Cumulant Generating Function of Codeword Lengths in Optimal Lossless Compression

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 analyzes the distribution of the codeword lengths of the optimal lossless compression code without prefix constraints both in the non-asymptotic regime and in the asymptotic regime. The technique we use is based on upper and lower bounding the cumulant generating function of the optimum codeword lengths. In the context of prefix codes, the normalized version of this quantity was proposed by Campbell in 1965 as a generalized average length. We then use the one-shot bounds to analyze the large deviations (reliability function) and small deviations (normal approximation) of the asymptotic fundamental limit in the case of memoryless sources. In contrast to other approaches based on the method of types or the Berry-Esséen inequality, we are able to deal with sources with infinite alphabets.**

I. Introduction

In this paper, we study the fundamental limits of optimal variable-length lossless data compressors without imposing prefix constraints, which coincide with those of almost-lossless data compressors. Recently, Kontoyiannis and Verdú [1] gave non-asymptotic upper and lower bounds on the distribution of codeword length, which they went on to use, along with Berry-Esséen's inequality, to prove rigorously the Gaussian approximation put forward by Strassen [2] for memoryless sources. In this paper, we follow an alternative approach based on the normalized cumulant generating function of the codeword lengths in order to obtain non-asymptotic bounds. We then show how to use those bounds to obtain simple proofs for the asymptotic normality and the reliability function of memoryless sources allowing countable source alphabets.

L. L. Campbell [3], [4] proposed the normalized cumulative generating function of the codeword lengths as an alternative to average length as a design criterion for lossless data compression codes subject to prefix constraints. He was able to upper and lower bound the minimum "generalized average length" of a prefix code in terms of the Rényi entropy.

II. Rényi Entropy

For $\alpha \geq 0$, $\alpha \neq 1$, and a discrete probability measure P_X, the *Rényi entropy* of order α is defined as ($\log = \log_2$ and $\exp(a) = 2^a$ throughout)

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \mathbb{E}\left[2^{(1-\alpha)\imath_X(X)}\right], \tag{1}$$

where $\imath_X(x) \triangleq \log \frac{1}{P_X(x)}$ is the *information* in x (with respect to P_X). The mean and variance of $\imath_X(X)$ are the entropy $H(X)$ and varentropy $V(X)$, respectively. Customarily, $H_1(X) = H(X)$ (which coincides with $\lim_{\alpha\to 1} H_\alpha(X)$, provided $H_{\alpha_0}(X) < \infty$ for some $\alpha_0 < 1$). Furthermore, $H_\infty(X) = \min_{x\in\mathcal{X}} \imath_X(x)$.

For a discrete random variable $X \sim P_X$ and $\alpha \in (0,1)$, the *α-scaled version of X*, denoted X_α, is defined by the relation

$$\imath_{X_\alpha}(x) \triangleq \alpha\, \imath_X(x) + (1-\alpha) H_\alpha(X). \tag{2}$$

By rearranging (2) and taking expectations, we make note of the following identity which will be useful later

$$\alpha D(X_\alpha \| X) = (1-\alpha)\left(H(X) - H(X_\alpha)\right). \tag{3}$$

The Rényi entropy can be seen to be a reparametrized version of the cumulant generating function of the information random variable $\imath_X(X)$, denoted $\Lambda_{\imath_X(X)}(\cdot)$:

$$H_\alpha(X) = \frac{1}{1-\alpha} \Lambda_{\imath_X(X)}(1-\alpha) \tag{4}$$

Accordingly, it can be shown that if $\zeta_n(X)$ denotes the nth cumulant of $\imath_X(X)$ (in units of bits^n), then the following series expansion holds about $t = 0$

$$t H_{1-t}(X) = t H(X) + \log e \sum_{n=2}^{\infty} \frac{\zeta_n(X)}{n!} \left(\frac{t}{\log e}\right)^n \tag{5}$$

III. Optimal Source Code

A *lossless source code* is an injective function $\mathsf{f}\colon \mathcal{X} \to \{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, \ldots\}$. If $x \in \{0,1\}^*$, $\ell(x)$ denotes its length.

For a countable set $\mathcal{X}$ equipped with probability measure P_X, a *P_X-optimal source code* is a lossless source code that satisfies:

(i) $P_X(x) \geq P_X(x') \iff \ell(\mathsf{f}_X(x)) \leq \ell(\mathsf{f}_X(x'))$ for all $x, x' \in \mathcal{X}$.

