Lossless Compression with Moderate Error Probability

Yücel Altuğ
School of Elect. and Comp. Eng.
Cornell University
Ithaca, NY, USA
Email: ya68@cornell.edu

Aaron B. Wagner
School of Elect. and Comp. Eng.
Cornell University
Ithaca, NY, USA
Email: wagner@ece.cornell.edu

Ioannis Kontoyiannis
Department of Informatics
Athens Univ. of Econ. and Business
Athens, Greece
Email: yiannis@aueb.gr

***Abstract*—For the problem of lossless compression of a memoryless source, we give a detailed, precise characterization of the best achievable error probability, in the "moderate error probability" regime. This is the asymptotic setting where the probability of error decays to zero while at the same time the rate converges to the entropy at a speed no faster than $1/\sqrt{N}$. These results combine some of the essential benefits of earlier analyses in terms of error exponents and of Gaussian approximation. Analogous results for the problem of hypothesis testing are also established.**

I. Introduction

Perhaps the most classical information-theoretic question is that of determining the best achievable communication rate while processing large blocks of data. This results in the determination of the entropy rate and the rate-distortion function of a random source, and of the capacity of a noisy channel; see, e.g., [1] and references therein. In recent years, a significant amount of effort has been devoted to the examination of the best possible "second order" performance, namely to the development and analysis of finer asymptotic (large blocklength) results. One way to categorize these refined asymptotics is by dividing them into three cases: The small, moderate and large error probability regimes, respectively. In the small error probability regime the goal is to minimize the error probability under a fixed rate constraint. This regime is classical; it is well-known that the error probability typically vanishes exponentially fast, and the best possible error exponent is well-understood in various information-theoretic scenarios; see, e.g., [1]–[5] and references therein. Moreover, for the problem of lossless compression of a memoryless source, the optimal order of the sub-exponential pre-factor is also known [6], and corresponding results for channel coding have recently been developed [7]–[10].

In the large error probability regime, one considers a positive error threshold and the aim is to characterize the fastest speed at which the rate can approach the ideal asymptotic limit, with the constraint that error probability does not exceed this threshold. Several classical problems have been investigated from this perspective, see, e.g., [11]–[23] and references therein. Indeed, detailed expansions that not only give the leading term, but also the lower order terms have been reported in, e.g., [11], [16], [17], [19], [20], [22], [23].

The *moderate error probability regime* has only been examined recently; it aims to combine the desired features of both the large and small probability regimes, by considering the case where the error probability vanishes and at the same time the rate approaches the ideal asymptotic limit. E.g., in the case of lossless compression, we would allow the rate to approach the entropy at a speed slower than in the large error probability regime, and investigate the behavior of the smallest possible error probability. Previous work along these lines has been described in [24]–[29]. These studies report that error probability vanishes sub-exponentially fast, and they characterize the rate of this decay.

Unlike with large and small error probability, little is known about the lower-order terms in the moderate error probability regime. In this paper we consider the problem of obtaining more accurate results in this direction, for the problem of losslessly compressing a block $X^N = (X_1, X_2, \ldots, X_N)$ generated by a discrete memoryless source $\{X_N\}$ with marginal distribution p on a finite alphabet $\mathcal{X}$. Our main contribution is the derivation of a precise asymptotic characterization of the optimal error probability in the moderate error probability regime. Specifically, let $\mathrm{P_e}(p, N, k)$ denote the smallest achievable probability that the compressed lengths of any lossless compressor exceed the rate k/N. Let R_N denote a sequence of rates that converge to the entropy $\mathrm{H}(p)$ of the source slowly enough so that $R_N - \mathrm{H}(p) \geq c/\sqrt{N}$ for some positive constant c. In Theorem 2.1 we derive an explicit function $F_N(R_N)$ and we show that, asymptotically, $\mathrm{P_e}(p, N, \lceil NR_N \rceil) \approx F_N(R_N)$. The function $F_N(R)$ is of the form,

$$F_N(R) = G(R,p)\frac{1}{N^{(1+\mathrm{e}'_{\mathrm{S}}(R))/2}}e^{-N\mathrm{e}_{\mathrm{S}}(R)},$$

where $\mathrm{e_S}(R)$ is the usual source coding error exponent and the constant $G(R,p)$ is explicitly identified; see equation (7). Several consequences of this result are given and discussed in the following section. For example, this result recovers and refines the Moderate Deviations Principle (MDP) for this problem (see Corollary 2.1), which had earlier been established by other means [24]. We argue that deducing the MDP in this manner is particularly illuminating (see Remark 2.2). Although it is not our focus, our result also gives bounds in the large error probability regime. There it is order optimal, although its bounds on the relevant constants are tight only when the error probabilities are very small (see Corollary 2.2). Note that, both results stated above, are established in three cases: (i) when in the definition of $\mathrm{P}_{\mathrm{e}}(p, N, k)$ we consider all variable-length prefix codes; (ii) when the prefix constraint is removed and we examine all possible injective compressors; (iii) and when we consider all fixed-length codes.

