(\frac{1 + \log_e |\mathcal{S}|}{(2 - 2^t)^{1/t}} \right), \tag{35}$$

where (35) follows from Lemma 5 in the appendix. Supremizing over P_X only weakens the inequality; recalling (4) proves the theorem for $t \in (0, 1)$. The proof for $t < 0$ follows mutatis mutandis. ∎

Appendix

A. Fixed-to-Variable Lossy Source Coding Lemmas

The following result follows similarly to [4, Lemma 1]:

Lemma 1. *Let $\mathsf{f} \colon \mathcal{X} \to \{0,1\}^*$ be injective. Then*

$$\sum_{x \in \mathcal{X}} 2^{-\ell(\mathsf{f}(x))} \le \log(1 + |\mathcal{X}|). \tag{36}$$

Lemma 2. *[2] Fix $d > d_{\min}$. The following hold:*

1) *For $P_{Z^\star}$-almost every z*

$$\jmath_S(s, d) = \imath_{S;Z^\star}(s; z) + \lambda^\star \mathsf{d}(s, z) - \lambda^\star d. \tag{37}$$

2) *For all $z \in \mathcal{Z}$ and $S \sim P_S$,*

$$\mathbb{E}\left[\exp\left\{ \jmath_S(S, d) - \lambda^\star \mathsf{d}(S, z) + \lambda^\star d \right\} \right] \le 1. \tag{38}$$

Lemma 3. *Fix $d > d_{\min}$. For any nonzero $t > -1$, we have $P_{Z^\star}$-a.e.*

$$\frac{1+t}{t} \log \mathbb{E}\left[\exp\left\{ \frac{t}{1+t} \jmath_S(S, d) \right\} \right]$$

$$\le \log \mathbb{E}\left[\exp\left\{ \lambda^\star (\mathsf{d}(S, Z^\star) - d) - \ell(\mathsf{f}(S)) \right\} \middle| Z^\star \right]$$

$$+ \frac{1}{t} \log \mathbb{E}\left[\exp\left\{ t\, \ell(\mathsf{f}(S)) \right\} \right], \tag{39}$$

Proof: Denote

$$\alpha(s) = \Big(\exp\Big\{ -\imath_{S|Z^\star}(s|z) + \lambda^\star (\mathsf{d}(s, z) - d) - \ell(\mathsf{f}(s)) \Big\} \Big)^{\frac{t}{1+t}}$$

$$\beta(s) = \Big(\exp\Big\{ -\imath_S(s) + t\, \ell(\mathsf{f}(s)) \Big\} \Big)^{\frac{1}{1+t}}. \tag{40}$$

For $t > 0$, Hölder's inequality states

$$\sum_{s \in \mathcal{S}} \alpha(s)\beta(s) \le \left(\sum_{s \in \mathcal{S}} \alpha^{\frac{1+t}{t}}(s) \right)^{\frac{t}{1+t}} \left(\sum_{s \in \mathcal{S}} \beta^{1+t}(s) \right)^{\frac{1}{1+t}}. \tag{41}$$

Taking the logarithm of both sides, we find:

$$\log \mathbb{E}\left[\exp\left\{ \frac{t}{1+t} \Big(\imath_{S;Z^\star}(S; z) + \lambda^\star \mathsf{d}(S, z) - \lambda^\star d \Big) \right\} \right]$$

$$\le \frac{t}{1+t} \log \mathbb{E}\left[\exp\left\{ \lambda^\star (\mathsf{d}(S, Z^\star) - d) - \ell(\mathsf{f}(S)) \right\} \middle| Z^\star = z \right]$$

$$+ \frac{1}{1+t} \log \mathbb{E}\left[\exp\left\{ t\, \ell(\mathsf{f}(S)) \right\} \right]. \tag{42}$$

Multiplying both sides of (42) by $(1+t)/t$ and applying the first claim of Lemma 2 proves (39) for $t > 0$ as desired. If $-1 < t < 0$, Hölder's inequality is reversed. Nonetheless, running through the argument again, we arrive at the same inequality since multiplying through by $(1+t)/t$ reverses the inequality a second time. ∎

