Universal Compression at Pragmatic Rates

Andreas Theocharous Lampros Gavalakis Ioannis Kontoyiannis
University of Cambridge
{at771, lg560, ik355}@cam.ac.uk

***Abstract*—We consider the problem of optimal universal lossless data compression subject to a constraint on the excess-rate probability. In the non-universal setting, in the regime where the excess-rate probability is required to be below a fixed $\epsilon > 0$, the best asymptotic rate that can be achieved by one-to-one codes on memoryless sources is given by the source entropy. More accurate approximations relying on the central limit theorem provide sharper characterizations of the optimal rate at finite blocklengths. But when the constraint ϵ on the excess-rate probability is small, recent work showed that these approximations fail to be informative and can be quite inaccurate. In the small-excess-rate-probability regime, the best achievable asymptotic rate is given by the inverse of the error-exponent function, and an accurate approximation at finite blocklengths was obtained based on the central limit theorem and large deviations techniques. In this work, we show that, in this small-excess-rate-probability regime, the same performance as when the source distribution is known can be obtained universally over the class of all memoryless sources on a fixed alphabet of size m, at a cost of no more than $(\frac{m-2}{2})\log n$ bits, where n denotes the blocklength. The main new technical tool in the derivation of this result is a fine combinatorial estimate on the number of strings with empirical entropy below a given threshold, which may be of independent interest.**

I. Introduction

A. Variable-rate data compression

We consider the problem of compressing data generated by a memoryless source $\boldsymbol{X} = \{X_n \ ; \ n \geq 1\}$, consisting of independent and identically distributed (i.i.d.) random variables X_n, with values in a finite alphabet A, and with common distribution described by a probability mass function (p.m.f.) P: $\mathbb{P}(X_n = x) = P(x)$, $x \in A$. A *variable-rate code*, or *fixed-to-variable compressor*, for strings of length n from A is an injective function $f_n : A^n \to \{0,1\}^*$, where $\{0,1\}^*$ denotes the set of all finite-length binary strings:

$$\{0,1\}^* = \{\varnothing, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}.$$

Writing $\ell(s)$ for the length of a string $s \in \{0,1\}^*$, the compressor f_n maps strings $x^n = (x_1, \ldots, x_n) \in A^n$ into binary strings $f_n(x^n)$ of length $\ell(f_n(x^n))$ bits. We call such compressors *one-to-one codes*.

The best achievable performance among all fixed-to-variable compressors is quantified as follows. For a blocklength $n \geq 1$ and rate $R > 0$, the best achievable excess-rate probability $\epsilon_n^*(R, P)$ for a source $\boldsymbol{X}$ with distribution P is:

$$\epsilon_n^*(R, P) := \min_{f_n} \mathbb{P}\big(\ell(f_n(X^n)) \geq nR\big),$$

where X^n denotes the block of random variables $(X_1, \ldots, X_n)$. This minimum is achieved by an optimal compressor f_n^* independent of the rate R [17].

Similarly, for a blocklength $n \geq 1$ and excess-rate probability $\epsilon > 0$, the best achievable rate $R_n^*(\epsilon, P)$, is:

$$R_n^*(\epsilon, P) := \inf\big\{R > 0 : \min_{f_n} \mathbb{P}\big(\ell(f_n(X^n)) \geq nR\big) \leq \epsilon\big\}.$$

B. The Shannon regime

Suppose that the source distribution P is known, and that the excess-rate probability $\epsilon > 0$ is fixed. In this case, the evaluation of the first-order asymptotic behaviour of the optimal rate $R_n^*(\epsilon, P)$ essentially goes back to Shannon:

$$nR_n^*(\epsilon, P) = nH(P) + o(n) \ \text{ bits}, \quad \text{as } n \to \infty, \qquad (1)$$

where $H(P) = -\sum_x P(x)\log P(x)$ denotes the entropy, and, throughout the paper, 'log' denotes $\log_2$, the logarithm taken to base 2. The first-order asymptotic expansion (1) was later refined [17], [27], to the following: As $n \to \infty$,

$$nR_n^*(\epsilon, P) = nH(P) + \sqrt{n}\sigma(P)Q^{-1}(\epsilon) - \frac{1}{2}\log n + O(1) \ \text{ bits}, \qquad (2)$$

where $\sigma^2(P) = \mathrm{Var}_P(-\log P(X))$ is the *minimal coding variance* [14], [15] or *source dispersion* [17], and $Q(x) = 1 - \Phi(x)$, $x \in \mathbb{R}$, denotes the standard normal tail function. See [17] for a detailed account of the history of these results.