(ii) If $\ell(\mathsf{f}_X(x')) = k \in \{0,1,2,\ldots\}$ for an $x' \in \mathcal{X}$, then $|\{x \in \mathcal{X}\colon \ell(\mathsf{f}_X(x)) < k\}| = 2^k - 1$.

Note that a P_X-optimal source code is not unique. Indeed, swapping codewords of the same length preserves P_X-optimality. Nonetheless, the distribution of the length is the same for any choice of the optimal source code. With this in mind, it is convenient to adopt the code $\mathsf{f}_X^\star$ that assigns the lexicographically ordered strings in $\{0,1\}^*$ to $\mathcal{X}$ ordered in descending probabilities under P_X.

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

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

Convention 1. *We assume that $\mathcal{X} = \{x_1, x_2, x_3, \dots\}$ and P_X satisfies*

$$P_X(x_i) \geq P_X(x_j) \Longleftrightarrow i \leq j. \tag{6}$$

It is easy to show that [1]

$$\ell\left(\mathsf{f}^\star_X(x_k)\right) = \lfloor \log k \rfloor. \tag{7}$$

The normalized cumulant generating function of the lengths of the optimal code satisfies the following bounds:

Theorem 1. *For nonzero $t > -1$,*

$$H_{\frac{1}{1+t}}(X) - \log\log(1 + |\mathcal{X}|) \leq \frac{1}{t}\log \mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_X(X))}\right] \tag{8}$$

$$\leq H_{\frac{1}{1+t}}(X) \tag{9}$$

If $t \leq -1$

$$H_\infty(X) - \log\log(1 + |\mathcal{X}|) \leq -\log \mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_X(X))}\right] \tag{10}$$

$$\leq H_\infty(X) \tag{11}$$

In contrast, for prefix codes, Campbell [3] showed that for nonzero $t > -1$, the normalized cumulant generating function belongs to $[H_{\frac{1}{1+t}}(X), H_{\frac{1}{1+t}}(X) + 1)$. Bounds for $t \leq -1$ were outside the scope of [3], but have appeared in the context of guesswork (e.g., [5, Lemma 1]). To be precise, we note that guesswork effort under optimal guessing is in exact correspondence with the codeword lengths given in (7). Indeed, Arikan rediscovered Campell's bounds roughly three decades later in the context of guessing in [6]. The relationship was later discussed by Arikan and Merhav [7], and Hanawal and Sundaresan [8].

In light of these connections, we remark that Theorem 1 has appeared in various forms dating back to Campbell's original 1965 paper. Our goal in the present paper is pedagogical in nature. Using Theorem 1, we give a self-contained and rigorous derivation of Strassen's Gaussian codeword-length approximation, and also recover the source reliability function for lossless compression. In both cases, we do not restrict ourselves to finite alphabets, which distinguishes the present paper from prior work on lossless compression (e.g,. [1] and references therein), and the studies on guesswork discussed above.

IV. Applications of Theorem 1

We now proceed toward demonstrating how Theorem 1 can be used as a key tool for establishing two of the fundamental results for lossless source coding: the source reliability function and the asymptotic normality of codeword lengths under optimal encoding. For both results, Theorem 1 contains all the information theoretic ideas, while the remainder of the effort consists of invoking standard limiting results. The proofs can be found in Section V.

In order to simplify the presentation, we restrict our attention to memoryless sources in this paper. Thanks to the non-asymptotic nature of Theorem 1, with modest effort, our arguments can be modified to handle more general sources under suitable constraints (e.g., finite-state Markov sources).

A. Reliability function

For a countable set $\mathcal{X}$ and a sequence of distributions P_{X^n} on $\mathcal{X}^n$, we define the source reliability function

$$E(R) \triangleq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left\{\ell\left(\mathsf{f}^\star_{X^n}(X^n)\right) > nR\right\}}. \tag{12}$$

In other words, the reliability function $E(R)$ characterizes the large-deviations behavior of codeword lengths for the optimal encoder $\mathsf{f}^\star_{X^n}$. In the equivalent problem of almost-lossless fixed-length data compression the following result was announced by Shannon [9] in 1956 for finite-alphabet sources.