Lemma 4. *Fix $d > d_{\min}$ and let $\{\mathsf{f}, \mathsf{c}\}$ be a variable-length rate-distortion code operating at distortion d. For $P_{Z^\star}$-a.e.*

$$\mathbb{E}\left[\exp\left\{ \lambda^\star (\mathsf{d}(S, Z^\star) - d) - \ell(\mathsf{f}(S)) \right\} \middle| Z^\star \right]$$

$$\le \log(1 + \min\{|\mathcal{S}|, |\mathcal{Z}|\}), \tag{43}$$

Proof: For notational convenience, define $\hat{s} = \mathsf{c}(\mathsf{f}(s)) \in \mathcal{Z}$, and let $\bar{\mathcal{Z}} = \{\mathsf{c}(\mathsf{f}(s)) : s \in \mathcal{S}\} \subseteq \mathcal{Z}$. Assuming, without essential loss of generality that the decoder mapping is injective, define the probability measure Q_Z on $\bar{\mathcal{Z}}$ by

$$Q_Z(z) = \frac{1}{\kappa} 2^{-\ell(\mathsf{c}^{-1}(z))} \tag{44}$$

with $\kappa = \sum_{z \in \bar{\mathcal{Z}}} 2^{-\ell(\mathsf{c}^{-1}(z))}$.

By definition of $\{\mathsf{f},\mathsf{c}\}$, we have

$$2^{-\ell(\mathsf{f}(s))} = 2^{\lambda^\star(\mathsf{d}(s,\hat{s})-d)-\lambda^\star(\mathsf{d}(s,\hat{s})-d)-\ell(\mathsf{f}(s))} \tag{45}$$
$$\leq 2^{-\lambda^\star(\mathsf{d}(s,\hat{s})-d)-\ell(\mathsf{f}(s))} \tag{46}$$
$$= \kappa 2^{-\lambda^\star(\mathsf{d}(s,\hat{s})-d)} Q_Z(\hat{s}) \tag{47}$$

since $\lambda^\star(\mathsf{d}(s,\hat{s})-d) \leq 0$. Summing (47) over all $\hat{s}$ yields

$$2^{-\ell(\mathsf{f}(s))} \leq \kappa \mathbb{E}\left[\exp\left\{-\lambda^\star(\mathsf{d}(s,Z)-d)\right\}\right], \tag{48}$$

where $Z \sim Q_Z$. Next, let $(S, Z^\star, Z) \sim P_{SZ^\star} \times Q_Z$. Then,

$$\mathbb{E}\left[\exp\left\{\lambda^\star(\mathsf{d}(S,Z^\star)-d)-\ell(\mathsf{f}(S))\right\}|Z^\star = z\right]$$
$$= \mathbb{E}\left[\exp\left\{\imath_{S;Z^\star}(S;z)+\lambda^\star(\mathsf{d}(S,z)-d)-\ell(\mathsf{f}(S))\right\}\right] \tag{49}$$
$$= \mathbb{E}\left[\exp\left\{\jmath_S(S,d)-\ell(\mathsf{f}(S))\right\}\right] \tag{50}$$
$$\leq \kappa \mathbb{E}\left[\exp\left\{\jmath_S(S,d)-\lambda^\star(\mathsf{d}(S,Z)-d)\right\}\right] \tag{51}$$
$$\leq \kappa \tag{52}$$
$$\leq \log(1+\min\{|\mathcal{S}|,|\mathcal{Z}|\}), \tag{53}$$

where (49), (50), (51), (52), (53) follow from a change of measure, the first claim of Lemma 2, (48), the second part of Lemma 2, and Lemma 1, respectively ∎

B. Variable-to-Fixed Channel Coding Lemmas

Lemma 5. *[6], [11] Let $t < 1$ be nonzero. For given $P_X, P_{Y|X}$ and corresponding finite spaces $\mathcal{X}, \mathcal{Y}$,*

$$\inf_{Q_Y} D_{\frac{1}{1-t}}\left(P_{Y|X}P_X \middle\| Q_Y P_X\right)$$
$$= \frac{1}{t}\log\left(\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}} P^{\frac{1}{1-t}}_{Y|X=x}(y)P_X(x)\right)^{1-t}\right). \tag{54}$$