C. The error exponents regime

When the excess-rate probability is not fixed but is instead required to decay to zero exponentially fast, the best rate that can be achieved turns out to be higher than the entropy. Specifically, for $\delta > 0$ in an appropriate range, as $n \to \infty$,

$$nR_n^*(2^{-n\delta}, P) = nH(P_{\alpha^*}) + o(n) \ \text{ bits}, \qquad (3)$$

for an $\alpha^* \in (0,1)$ depending on δ, where, for each $\alpha \in (0,1)$, P_α denotes the tilted distribution,

$$P_\alpha(x) = \frac{P(x)^\alpha}{\sum_{y \in A} P(y)^\alpha}, \quad x \in A.$$

The first-order asymptotic relation (3) is simply the "rate version" of the well-known error-exponents result that was established in the early works [9], [13] and in the present form by Blahut [4]. The rate, $H(P_{\alpha^*})$ can be expressed as the inverse of the error-exponent function,

$$\Delta_P(R) := \inf_{P' : H(P') \geq R} D(P' \| P),$$

as

$$H(P_{\alpha^*}) = \Delta_P^{-1}(\delta) = \sup_{P' : D(P'\|P) \leq \delta} H(P'),$$

where $D(P'\|P)$ is the relative entropy between P' and P.

Recently, a nonasymptotic refinement of the evaluation of the fundamental compression limit R_n^* in the small excess-rate probability regime was obtained. It was shown in [31] that, as $n \to \infty$,

$$nR_n^*(2^{-n\delta}, P) = nH(P_{\alpha^*}) - \frac{1}{2(1-\alpha^*)} \log n + O(1) \text{ bits.} \tag{4}$$

Following [31], we call the expression

$$\mathcal{R}_n(\epsilon, P) := H(P_{\alpha^*}) - \frac{1}{2n(1-\alpha^*)} \log n \text{ bits/symbol,} \tag{5}$$

the best *pragmatic rate* that can be achieved at blocklength n with excess-rate probability no greater than $\epsilon = 2^{-n\delta}$.

Despite their technical nature, the approximations to R_n^* provided by all these different approaches are of as much practical relevance as they are of mathematical interest. In particular, each one is useful in different regimes of blocklength and excess-rate probability requirements. For example, the Strassen-style approximation suggested by (2) is only relevant for moderate values of ϵ. If ϵ is small, then the second term dominates and the approximation is no longer valid or useful.

In that case, the operational utility of (4) can be described as follows. Given the blocklength n and a small target excess-rate probability ϵ, we can compute $\delta = (1/n)\log(1/\epsilon)$ and solve for the corresponding α^*. Then the best achievable rate with these parameters is approximately $\mathcal{R}_n(\epsilon, P)$ bits/symbol. A detailed, explicit example demonstrating the superiority of this approximation for a certain range of parameter values is given in [31].

D. Universal compression at pragmatic rates

In this paper, we examine the best rate that can be *universally* achieved, by a single one-to-one compressor applied to any memoryless source $\boldsymbol{X}$ on a finite alphabet A of size m.

In the Shannon regime, Kosut and Sankar [19] showed that the same optimal performance as for known sources [17], [27] could be achieved universally by one-to-one codes, at a cost of approximately

$$\Big(\frac{m-2}{2}\Big) \log n \quad \text{bits,} \tag{6}$$

over all memoryless sources $\boldsymbol{X}$ on a given alphabet A of size $m = |A| \geq 2$. In Theorem 14 in Section IV, we state an analogous result in the error exponents regime. Namely, we describe a specific sequence of one-to-one compressors $\{\phi_n^*\}$ that simultaneously achieve excess-rate probability no greater than $2^{-n\delta}$ on any memoryless source with full support on A, at a rate no greater than:

$$H(P_{\alpha^*}) + \Big(\frac{m-2}{2} - \frac{1}{2(1-\alpha^*)}\Big)\frac{\log n}{n} + O\Big(\frac{1}{n}\Big) \text{ bits per symbol.} \tag{7}$$

Compared to $\mathcal{R}_n(\epsilon, P)$ in (5), this includes a universality penalty of $(\frac{m-2}{2})\log n$ bits, identical to the corresponding term in the optimal universal rate obtained in the Shannon regime, c.f. (6).

Interestingly, the code used to prove Theorem 14 is somewhat simpler than the compressors employed in [19]. The idea is to simply order all strings $x^n \in A^n$ in order of increasing empirical entropy, and take $\phi_n^*(x^n)$ to be the binary description of the index of x^n in this ordering.

On the other hand, in the present setting the analysis of their performance requires a very accurate combinatorial estimate on the number of strings with low empirical entropy. Specifically, in Theorem 8 of Section III we show that the number of $x^n \in A^n$ such that $H(\hat{P}_{x^n}) \leq H(Q)$ is:

$$\Theta\big(n^{\frac{m-3}{2}} 2^{nH(Q)}\big);$$

see Section III for precise definitions. The proof of this estimate is the most technical part of this paper, and the result itself may be of independent interest.

Although we do not provide a matching converse to the achievability statement in Theorem 14, we conjecture that the rate obtained there, given in (7), is optimal.