Theorem 2. *Let P_X be a discrete probability distribution on a countable set $\mathcal{X}$ with $H(X) < \infty$. For $H(X) < R < \log|\mathcal{X}|$, the reliability function for the memoryless source with distribution P_X is given parametrically by*

$$E(R) = D(X_\alpha \| X) \quad R = H(X_\alpha) \quad \textit{for } \alpha \in (0, 1). \tag{13}$$

It is important to note that the reliability function for lossless source coding is traditionally established using the method of types, and hence the source alphabet is typically assumed to be finite (e.g., [10, Chapter 2]). Following our approach, the source alphabet can be countably infinite, and the only hypothesis required is that $H(X) < \infty$ which is necessary for the reliability function to be meaningful. Therefore, within the context of memoryless sources, Theorem 2 has full generality.

We appeal to the Gärtner-Ellis Theorem in our proof of Theorem 2. This is the same approach employed in the large deviations analyses of guesswork (e.g., [5], [8]) under a finite-alphabet assumption. There, the established large deviations principle was shown to hold for a general class of finite-alphabet sources under modest assumptions [5], [8].

B. Asymptotic normality of codeword lengths

Let $\mathcal{X}$ be a countable set. For a sequence of distributions P_{X^n} on $\mathcal{X}^n$, we define

$$R^\star(n, \epsilon) \triangleq \inf\left\{R \colon \mathbb{P}\left\{\ell\left(\mathsf{f}^\star_{X^n}(X^n)\right) > nR\right\} \leq \epsilon\right\}. \tag{14}$$

In words, $R^\star(n, \epsilon)$ is smallest R for which the best code (with respect to source distribution P_{X^n}) exceeds rate R with probability no larger than ϵ. As with the reliability function, Theorem 1 provides an effective tool for characterizing $R^\star(n, \epsilon)$, even in the setting where $\mathcal{X}$ is countably infinite. The finite-alphabet version of Theorem 3 was given in [1] following a more cumbersome approach based on the Berry-Esséen non-asymptotic bound.

Theorem 3. *Let $\mathcal{X}$ be a countable set equipped with probability measure P_X which satisfies $H(X) < \infty$, $0 < V(X) < \infty$, and, under the assumptions of Convention 1,*

$$\sum_{k=2^n}^{\infty} P_X(x_k) \log \frac{1}{P_X(x_k)} = o\left(\frac{1}{\sqrt{n}}\right). \tag{15}$$

For the discrete memoryless source with distribution P_X,

$$R^\star(n, \epsilon) = H(X) + \sqrt{\frac{V(X)}{n}} Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right). \tag{16}$$

We remark that the constraint (15), which is effectively a constraint on the tail probabilities of P_X, can be significantly weakened at the expense of a less concise form.

V. Proofs

A. Proof of Theorem 1

In light of our comments following Theorem 1, we remark that various proofs have appeared previously (e.g., [3]–[8]). Nonetheless, the proof is elementary and since our goal is to give a self-contained treatment of lossless compression, we provide a short proof here for completeness.

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

$$\sum_{x\in\mathcal{X}} 2^{-\ell(\mathsf{f}(x))} \le \log(1+|\mathcal{X}|). \tag{17}$$

Proof. The claim follows by greedily assigning 2^ℓ elements of $\mathcal{X}$ to codewords of length ℓ, with $\ell = 0,\ldots,n-1$ and assigning the remaining $|\mathcal{X}| - 2^n + 1 < 2^n$ elements to codewords of length $n = \lfloor \log_2(1+|\mathcal{X}|)\rfloor$. In such case,

$$\sum_{a\in\mathcal{X}} 2^{-\ell(\mathsf{f}(a))} = (|\mathcal{X}| - 2^n + 1)2^{-n} + \sum_{\ell=0}^{n-1} 2^\ell 2^{-\ell} \tag{18}$$

$$= \log_2(1+|\mathcal{X}|) - (1+\Delta - 2^\Delta) \tag{19}$$

$$\le \log_2(1+|\mathcal{X}|) \tag{20}$$

where $\Delta = \log_2(1+|\mathcal{X}|) - \lfloor\log_2(1+|\mathcal{X}|)\rfloor \in [0,1)$. □

We can now prove Theorem 1 by considering two cases:

1) $t \le -1$: Since the most likely element of $\mathcal{X}$ is mapped to the null string under $\mathsf{f}_X^\star$, $\log\mathbb{E}\left[2^{t\ell(\mathsf{f}_X^\star(X))}\right] \ge \log\left[\max_{x\in\mathcal{X}} P_X(x)\right]$ which is (11). Now, assume $\mathcal{X}$ is finite, and note that we also have

$$\log\mathbb{E}\left[2^{t\ell(\mathsf{f}_X^\star(X))}\right] \le \log\mathbb{E}\left[2^{-\ell(\mathsf{f}_X^\star(X))}\right] \tag{21}$$

$$\le \log\left[\max_{x\in\mathcal{X}} P_X(x)\sum_{x\in\mathcal{X}} 2^{-\ell(\mathsf{f}_X^\star(x))}\right] \tag{22}$$

$$\le -H_\infty(X) + \log\log(1+|\mathcal{X}|), \tag{23}$$

where (23) follows from Lemma 1. (See also [5, Lemma 1].)