The potential utility of these results for practically approximating the fundamental limit $\mathrm{P}_{\mathrm{e}}(p, N, k)$ at finite blocklengths N will be examined in future work.

II. Main Results

Let $\{X_N\}$ be a memoryless source with marginal distribution p on the finite alphabet $\mathcal{X}$. A fixed-to-variable lossless source code for blocks $\mathbf{x}^N = (x_1, x_2, \ldots, x_N) \in \mathcal{X}^N$ of length N generated by the source $\{X_N\}$, is a pair (f_N, ϕ_N) where the encoder is an injective function $f_N : \mathcal{X}^N \to \{0,1\}^* := \{\emptyset, 0, 1, 00, 01, 10, 11, \ldots\}$ and the decoder $\phi_N : \{0,1\}^* \to \mathcal{X}^N$ is such that $\phi_N(f_N(\mathbf{x}^N)) = \mathbf{x}^N$ for all $\mathbf{x}^N$. Similarly, given positive integers N, k, an N-to-k fixed-to-fixed source code (g_N, φ_N): consists of an encoder $g_N : \mathcal{X}^N \to \{0,1\}^k$ and a decoder $\varphi_N : \{0,1\}^k \to \mathcal{X}^N$.

For any N, k and any source distribution p, the best achievable compression performance at finite blocklengths N can be described in terms of the following fundamental limits:

$$\mathrm{P}_{\mathrm{e},1}(p, N, k) := \min \Pr\left\{\ell(f_N(\mathbf{X}^N)) \geq k\right\}, \tag{1}$$
$$\mathrm{P}_{\mathrm{e},2}(p, N, k) := \min \Pr\left\{\ell(f_N(\mathbf{X}^N)) \geq k\right\}, \tag{2}$$
$$\mathrm{P}_{\mathrm{e},3}(p, N, k) := \min \Pr\left\{\varphi_N(g_N(\mathbf{X}^N)) \neq \mathbf{X}^N\right\}. \tag{3}$$

The minima in (1), (2) and (3) above are over all variable-length compressors, all prefix-free variable-length compressors, and all fixed-length compressors, respectively, where $\ell(\mathbf{a})$ denotes the length of a binary string $\mathbf{a} \in \{0,1\}^*$.

As shown in [23, Theorem 1], for any N, k with $k < N \log_2 |\mathcal{X}|$ and any source distribution p, we have the following simple relationships:

$$\mathrm{P}_{\mathrm{e},3}(p, N, k) \leq \mathrm{P}_{\mathrm{e},1}(p, N, k) = \mathrm{P}_{\mathrm{e},2}(p, N, k+1). \tag{4}$$

To avoid trivialities, in the sequel we assume that p has full support and that it is not the uniform distribution on $\mathcal{X}$. The varentropy of the source is defined (in nats) as $\sigma^2(p) := \mathrm{Var}_p(-\ln p(X))$. Note that our assumption that p is not uniform is equivalent to the requirement that the varentropy is nonzero.

Recall that the source coding error exponent is defined, for any $R \in [\mathrm{H}(p), \ln|\mathcal{X}|]$, is,

$$\mathrm{e}_{\mathrm{S}}(R) := \min_{q \,:\, \mathrm{H}(q) \geq R} \mathrm{D}(q \| p),$$

where the entropy and relative entropy are defined (in nats) as usual

We are now in a position to state our main result; its proof is given in Section III.

Theorem 2.1: Consider a sequence of rates $\{R_N\}$ such that the difference $\tau_N = R_N - \mathrm{H}(p)$ converges to zero slowly enough that $\tau_N \geq c/\sqrt{N}$ for some constant $c > 0$. Then for each $i = 1, 2, 3$, we have:

$$\limsup_{N\to\infty} \frac{\mathrm{P}_{\mathrm{e},i}(p, N, \lceil N R_N \rceil)}{F_N(R_N)} \leq 1, \tag{5}$$
$$\liminf_{N\to\infty} \frac{\mathrm{P}_{\mathrm{e},i}(p, N, \lceil N R_N \rceil)}{F_N(R_N)} \geq 1 - \sqrt{\frac{\sigma^2(p)}{e}} \limsup_{N\to\infty} \frac{1}{\tau_N \sqrt{N}}, \tag{6}$$

where,

$$F_N(R) := \frac{e^{-N \mathrm{e}_{\mathrm{S}}(R)} \left[2\pi N \sigma^2(\gamma(R), p)\right]^{-0.5(1+\mathrm{e}'_{\mathrm{S}}(R))}}{\gamma(R)(1-\gamma(R))^{\mathrm{e}'_{\mathrm{S}}(R)}}, \tag{7}$$

with $\gamma(R) := \frac{\mathrm{e}'_{\mathrm{S}}(R)}{1+\mathrm{e}'_{\mathrm{S}}(R)}$, where $\mathrm{e}'_{\mathrm{S}}(R)$ denotes the derivative of $\mathrm{e}_{\mathrm{S}}(\cdot)$ at R.