Lemma 6. *For all $\mathcal{E} \subseteq \mathcal{X}$,*

$$W_{X|Y=y}(\mathcal{E}) \begin{cases} \leq P_X^{1-t}(\mathcal{E}) & 0<t<1 \\ = P_X(\mathcal{E}) & t=0 \\ \geq P_X^{1-t}(\mathcal{E}) & t<0 \end{cases} \tag{55}$$

where $W_{X|Y}$ is defined in (22)*.*

Proof: It suffices to prove the claim when $\mathcal{E} = \{x\}$, the argument for general events $\mathcal{E} \subseteq \mathcal{X}$ with $W_{X|Y=y}(\mathcal{E}) > 0$ proceeds identically. To this end, first we assume $t \in (0,1)$ and we set out to show

$$\sum_{s:\mathsf{c}(s)=x} P_S(s)\exp\{t\ell(s,\mathsf{f}(y))\} \leq \frac{P_X^{1-t}(x)}{2-2^t}. \tag{56}$$

Fix $\mathsf{f}(y)$, and denote $\mathcal{S}_x = \{s \in \mathcal{S} : \mathsf{c}(s) = x\}$, and let integers k, Δ be such that

$$2^k \leq |\mathcal{S}_x| = 2^k + \Delta < 2^{k+1}. \tag{57}$$

Observe that the left side of (56) is maximized (over f) when the elements in $\{0,1\}^m$ matching the longest initial segments of $\mathsf{f}(y)$ are greedily assigned to $\mathcal{S}_x$. Though details are omitted due to space constraints, an analysis of this assignment yields

$$\frac{|\mathcal{S}_x|}{P_X(x)} \times \sum_{s:\mathsf{c}(s)=x} P_S(s)\exp\{t\ell(s,\mathsf{f}(y))\}$$
$$= |\mathcal{S}_x| \times \mathbb{E}\left[\exp\{t\ell(S,\mathsf{f}(y))\}\middle|\mathsf{c}(S)=x\right] \tag{58}$$
$$\leq \quad 2^{tm}\frac{2^{k(1-t)}+\Delta(2-2^t)2^{-t(k+1)}}{2-2^t}. \tag{59}$$

Since $P_X(x) = (2^k+\Delta)/2^m$, we have for $t \in (0,1)$

$$(2-2^t)P_X^{t-1}(x)\sum_{s:\mathsf{c}(s)=x} P_S(s)\exp\{t\ell(s,\mathsf{f}(y))\}$$
$$= (2-2^t)\frac{(2^k+\Delta)^t}{2^{tm}}\sum_{s:\mathsf{c}(s)=x} P_S(s)\exp\{t\ell(s,\mathsf{f}(y))\} \tag{60}$$
$$\leq \frac{2^{k(1-t)}+\Delta(2-2^t)2^{-t(k+1)}}{(2^k+\Delta)^{1-t}}. \tag{61}$$

If $t < 0$, then $\sum_s P_{S|X=x}(s)\exp\{t\ell(s,\mathsf{f}(y))\}$ is minimized by repeating the same argument, reversing the inequality (61). Therefore, to complete the proof, it suffices to verify that

$$t2^{k(1-t)} \leq t\left((2^k+\Delta)^{1-t} - \Delta(2-2^t)2^{-t(k+1)}\right) \tag{62}$$

for $t < 1$. To this end, note that equality holds at the endpoints $\Delta \in \{0, 2^k\}$, and that the right side of (62) is concave in Δ. ∎

References

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

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

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

[4] T. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," *2014 IEEE Int. Symposium on Information Theory*, July 2014.

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

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

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

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

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

[10] S. Verdú and S. Shamai, "Variable-rate channel capacity," *IEEE Trans. on Inf. Theory*, vol. 56, no. 6, pp. 2651–2667, 2010.

[11] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[12] R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. on Inf. Theory*, vol. 11, pp. 3–18, 1965.

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

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

[15] J. K. Omura, "A lower bounding method for channel and source coding probabilities," *Information and Control*, vol. 27, pp. 148–177, 1975.

[16] S. T. McCormick, "Submodular function minimization," *Handbooks in operations res. and management sci.*, vol. 12, pp. 321–391, 2005.