2) $t > -1$: The remainder of the proof follows Campbell's original argument [3]. Inequality (8) follows by Lemma 1 and choosing $f(x) := 2^{\ell(\mathsf{f}(x))}$ in the following:

Lemma 2. *Let $f\colon \mathcal{X} \to [0,\infty)$. For nonzero $t > -1$,*

$$\frac{1}{t}\log\mathbb{E}\left[f^t(X)\right] \ge H_{\frac{1}{1+t}}(X) - \log\sum_{x\in\mathcal{X}}\frac{1}{f(x)}. \tag{24}$$

Proof. Set $\alpha(x) = f^{-\frac{t}{1+t}}(x)$, $\beta(x) = P_X^{\frac{1}{1+t}}(x) f^{\frac{t}{1+t}}(x)$. The claim for $t > 0$ is proved by invoking Hölder's inequality

$$\sum_{x\in\mathcal{X}}\alpha(x)\beta(x) \le \left(\sum_{x\in\mathcal{X}}\alpha^{\frac{1+t}{t}}(x)\right)^{\frac{t}{1+t}}\left(\sum_{x\in\mathcal{X}}\beta^{1+t}(x)\right)^{\frac{1}{1+t}},$$

and the reverse Hölder inequality for $-1 < t < 0$. □

To show (9), recall from (7) that $\mathsf{f}_X^\star$ satisfies $\ell(\mathsf{f}_X^\star(x_k)) \le \log k$. Hence for $t > 0$, we have the Chernoff bound

$$2^{t\ell(\mathsf{f}_X^\star(x_k))} \le k^t \le \left[\sum_{x'\in\mathcal{X}}\left(\frac{P_X(x')}{P_X(x_k)}\right)^{1/(1+t)}\right]^t, \tag{25}$$

from which it easily follows that

$$\mathbb{E}\left[2^{t\ell(\mathsf{f}_X^\star(X))}\right] \le \sum_{x\in\mathcal{X}} P_X(x)\left[\sum_{x'\in\mathcal{X}}\left(\frac{P_X(x')}{P_X(x)}\right)^{1/(1+t)}\right]^t$$

$$= \left(\sum_{x\in\mathcal{X}}\left(P_X(x)\right)^{1/(1+t)}\right)^{1+t}. \tag{26}$$

Taking logarithms and dividing through by t gives the desired result. On the other hand, if $t \in (-1,0)$, then

$$2^{t\ell(\mathsf{f}_X^\star(x_k))} \ge k^t \ge \left[\sum_{x'\in\mathcal{X}}\left(\frac{P_X(x')}{P_X(x_k)}\right)^{1/(1+t)}\right]^t, \tag{27}$$

and therefore $\log\mathbb{E}\left[2^{t\ell(\mathsf{f}_X^\star(X))}\right] \ge tH_{\frac{1}{1+t}}(X)$. Dividing through by t yields (9).

B. Proof of Theorem 2

To handle the case when $|\mathcal{X}| = \infty$, we invoke Lemmas 4 and 5 from Appendix A. We define the monotonically non-decreasing function

$$\Lambda(t) \triangleq \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}_{X^n}^\star(X^n))}\right] \tag{28}$$

$$= \begin{cases} tH_{\frac{1}{1+t}}(X) & t > -1 \\ -H_\infty(X) & t \le -1, \end{cases} \tag{29}$$

where (29) follows from Theorem 1 and Lemma 4. Furthermore, we denote the critical threshold

$$t_c \triangleq \sup_{t>-1}\{t\colon \Lambda(t) < \infty\}, \tag{30}$$

which is finite only if $|\mathcal{X}| = \infty$. Invoking Lemma 5 if necessary (i.e., if $|\mathcal{X}| = \infty$), we find that $\Lambda(t)$ is differentiable on the interval $(-1, t_c)$ containing the origin, with derivative given by

$$\Lambda'(t) = H\left(X_{\frac{1}{1+t}}\right) \quad \text{for } -1 < t < t_c. \tag{31}$$

This implies

$$\mathcal{G} = \{\Lambda'(t)\colon -1 < t < t_c\} = (H(X_\infty), \log|\mathcal{X}|) \tag{32}$$

which holds regardless of whether $|\mathcal{X}| = \infty$.