Remark 2.1: In the sequel, it will be evident that we actually have, $\lim_{N\to\infty}\left((1-\gamma(R_N))\sqrt{2\pi\sigma^2(\gamma(R_N), p)}\right)^{-\mathrm{e}'_{\mathrm{S}}(R_N)} = 1$, for R_N as given in Theorem 2.1. We include this term to have a better finite N characterization, which might be useful in future studies.

Next, we examine the consequences of Theorem 2.1 under more detailed assumptions on the rate sequence $\{R_N\}$. The following two corollaries are proved in Section III.

Corollary 2.1: Consider a sequence of rates $\{R_N\}$ such that the difference $\tau_N = R_N - \mathrm{H}(p)$ is positive and converges to zero slowly enough that $\sqrt{N}\tau_N \to \infty$ as $N \to \infty$. Then for each $i = 1, 2, 3$, we have:

$$\lim_{N\to\infty} \frac{\ln \mathrm{P}_{\mathrm{e},i}(p, N, \lceil N R_N \rceil)}{\tau_N^2 N} = -\frac{1}{2\sigma^2(p)}. \tag{8}$$

Remark 2.2: Theorem 2.1 sheds light on the following fact. If we use the crude approximation $\mathrm{P}_{\mathrm{e},i}(N, R) \approx e^{-N\mathrm{e}_{\mathrm{S}}(R)}$ and formally apply the scaling of Corollary 2.1, we obtain the correct answer given in (8). Yet if we use the seemingly more-accurate approximation $\mathrm{P}_{\mathrm{e},i}(N, R) \approx \frac{e^{-N\mathrm{e}_{\mathrm{S}}(R)}}{\sqrt{N}}$, which is suggested by the results of Csiszár and Longo [6] on the small error probability regime, we do not. The reason is that the constant that is ignored in the second approximation diverges as R approaches $\mathrm{H}(p)$, and this exactly cancels the $1/\sqrt{N}$ factor. Existing MDP proofs (including [26] for channel coding) obscure this fact, while Theorem 2.1 brings it to the surface.

Remark 2.3: Corollary 2.1 gives a complete MDP for all three types of compressors, and in particular it recovers the MDP for fixed-length lossless source coding result[1] in [24].

Corollary 2.2: Fix some $b > 0$ consider the rates $R_N :=$

[1] The result of [24] is actually for source coding with side information, which subsumes the fixed-length lossless source setting.

$\mathrm{H}(p)+\tau_N$, with $\tau_N := \sqrt{\frac{\sigma^2(p)}{N}}b$. For each $i = 1, 2, 3$ we have:

$$\limsup_{N\to\infty} \mathrm{P}_{\mathrm{e},i}(p, N, \lceil NR_N\rceil) \leq \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}, \tag{9}$$

$$\liminf_{N\to\infty} \mathrm{P}_{\mathrm{e},i}(p, N, \lceil NR_N\rceil) \geq \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}\left(1 - \frac{1}{\sqrt{eb^2}}\right). \tag{10}$$

Remark 2.4: The subsequential limits in (9) and (10) are both equal to $Q(b)$, where $Q(\cdot)$ is the Gaussian Q-function, because of the fact that the second-order term in the best achievable compression rate is $\sqrt{\frac{\sigma^2(p)}{N}}Q^{-1}(\epsilon)$ [23]. Thus, this result is order optimal, but it does not determine the relevant constant exactly. It is close, however. Indeed, the right-hand side of (9) is a commonly-used upper bound on $Q(b)$, and the ratio of the two bounds tends to one as b tends to infinity.

III. Proofs

Boldface letters denote vectors, regular letters with subscripts denote their elements, capital letters represent random variables, lowercase letters denote their individual realizations. For a finite set $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ is the set of all probability mass functions on $\mathcal{X}$, $U_{\mathcal{X}}$ is the uniform distribution on $\mathcal{X}$, and p^N denotes the N-fold product distribution of $p \in \mathcal{P}(\mathcal{X})$ on $\mathcal{X}^N$.

We shall prove binary hypothesis testing results that will be used to prove the results of Section II. To this end, let $\mathcal{X}$ be an arbitrary finite set, and take $P \neq Q \in \mathcal{P}(\mathcal{X})$ to be arbitrary distributions of full support. For any $\lambda \in \mathbb{R}$ and $x \in \mathcal{X}$, let $\tilde{Q}_\lambda(x) := \frac{P(x)^{1-\lambda}Q(x)^\lambda}{\sum_{z\in\mathcal{X}} P(z)^{1-\lambda}Q(z)^\lambda}$, and $\sigma^2(\lambda, P, Q) := \mathrm{Var}_{\tilde{Q}_\lambda}[\ln(Q(X)/P(X))]$. Note that, since $P \neq Q$, we have $\sigma^2(\cdot, P, Q) \in \mathbb{R}^+$ on $\mathbb{R}$. Also, let $\sigma^2(P, Q) := \sigma^2(0, P, Q)$. Define $\mathrm{e_H}(A) := \min_{\hat{Q}\in\mathcal{P}(\mathcal{X}) : \mathrm{D}(\hat{Q}||Q)\leq A} \mathrm{D}(\hat{Q}||P)$, for $A \geq 0$.

Further, for $A \in (0, \mathrm{D}(P||Q)]$, let $\eta(A) := \frac{|\mathrm{e}'_\mathrm{H}(A)|}{1+|\mathrm{e}'_\mathrm{H}(A)|}$, where $\mathrm{e}'_\mathrm{H}(A)$ denotes the derivative of $\mathrm{e_H}(\cdot)$ at A, and,

$$J_N(A) := \frac{e^{-N\mathrm{e_H}(A)}\left[2\pi N\sigma^2(\eta(A), P, Q)\right]^{-0.5(1+|\mathrm{e}'_\mathrm{H}(A)|)}}{\eta(A)(1-\eta(A))^{|\mathrm{e}'_\mathrm{H}(A)|}}.$$

Consider any $N \in \mathbb{Z}^+$. Let P^N and Q^N denote the null and alternate hypothesis, respectively. For any (potentially randomized) binary hypothesis test T_N, $\alpha(T_N)$ and $\beta(T_N)$ denotes type-I (P^N probability of the event that the test decides Q^N) and type-II (Q^N probability of the event that the test decides P^N) error probability, respectively.

Theorem 3.1: Consider a positive sequence $\{\xi_N\}$ with $\xi_N = \Omega(1/\sqrt{N})$, $\xi_N = o(1)$. Let $A_N := \mathrm{D}(P||Q) - \xi_N$.

(i) There exists a sequence of deterministic hypothesis tests, $\{T_N\}$, such that $\beta(T_N) \leq e^{-NA_N}$ for all N and:

$$\limsup_{N\to\infty} \frac{\alpha(T_N)}{J_N(A_N)} \leq 1. \tag{11}$$

(ii) For any sequence of (potentially randomized) hypothesis tests $\{T_N\}$ with $\beta(T_N) \leq e^{-NA_N}$ for all N, we have:

$$\liminf_{N\to\infty} \frac{\alpha(T_N)}{J_N(A_N)} \geq 1 - \sqrt{\frac{\sigma^2(P,Q)}{e}} \limsup_{N\to\infty} \frac{1}{\xi_N\sqrt{N}}.$$

An outline of the proof of Theorem 3.1 given in the Appendix. The following corollaries are easy consequences of Theorem 3.1, once one notes that

$$\lim_{N\to\infty} \frac{\mathrm{e_H}(A_N)}{\xi_N^2} = \frac{1}{2\sigma^2(P,Q)},$$

which can be verified via Lemma A.2.

Corollary 3.1: Consider a positive sequence $\{\xi_N\}$ with $\xi_N = o(1)$, $\lim_{N\to\infty} \xi_N\sqrt{N} = \infty$. Let $A_N := \mathrm{D}(P||Q)-\xi_N$.

(i) There exists a sequence of deterministic hypothesis tests, $\{T_N\}$, such that $\beta(T_N) \leq e^{-NA_N}$ for all N and:

$$\limsup_{N\to\infty} \frac{\ln\alpha(T_N)}{\xi_N^2 N} \leq -\frac{1}{2\sigma^2(P,Q)}. \tag{12}$$

(ii) For any sequence of (potentially randomized) hypothesis tests $\{T_N\}$ with $\beta(T_N) \leq e^{-NA_N}$ for all N, we have:

$$\liminf_{N\to\infty} \frac{\ln\alpha(T_N)}{\xi_N^2 N} \geq -\frac{1}{2\sigma^2(P,Q)}. \tag{13}$$

Remark 3.1: Corollary 3.1 gives a complete MDP for the binary hypothesis testing problem, which is new[2] to the best of our knowledge.

Corollary 3.2: Fix some $b \in \mathbb{R}^+$ and define $A_N := \mathrm{D}(P||Q) - \xi_N$ with $\xi_N := \sqrt{\frac{\sigma^2(P,Q)}{N}}b$.

(i) There exists a sequence of deterministic hypothesis tests, $\{T_N\}_{N\geq 1}$, with $\beta(T_N) \leq e^{-NA_N}$ for all N and:

$$\limsup_{N\to\infty} \alpha(T_N) \leq \limsup_{N\to\infty} J_N(A_N) = \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}. \tag{14}$$

(ii) For any sequence of (potentially randomized) hypothesis tests $\{T_N\}$ with $\beta(T_N) \leq e^{-NA_N}$ for all N, we have

$$\begin{aligned}\liminf_{N\to\infty} \alpha(T_N) &\geq \liminf_{N\to\infty} J_N(A_N)\times \\ &\quad \left(1 - \sqrt{\frac{\sigma^2(P,Q)}{e}} \limsup_{N\to\infty} \frac{1}{\xi_N\sqrt{N}}\right) \\ &= \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}\left(1 - \frac{1}{\sqrt{eb^2}}\right).\end{aligned} \tag{15}$$

Comments similar to those made in Remark 2.4 also apply to Corollary 3.2.