Next, define the Fenchel-Legendre transform of $\Lambda(\cdot)$ by

$$\Lambda^*(R) = \sup_{t\in\mathbb{R}}\Big\{tR - \Lambda(t)\Big\} = \sup_{t>-1}\Big\{tR - \Lambda(t)\Big\}. \tag{33}$$

Recalling (3), and applying the identity (31) once more combined with convexity of $\Lambda^*(\cdot)$, we find that

$$\Lambda^*(R) = D\left(X_{\frac{1}{1+t^*}}\|X\right), \tag{34}$$

where $t^* \in (-1, t_c)$ attains the supremum in (33) and satisfies $R = H\left(X_{\frac{1}{1+t^*}}\right)$. By letting $t^* \in (-1,0)$, we sweep the whole interval $H(X) < R < \log|\mathcal{X}|$.

For any $H(X) < R < \log|\mathcal{X}|$, we apply the Gärtner-Ellis Theorem (see Theorem 4 in Appendix B) to the sequence of random variables $Z_n \triangleq \frac{1}{n}\ell\left(\mathsf{f}_{X^n}^\star(X^n)\right)$ to obtain:

$$\begin{aligned} -\inf_{x\geq R}\Lambda^*(x) &\geq \limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left[\frac{1}{n}\ell\left(\mathsf{f}^\star_{X^n}(X^n)\right)\geq R\right] && (35)\\ &\geq \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left[\frac{1}{n}\ell\left(\mathsf{f}^\star_{X^n}(X^n)\right)> R\right] && (36)\\ &\geq -\inf_{x\in\mathcal{G}\cap(R,\infty)}\Lambda^*(x) = -\inf_{x\geq R}\Lambda^*(x), && (37)\end{aligned}$$

where (37) follows from the fact that $\mathcal{G}\cap(R,\infty) = (R,\log|\mathcal{X}|)$ and the continuity of $\Lambda^*(\cdot)$. Since $\Lambda^*(\cdot)$ is monotone increasing, we can conclude that

$$\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\left[\frac{1}{n}\ell\left(\mathsf{f}^\star_{X^n}(X^n)\right)> R\right] = \Lambda^*(R). \qquad (38)$$

Recalling (34) and reparameterizing in terms of $\alpha = \frac{1}{1+t^*}$ completes the proof.

C. Proof of Theorem 3

Assume first that $|\mathcal{X}|<\infty$. Our proof consists of showing the following convergence in distribution

$$\frac{\ell(\mathsf{f}^\star_{X^n}(X^n)) - H(X^n)}{\sqrt{V(X^n)}} \xrightarrow{D} \mathcal{N}(0,1). \qquad (39)$$

To that end, Curtiss' Theorem [11] asserts that it suffices to show pointwise convergence of the moment generating function

$$\lim_{n\to\infty}\mathbb{E}\left[\exp_e\left\{\lambda\left(\frac{\ell(\mathsf{f}^\star_{X^n}(X^n)) - H(X^n)}{\sqrt{V(X^n)}}\right)\right\}\right] = e^{\lambda^2/2},$$

where $\exp_e\{x\}\triangleq e^x$, and convergence need only occur in a neighborhood of the origin. For notational convenience, define

$$t_n = \frac{\lambda\log e}{\sqrt{V(X^n)}}, \qquad (40)$$

in view of the cumulant expansion (5), and the fact that $\zeta_2(X^n) = V(X^n) = nV(X)$, we obtain

$$\begin{aligned} t_n H_{\frac{1}{1+t_n}}(X^n) - t_n H(X^n) &= \frac{V(X^n)}{2\log e}\frac{t_n^2}{1+t_n} + O\left(\frac{1}{\sqrt{n}}\right)\\ &= \frac{\lambda^2\log e}{2} + O\left(\frac{1}{\sqrt{n}}\right). \qquad (41)\end{aligned}$$

Invoking Theorem 1 with $\mathcal{X}\leftarrow\mathcal{X}^n$, $P_X\leftarrow P_{X^n}$ and $t\leftarrow t_n$ establishes the desired convergence and establishes the proof assuming $|\mathcal{X}|<\infty$.