Next, we prove the results in Section II by invoking the previous hypothesis testing results with $Q = U_{\mathcal{X}}$ and $P = p$. To that end, let $\{\tau_N\}$ be such that $\tau_N = \Omega(1/\sqrt{N})$, $\tau_N = o(1)$ and $\tau_N \in \mathbb{R}^+$. Define $R_N := \mathrm{H}(p) + \tau_N$ and $r_N := R_N + \zeta_N$ such that $|\zeta_N| = O(1/N)$. We note that,

$$\lim_{N\to\infty} \frac{F_N(R_N)}{F_N(r_N)} = 1, \tag{16}$$

whose proof is omitted due to space restrictions.

First, we prove Theorem 2.1 using Theorem 3.1. In light of (4), it suffices to prove (5) (resp. (6)) for $\mathrm{P}_{\mathrm{e},2}(N, \lceil NR_N\rceil)$ (resp. $\mathrm{P}_{\mathrm{e},3}(N, \lceil NR_N\rceil)$). The former (resp. latter) follows by applying item (i) (resp. (ii)) of Theorem 3.1 with $A_N =$

[2] The moderate deviations result in [29] refers to a different setup.

$\ln|\mathcal{X}| - R_N + \frac{\ln 2}{N} - \frac{\ln(1-e^{-NR_N+\ln 2})}{N}$ (resp. $A_N = \ln|\mathcal{X}| - R_N - \frac{\ln 2}{N}$) and recalling (16).

To prove Corollary 2.1, let R_N be as given in the statement of the corollary. From (4), it suffices to check,

$$\limsup_{N\to\infty} \frac{\ln \mathrm{P}_{\mathrm{e},2}(p, N, \lceil NR_N \rceil)}{\tau_N^2 N} \leq -\frac{1}{2\sigma^2(p)},$$
$$\liminf_{N\to\infty} \frac{\ln \mathrm{P}_{\mathrm{e},3}(p, N, \lceil NR_N \rceil)}{\tau_N^2 N} \geq -\frac{1}{2\sigma^2(p)}.$$

The former (resp. latter) inequality follows by applying item (i) (resp. (ii)) of Corollary 3.1 with $A_N = \ln|\mathcal{X}| - R_N + \frac{\ln 2}{N} - \frac{\ln(1-e^{-NR_N+\ln 2})}{N}$ (resp. $A_N = \ln|\mathcal{X}| - R_N - \frac{\ln 2}{N}$).

We conclude with the proof of Corollary 2.2. To this end, fix some $b \in \mathbb{R}^+$ and define $\tau_N := \sqrt{\frac{\sigma^2(p)}{N}}b$ and $R_N := \mathrm{H}(p) + \tau_N$. Note that to prove (9), it suffices to check,

$$\limsup_{N\to\infty} \mathrm{P}_{\mathrm{e},2}(p, N, \lceil NR_N \rceil) \leq \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}, \tag{17}$$

by (4). Equation (17) follows from (5) and invoking the equality in (14) with $P = p$, $Q = U_{\mathcal{X}}$ and $A_N = \ln 2 - R_N$.

Similarly, to prove (10), it suffices to check

$$\liminf_{N\to\infty} \mathrm{P}_{\mathrm{e},3}(p, N, \lceil NR_N \rceil) \geq \frac{e^{-b^2/2}}{\sqrt{2\pi b^2}}\left(1 - \frac{1}{\sqrt{eb^2}}\right), \tag{18}$$

due to (4). Eq. (18) follows from (6) and applying the equality in (15) with $P = p$, $Q = U_{\mathcal{X}}$ and $A_N = \ln 2 - R_N$.

APPENDIX

We start with a concentration result that will be used repeatedly; its proof, which is omitted due to space restrictions, resembles Dembo-Zeitouni's proof of exact asymptotics theorem of Bahadur-Rao (cf. [30, Theorem 3.7.4]) with the main difference that the proof of Lemma A.1 uses Berry-Esseen theorem (e.g. [31, Theorem III.1]) whereas the proof of [30, Theorem 3.7.4] uses Berry-Esseen expansion (e.g. [32, pg. 538–540]). This difference enables us to prove a result that is valid for all $n \in \mathbb{Z}^+$ at the expense of slightly looser constants.