Now, assume $\mathcal{X}$ is countably infinite. We will only treat the case where $\lambda\geq 0$ (the case where $\lambda\leq 0$ follows similarly). Since the upper bound (9) does not involve $|\mathcal{X}|$, the previous argument also applies to yield the pointwise bound for $\lambda\geq 0$:

$$\limsup_{n\to\infty}\mathbb{E}\left[\exp_e\left\{\lambda\left(\frac{\ell(\mathsf{f}^\star_{X^n}(X^n)) - H(X^n)}{\sqrt{V(X^n)}}\right)\right\}\right] \leq e^{\lambda^2/2}.$$

Define $\mathcal{Y} = \{x_1,\dots,x_{k_n}\}$, and let Y be equal to X conditioned on the event $X\in\mathcal{Y}$. Since we have assumed $\lambda\geq 0$, Lemma 3, Theorem 1, and (5) together imply

$$\begin{aligned} &\log\mathbb{E}\left[\exp_e\left\{\lambda\left(\frac{\ell(\mathsf{f}^\star_{X^n}(X^n)) - H(X^n)}{\sqrt{V(X^n)}}\right)\right\}\right]\\ &\geq \log\mathbb{E}\left[\exp_e\left\{\lambda\left(\frac{\ell(\mathsf{f}^\star_{Y^n}(Y^n)) - H(X^n)}{\sqrt{V(X^n)}}\right)\right\}\right] && (42)\\ &\geq t_n H_{\frac{1}{1+t_n}}(Y^n) - t_n H(X^n) - O\left(\frac{\log(n\log k_n)}{\sqrt{n}}\right) && (43)\\ &= \frac{\lambda^2\log e}{2}\frac{V(Y)}{V(X)} + O\left(\frac{\log(n\log k_n)}{\sqrt{n}}\right)\\ &\quad + O\left(\sqrt{n}(H(Y)-H(X))\right). && (44)\end{aligned}$$

If we choose the sequence $k_n = 2^n$, $n=1,2,\dots$, then the tail bound (15) implies $\lim_{n\to\infty}\sqrt{n}(H(Y)-H(X)) = 0$, and we obtain the desired convergence.

Appendix A
Technical Lemmas $|\mathcal{X}|=\infty$

When $|\mathcal{X}|=\infty$, the proof of Theorem 2 uses the following technical lemmas.

Lemma 3. *Let P_X be a discrete probability distribution on $\mathcal{X}$, and let P_Y be the discrete probability distribution on the finite subset $\mathcal{Y} = \{x_1,x_2,\dots,x_k\}\subseteq\mathcal{X}$ defined by*

$$P_Y(y) = \mathbb{P}\left[X=y|X\in\mathcal{Y}\right]. \qquad (45)$$

Then,

$$\frac{1}{t}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_X(X))}\right] \geq \frac{1}{t}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_Y(Y))}\right] \quad \textit{for } t\neq 0. \qquad (46)$$

Proof. By Convention 1, $\mathsf{f}^\star_Y(y) = \mathsf{f}^\star_X(y)$ for all $y\in\mathcal{Y}$. Since $\mathcal{Y}$ consists of the k most likely elements of $\mathcal{X}$, the claim is immediate. □

Lemma 4. *Let P_X be a discrete probability measure on $\mathcal{X}$ with $H(X)<\infty$, and let $X^n\sim P_X\times\cdots P_X$. If $t>-1$ is nonzero,*

$$\liminf_{n\to\infty}\frac{1}{nt}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_{X^n}(X^n))}\right] \geq H_{\frac{1}{1+t}}(X). \qquad (47)$$

On the other hand, if $t\leq -1$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_{X^n}(X^n))}\right] \geq H_\infty(X). \qquad (48)$$

Proof. We assume $|\mathcal{X}|=\infty$ since otherwise the claim is a direct application of Theorem 1. Consider first the case $t>-1$. For each n, define the truncated alphabet $\mathcal{X}_{k_n} = \{x_1,x_2,\dots,x_{k_n}\}\subset\mathcal{X}$, and let X_{k_n} be the random variable X conditioned on the event $X\in\mathcal{X}_{k_n}$. For a given n, define the random variable $Y^n\in(\mathcal{X}_{k_n})^n$ according to

$$P_{Y^n}(y^n) = \mathbb{P}\left[X^n = y^n|X^n\in(\mathcal{X}_{k_n})^n\right]. \qquad (49)$$