Let $\{Z_i\}$ be independent and identically distributed random variables with law μ, such that $\mathrm{Var}[Z_1] > 0$. Let $\Lambda(\cdot)$ (resp. $\Lambda^*(\cdot)$) denote the log-moment generating function (resp. Fenchel-Legendre transform) of Z_1 (resp. $\Lambda(\cdot)$). Assume the existence of some $q \in \mathbb{R}$ and a corresponding $\eta(q) \in \mathbb{R}^+$ such that: (i) There exists a neighborhood of $\eta(q)$ where $\Lambda(\cdot) < \infty$; and (ii) $\Lambda'(\eta(q)) = q$. Define the "tilted" probability measure via $\frac{d\tilde{\mu}_{\eta(q)}}{d\mu}(z) := e^{\eta(q)z - \Lambda(\eta(q))}$, $T_i := \frac{Z_i - q}{\sqrt{\Lambda''(\eta(q))}}$, and $m_3(\eta(q)) := \mathrm{E}_{\tilde{\mu}_{\eta(q)}}[|T_1|^3]$. Set $t(a,q) := a\eta(q)m_3(\eta(q))\sqrt{2\pi\Lambda''(\eta(q))}$ for any $a \geq 1$.

Lemma A.1: For any $n \in \mathbb{Z}^+$ and $a > 1$,

$$\frac{e^{-n\Lambda^*(q)}[1 + \eta(q)\sqrt{2\pi\Lambda''(\eta(q))}m_3(\eta(q))]}{\eta(q)\sqrt{2\pi n\Lambda''(\eta(q))}} \geq$$
$$\Pr\left[\frac{1}{n}\sum_{i=1}^n Z_i \geq q\right] \geq \frac{e^{-n\Lambda^*(q)}e^{-t(a,q)}\left(1-\frac{1}{a}\right)(1+t(a,q))}{\eta(q)\sqrt{2\pi n\Lambda''(\eta(q))}}$$
$$\left\{1 - \frac{[1 + (1+t(a,q))^2]}{(1+t(a,q))\eta(q)\left(1-\frac{1}{a}\right)\sqrt{en\Lambda''(\eta(q))}}\right\}.$$

Define $\Lambda_0(\lambda) := \ln \mathrm{E}_P\left[e^{\lambda \ln\frac{Q(X)}{P(X)}}\right]$ and $\Lambda_1(\lambda) := \Lambda_0(1-\lambda)$, $\lambda \in \mathbb{R}$ and let $\Lambda_i^*(\cdot)$ be the Fenchel-Legendre transform of Λ_i, $i = 0, 1$. The following properties are easy to verify:

Lemma A.2: For any $A \in (0, \mathrm{D}(P||Q)]$:

(i) There exists a unique $\eta(A) \in [0,1)$, such that $\Lambda_0'(\eta(A)) = \mathrm{e}_{\mathrm{H}}(A) - A$ and $\Lambda_1'(1-\eta(A)) = A - \mathrm{e}_{\mathrm{H}}(A)$.
(ii) $\Lambda_0^*(\mathrm{e}_{\mathrm{H}}(A) - A) = \mathrm{e}_{\mathrm{H}}(A)$ and $\frac{\eta(A)}{1-\eta(A)} = |\mathrm{e}_{\mathrm{H}}'(A)|$.
(iii) $\Lambda_1^*(A - \mathrm{e}_{\mathrm{H}}(A)) = A$.
(iv) $\eta'(A) = \frac{\mathrm{e}_{\mathrm{H}}'(A)-1}{\Lambda_0''(\eta(A))}$.

Lemma A.3: For any $A \in (0, \mathrm{D}(P||Q))$:

(i) ${\Lambda_0^*}'(\mathrm{e}_{\mathrm{H}}(A) - A) = \eta(A)$, ${\Lambda_1^*}'(A - \mathrm{e}_{\mathrm{H}}(A)) = 1 - \eta(A)$.
(ii) ${\Lambda_0^*}''(\mathrm{e}_{\mathrm{H}}(A) - A) = {\Lambda_1^*}''(A - \mathrm{e}_{\mathrm{H}}(A)) = \frac{1}{\Lambda_0''(\eta(A))}$.

Outline of the Proof of Theorem 3.1: First, define,

$$m_{0,3}(\lambda) := \mathrm{E}_{\tilde{Q}_\lambda}\left[\left|\ln\frac{Q(X)}{P(X)} - \mathrm{E}_{\tilde{Q}_\lambda}\left[\ln\frac{Q(X)}{P(X)}\right]\right|^3\right],$$
$$m_{1,3}(\lambda) := \mathrm{E}_{\tilde{Q}_\lambda}\left[\left|\ln\frac{P(X)}{Q(X)} - \mathrm{E}_{\tilde{Q}_\lambda}\left[\ln\frac{P(X)}{Q(X)}\right]\right|^3\right],$$
$$t_i(A,a) := a(i + (-1)^i\eta(A))m_{i,3}(\eta(A))\sqrt{2\pi\sigma^2(\eta(A), P, Q)}$$

for any $i \in \{0,1\}$, $\lambda \in \mathbb{R}$, $a \in \mathbb{R}^+$ and $A \in (0, \mathrm{D}(P||Q))$.

To prove (i), fix $\delta \in \mathbb{R}^+$ and define $r_N := A_N - \epsilon_N$ with,

$$\epsilon_N := \frac{1}{N}\ln\left(\frac{(1-\eta(A_N))\sqrt{2\pi N\Lambda_0''(\eta(A_N))}}{(1+\delta)(1+t_1(A_N,1))}\right).$$

Consider large enough N s.t. $r_N \in \mathbb{R}^+$. Define,

$$\mathcal{A}_N := \left\{\mathbf{x}^N \in \mathcal{X}^N : \frac{1}{N}\ln\frac{P^N(\mathbf{x}^N)}{Q^N(\mathbf{x}^N)} \geq r_N - \mathrm{e}_{\mathrm{H}}(r_N)\right\},$$

and consider a deterministic hypothesis test that decides P if the observed sequence is in $\mathcal{A}_N$ and decides Q otherwise. Let,

$$\alpha_N := P^N\{\mathcal{A}_N^c\} \text{ and } \beta_N := Q^N\{\mathcal{A}_N\}.$$

Combining Lemma A.1 with Lemma A.2 (iii), it can be shown,

$$\beta_N \leq e^{-Nr_N}\frac{[1 + t_1(r_N,1)]}{(1-\eta(r_N))\sqrt{2\pi N\Lambda_0''(\eta(r_N))}} \leq e^{-NA_N},$$

for N large enough, so the type-II error constraint is satisfied. And using Lemma A.1 again,

$$\alpha_N \leq e^{-N\Lambda_0^*(\mathrm{e}_{\mathrm{H}}(r_N) - r_N)}\frac{[1 + t_0(r_N,1)]}{\eta(r_N)\sqrt{2\pi N\Lambda_0''(\eta(r_N))}}. \tag{19}$$

We can show that $\limsup_{N\to\infty}\frac{\alpha_N}{J_N(A_N)} \leq 1$ by combining Lemmas A.2 and A.3 with (19), giving item (i) of the theorem.

We conclude with the proof of item (ii). Fix some $\delta \in \mathbb{R}^+$ and $a > 1$. Further, consider a sufficiently large N such that both $A_N \in (0, \mathrm{D}(P||Q))$ and $\delta_N := \left[1 - \frac{[1+(1+t_1(A_N,a))^2]}{(1+t_1(A_N,a))(1-\eta(A_N))\left(1-\frac{1}{a}\right)\sqrt{eN\Lambda_0''(\eta(A_N))}}\right] > 0$. Define

$$\epsilon_N := \frac{1}{N} \ln \left(\frac{(1+\delta)(1-\eta(A_N))\sqrt{2\pi N \Lambda_0''(\eta(A_N))}}{e^{-t_1(A_N,a)}\left(1-\frac{1}{a}\right)(1+t_1(A_N,a))\delta_N} \right).$$

Also, let $\tilde{\epsilon}_N := \epsilon_N\left(1+\frac{1}{N}\right)$, $r_N := A_N - \epsilon_N$ and $\tilde{r}_N := A_N - \tilde{\epsilon}_N$. Consider sufficiently large N such that $\tilde{r}_N > 0$ and let $\mathcal{A}_N$ be as before, but with this r_N.

Combining Lemma A.1 with Lemma A.2 (iii) it can be shown that, for large enough N, $\beta_N > e^{-NA_N}$, which implies that this test violates the type-II error constraint. Since it is a likelihood ratio test, this violation can only improve the type-I error probability of the best test that satisfies the constraint, therefore the type-I error probability α_N gives a lower bound on the type-I error probability of the best feasible test. Hence, in order to conclude the proof, we bound α_N below.

By applying Lemma A.1 once more, we have,

$$\alpha_N \geq \frac{e^{-N\Lambda_0^*(\mathrm{e_H}(\tilde{r}_N)-\tilde{r}_N)}e^{-t_0(\tilde{r}_N,a)}\left(1-\frac{1}{a}\right)(1+t_0(\tilde{r}_N,a))}{\eta(\tilde{r}_N)\sqrt{2\pi N\Lambda_0''(\eta(\tilde{r}_N))}} \times \left\{1 - \frac{[1+(1+t_0(\tilde{r}_N,a))^2]}{(1+t_0(\tilde{r}_N,a))\eta(\tilde{r}_N)\left(1-\frac{1}{a}\right)\sqrt{eN\Lambda_0''(\eta(\tilde{r}_N))}}\right\}.$$

Using Lemma A.2 and A.3, it can be verified that this implies $\liminf_{N\to\infty} \frac{\alpha_N}{J_N(A_N)} \geq 1 - \sqrt{\frac{\sigma^2(P,Q)}{e}} \limsup_{N\to\infty} \frac{1}{\xi_N\sqrt{N}}$, which implies item (ii) of the theorem.

Acknowledgment

The first author thanks Emre Telatar and Paul Cuff for their hospitality while portions of this work were being completed during his visits to ÉPFL and Princeton University. This research was supported by the National Science Foundation under grant CCF-1218578. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) Research Funding Program THALES: Investing in knowledge society through the European Social Fund.