Noting that Y^n consists of n i.i.d. copies of X_{k_n} and satisfies the conditions of Lemma 3,

$$\begin{aligned} \frac{1}{nt}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_{X^n}(X^n))}\right] &\geq \frac{1}{nt}\log\mathbb{E}\left[2^{t\ell(\mathsf{f}^\star_{Y^n}(Y^n))}\right] && (50)\\ &\geq \frac{1}{n}H_{\frac{1}{1+t}}(Y^n) - \frac{\log(n\log(1+k_n))}{n} && (51)\\ &= H_{\frac{1}{1+t}}(X_{k_n}) - \frac{\log(n\log(1+k_n))}{n}, && (52)\end{aligned}$$

where (51) follows from (8). Taking $k_n = n$, for example, it is easy to see that $X_{k_n} \xrightarrow{TV} X$. Using the lower-semicontinuity property of Rényi entropy [12, Theorem 5], we find $\liminf_{n\to\infty} H_{\frac{1}{1+t}}(X_{k_n}) \geq H_{\frac{1}{1+t}}(X)$, which proves the claim for $t > -1$. If $t \leq -1$, the proof proceeds in a similar manner, except we invoke (10) instead of (8). □

Lemma 5. *Define $\tau_0 \triangleq \sup_{t>-1}\{t\colon H_{\frac{1}{1+t}}(X) < \infty\}$, assuming $H(X) < \infty$. The function $tH_{\frac{1}{1+t}}(X)$ is finite and differentiable on the interval $t \in (-1, \tau_0)$ with derivative*

$$\frac{d}{dt}\left\{tH_{\frac{1}{1+t}}(X)\right\} = H\left(X_{\frac{1}{1+t}}\right). \tag{53}$$

Proof. First, we remark that [12, Theorem 1] implies that $\tau_0 > 0$. Consider any interval $[a, b]$ with $-1 < a < b < \tau_0$. Let $X^{(n)}$ be defined by the distribution

$$P_{X^{(n)}} = \left(P_X(x_1), P_X(x_2), \ldots, P_X(x_{n-1}), \sum_{k=n}^{\infty} P_X(x_k)\right).$$

Note that $tH_{\frac{1}{1+t}}(X^{(n)})$ converges uniformly to $tH_{\frac{1}{1+t}}(X) < \infty$ on $[a, b]$, due to uniform convergence of the series

$$\sum_{k=1}^{\infty} P_X^{\frac{1}{1+t}}(x_k) \tag{54}$$

since $\sum_{k=1}^{\infty} P_X^{\frac{1}{1+t}}(x_k) < \sum_{k=1}^{\infty} P_X^{\frac{1}{1+\tau_0}}(x_k) < \infty$ for $t < b < \tau_0$. Since $X^{(n)}$ has finite support, we obtain the identity

$$\frac{d}{dt}\left\{t\,H_{\frac{1}{1+t}}\left(X^{(n)}\right)\right\} = H\left(X^{(n)}_{\frac{1}{1+t}}\right). \tag{55}$$

Note that (2) implies

$$H\left(X_{\frac{1}{1+t}}\right) = \frac{1}{1+t}H(X) - \frac{t}{1+t}H_{\frac{1}{1+t}}(X) < \infty \quad \text{on } [a, b],$$

where finiteness follows by our assumption that $H(X) < \infty$ and monotonicity of Rényi entropy. Therefore,

$$\begin{aligned}
&\frac{d}{dt}\Big\{tH_{\frac{1}{1+t}}\left(X^{(n)}\right)\Big\} - H\left(X_{\frac{1}{1+t}}\right) && (56)\\
&= H\left(X^{(n)}_{\frac{1}{1+t}}\right) - H\left(X_{\frac{1}{1+t}}\right) && (57)\\
&= \frac{1}{1+t}\Big(H(X^{(n)}) - H(X)\Big)\\
&\quad + \frac{t}{1+t}\Big(H_{\frac{1}{1+t}}(X) - H_{\frac{1}{1+t}}(X^{(n)})\Big), && (58)
\end{aligned}$$

which converges uniformly to zero on $[a, b]$ (again, due to uniform convergence of (54)). Thus, $tH_{\frac{1}{1+t}}(X)$ is differentiable on $[a, b]$, and the derivative is given by (53) (cf. [13]). □