References

[1] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[2] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels," *Inform. Contr.*, vol. 10, pp. 65–103, Jan. 1967

[3] K. Marton, "Error Exponent for Source Coding with a Fidelity Criterion," *IEEE Trans. Inform. Theory*, vol. IT–20, pp. 197–199, Mar. 1974.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[5] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT–28, pp. 823–828, Nov. 1982.

[6] I. Csiszár and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses," *Studia Sci. Math. Hung.*, vol. 6, pp. 181–191, 1971.

[7] Y. Altuğ and A. B. Wagner, "Refinement of the Sphere Packing Bound for Symmetric Channels," in *Proc. 49th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2011.

[8] Y. Altuğ and A. B. Wagner, "A Refinement of the Random Coding Bound," in *Proc. 50th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2012.

[9] Y. Altuğ and A. B. Wagner, "Refinement of the Sphere-Packing Bound," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, Boston, MA, July 2012.

[10] Y. Altuğ and A. B. Wagner, "Refinement of the Sphere-Packing Bound: Asymmetric Channels," *submitted to IEEE Trans. on Inform. Theory.* Available from: http://arxiv.org/pdf/1211.6697v1.pdf.

[11] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," *Trans. Third Prague Conf. Information Theory*, 1962, Czechoslovak Academy of Sciences, Prague, pp. 689-723.

[12] I. Kontoyiannis, "Second-Order Noiseless Source Coding Theorems," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1339–1341, July 1997.

[13] I. Kontoyiannis, "Pointwise Redundancy in Lossy Data Compression and Universal Lossy Data Compression," *IEEE Trans. Inform. Theory*, vol. IT–46, pp. 136–152, Jan. 2000.

[14] M. Hayashi, "Second–Order Asymptotics in Fixed-Length Source Coding and Intrinsic Randomness," *IEEE Trans. Inform. Theory*, vol. IT–54, pp. 4619–4637, Oct. 2008.

[15] M. Hayashi, "Information Spectrum Approach to Second–Order Coding Rate in Channel Coding," *IEEE Trans. on Inform. Theory*, vol. IT 55, pp. 4947–4966, November 2009.

[16] Y. Polyanskiy, H. V. Poor and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Inform. Theory*, vol. IT–56, pp. 2307–2359, May 2010.

[17] W. Szpankowski and S. Verdú, "Minimum Expected Length of Fixed-to-Variable Lossless Compression Without Prefix Constraints," *IEEE Trans. Inform. Theory*, vol. IT–57, pp. 4017–4025, July 2011.

[18] Y. Polyanskiy, H. V. Poor and S. Verdú, "Feedback in the Non-Asymptotic Regime," *IEEE Trans. on Inform. Theory*, vol. 57, pp. 4903–4925, August 2011.

[19] V. Kostina and S. Verdú, "Fixed-Length Lossy Compression in the Finite Blocklength Regime," *IEEE Trans. on Information Theory*, vol. IT 58, pp. 3309–3338, June 2012.

[20] P. Moulin, "The log-volume of optimal constant-composition codes for memoryless channels, within O(1) bits," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, Boston, MA, July 2012.

[21] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *Proc. 50th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2012.

[22] M. Tomamichel and V. Y. F. Tan, "A Tight Upper Bound for the Third-Order Asymptotics of Discrete Memoryless Channels." Available from: http://arxiv.org/pdf/1212.3689v2.pdf.

[23] I. Kontoyiannis and S. Verdú, "Lossless Data Compression at Finite Blocklengths." Available from: http://arxiv.org/pdf/1212.2668v1.pdf.

[24] J. Chen, D.-k. He, A. Jagmohan and L. A. Lastras-Montaño, "On the Redundancy-Error Tradeoff in Slepian-Wolf Coding and Channel Coding," in *Proc. 2007 IEEE Int. Symp. Inf. Theory*, June 2007, pp. 1326–1330.

[25] D.-k He, L. A. Lastras-Montaño, E.-h. Yang, A. Jagmohan and J. Chen, "On the Redundancy of Slepian–Wolf Coding," *IEEE Trans. on Inform. Theory*, vol. IT 55, pp. 5607–5627, December 2009.

[26] Y. Altuğ and A. B. Wagner, "Moderate Deviation Analysis of Channel Coding: Discrete Memoryless Case," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*, June 2010, pp. 265–269.

[27] Y. Polyanskiy and S. Verdú, "Channel Dispersion and Moderate Deviations Limits for Memoryless Channels," in *Proc. 48th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2010, pp. 1334–1339.

[28] V. Y. F. Tan, "Moderate–Deviations of Lossy Source Coding for Discrete and Gaussian Source," *arXiv:1111.2217*, Nov. 2011.

[29] I. Sason, "On Refined Versions of the Azuma–Hoeffding Inequality with Applications in Information Theory," *arXiv:1111.1977*, Nov. 2011.

[30] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, 2nd edition*. New York: Springer–Verlag, 1998.

[31] C.-G. Esseen, "Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law," *Acta Math.*, vol. 77, pp. 1–125, 1945.

[32] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II, 2nd edition*. New York: Wiley, 1971.