Appendix B
The Gärtner-Ellis Theorem

Here, we quote [14, Theorem 2.3.6], adapted slightly for our purposes. To this end, consider a sequence of random variables $Z_n \in \mathbb{R}$, where Z_n possesses the law μ_n and has cumulant generating function

$$\Lambda_n(\lambda) \triangleq \log \mathbb{E}\left[2^{\lambda Z_n}\right]. \tag{59}$$

Assume the limit $\Lambda(\lambda) \triangleq \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda)$ exists as an extended real number and that the origin belongs to the interior of $\mathcal{D}_\Lambda \triangleq \{\lambda \in \mathbb{R}\colon \Lambda(\lambda) < \infty\}$. Let

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}}\Big\{\lambda x - \Lambda(\lambda)\Big\} \tag{60}$$

be the Fenchel-Legendre transform of $\Lambda(\cdot)$. Let $\mathcal{D} \subseteq \mathcal{D}_\Lambda$ be any open set on which $\Lambda(\cdot)$ is differentiable, and put $\mathcal{G} = \{\Lambda'(\lambda)\colon \lambda \in \mathcal{D}\}$.

Theorem 4 (Gärtner-Ellis)**.** *Under the above assumptions, the following hold:*

(a) *For any closed set F,*

$$\limsup_{n\to\infty} \frac{1}{n}\log\mu_n(F) \leq -\inf_{x\in F}\Lambda^*(x). \tag{61}$$

(b) *For any open set G,*

$$\liminf_{n\to\infty} \frac{1}{n}\log\mu_n(G) \geq -\inf_{x\in G\cap\mathcal{G}}\Lambda^*(x). \tag{62}$$

Proof. Theorem 4 is a slightly modified version of [14, Theorem 2.3.6]. Claim (a) is precisely [14, Theorem 2.3.6(a)]. Next, [14, Theorem 2.3.6(b)] states that

$$\liminf_{n\to\infty} \frac{1}{n}\log\mu_n(G) \geq -\inf_{x\in G\cap\mathcal{F}}\Lambda^*(x), \tag{63}$$

where $\mathcal{F}$ is the set of *exposed points* of $\Lambda^*(\cdot)$ (cf. [14, Definition 2.3.3]). However, [14, Lemma 2.3.9(b)] implies that $\mathcal{G} \subseteq \mathcal{F}$, and therefore

$$\inf_{x\in G\cap\mathcal{G}}\Lambda^*(x) \geq \inf_{x\in G\cap\mathcal{F}}\Lambda^*(x), \tag{64}$$

which completes the proof of Claim (b). □

References

[1] 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.

[2] V. Strassen, “Asymptotische abschäzungen in Shannons informationstheorie,” in *Trans. Third Prague Conf. Information Theory, on Statistics, Decision Functions, Random Processes, (Liblice, 1962)*. Prague: Publ. House Czech. Acad. Sci., 1964, pp. 689–723.

[3] L. L. Campbell, “A coding theorem and Rényi’s entropy,” *Information and Control*, vol. 8, no. 4, pp. 423–429, 1965.

[4] ——, “Definition of entropy by means of a coding problem,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 6, pp. 113–118, 1966.

[5] M. Christiansen and K. Duffy, “Guesswork, large deviations, and shannon entropy,” *IEEE Trans. Inf. Thy.*, vol. 59, no. 2, pp. 796–802, 2013.

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

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

[8] M. K. Hanawal and R. Sundaresan, “Guessing revisited: A large deviations approach,” *IEEE Trans. Inf. Th.*, vol. 57, no. 1, pp. 70–78, 2011.

[9] C. E. Shannon, “Notes on the relation of error probability to delay in a noisy channel,” MIT, Cambridge, MA, August 30, 1956.

[10] I. Csiszár and J. Korner, *Information Theory - Coding Theorems for Discrete Memoryless Systems*. Cambridge, 2011.

[11] J. H. Curtiss, “A note on the theory of moment generating functions,” *The Annals of Mathematical Statistics*, vol. 13, no. 4, pp. 430–433, 1942.

[12] M. Kovačević, I. Stanojević, and V. Šenk, “Some properties of Rényi entropy over countably infinite alphabets,” *Problems of Information Transmission*, vol. 49, no. 2, pp. 99–110, 2013.

[13] W. Rudin, *Principles of mathematical analysis*, 3rd ed. McGraw-Hill New York, 1964.

[14] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., ser. Stochastic Modelling and Applied Probability. Springer, 1998, vol. 38.