At first glance, it may be surprising that the cost of universality, both in the Shannon regime and in the error exponents regime, is $(\frac{m-2}{2})\log n$ bits, instead of the more familiar $(\frac{m-1}{2})\log n$ bits seen in the classical universal compression literature on compression redundancy. This discrepancy comes from the fact that, unlike, e.g., in Shtarkov's [26] or Rissanen's [24] well-known results, we consider one-to-one codes instead of prefix-free codes. In the case of a known source, a result of [17] shows that this makes a difference of at most one bit. But in the case of universal compression, the difference between optimal prefix-free and optimal one-to-one codes has been found to be of the order of $\log n$ bits, both in the Shannon excess-rate regime [19], and in terms of redundancy rates [3].

One way to see why there is indeed a difference in universal coding with one-to-one versus with prefix-free codes, is to consider binary memoryless sources. As described in [17], the optimal one-to-one code of any such source simply orders all strings of length n in decreasing probability, and assigns binary descriptions of increasing lengths to their indices lexicographically. But for binary $\mathrm{Bern}(p)$ sources there are only two optimal orderings, depending on whether p is greater or smaller than 1/2. Therefore, a universal one-to-one code only needs one extra bit to differentiate between these two cases.

On the other hand, the optimal prefix-free universal code needs to describe the maximum likelihood estimate of p based on x^n, to $\frac{1}{\sqrt{n}}$ accuracy, which takes approximately $\frac{1}{2}\log n$ bits; see, for example, the discussions in [1], [11], [16]. So there is an $O(\log n)$ universality penalty for prefix-free codes, whereas there is only an $O(1)$ universality penalty for one-to-one codes.

We do not consider the problem of prefix-free universal codes in the error exponents regime in this work, but we conjecture that, compared to the optimal rate in (4), they incur a penalty of, at best, $\frac{m}{2}\log n$ bits; cf. [19].

E. Prior related work

Extensive discussions of different aspects of the fundamental limits of lossless data compression can be found in Csiszár

and Körner's classic text, [6], Han's book on information spectrum methods [12], and Tan's monograph [28]. Universal lossless data compression also has a rich history. The texts [5], [8], [11], [25] and the papers [2], [24] provide comprehensive overviews.

All of our results are stated without their proofs; these can be found in the full version of the paper at [30].

II. Preliminaries and auxiliary results

Throughout, $\boldsymbol{X} = \{X_n\}$ denotes a memoryless source with distribution P that has full support on a finite alphabet A. For $\alpha \in (0,1)$, the tilted distribution P_α is given by

$$P_\alpha(x) = \frac{P(x)^\alpha}{Z_\alpha}, \quad x \in A, \tag{8}$$

where the normalising constant Z_α is

$$Z_\alpha = \sum_{x \in A} P(x)^\alpha.$$

The standard normal cumulative distribution function is denoted by Φ, 'log' denotes the logarithm taken to base 2, and the relative entropy between two p.m.f.s P, Q on the same finite alphabet A is $D(P\|Q) = \sum_{x\in A} P(x) \log[P(x)/Q(x)]$.

A. Normal approximation, compression and hypothesis testing

Here we collect some classical normal approximation inequalities and some simple lemmas on moments and derivatives of entropies and relative entropies, all of which are employed in the proof of our main universal compression result in Theorem 14.

First, we recall the classical Berry-Esséen bound [18], [22] for sums of i.i.d. random variables.

Theorem 1 (Berry-Esséen bound)**.** *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random variables with mean $\mu = \mathbb{E}(Z_1)$, variance $\sigma^2 = \mathrm{Var}(Z_1)$, and $\rho = \mathbb{E}(|Z_1 - \mu|^3) < \infty$. Then:*

$$\left| \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\Big(\frac{1}{n}\sum_{i=1}^n Z_i - \mu\Big) \le x \right) - \Phi(x) \right| \le \frac{\rho}{2\sigma^3\sqrt{n}},$$

for all $x \in \mathbb{R}$, $n \ge 1$.

The following is a simple corollary of Theorem 1, cf. [20], [21], [23].

Lemma 2. *Let $Z_1, Z_2, \ldots, Z_n$ be as in Theorem 1. Then:*

$$\mathbb{E}\left\{ \exp\Big(-\sum_{i=1}^n Z_i\Big) \mathbb{I}_{\{\sum_{i=1}^n Z_i \ge x\}} \right\} \le \Big(\frac{1}{\sqrt{2\pi}} + \frac{\rho}{\sigma^2}\Big)\frac{1}{\sigma\sqrt{n}} e^{-x},$$

for all $x \in \mathbb{R}$, $n \ge 1$.

B. Moments and derivatives identities

We close this section with some simple technical lemmas. Let X have p.m.f. P on the alphabet A, where P has full support on A, and recall the definition of the tilted distribution P_α in (8), for $\alpha \in (0,1)$. The proofs of the next five lemmas follow by direct computation; see for example, [7] and [20], [21].

Lemma 3. *Define the tilted second and third moments as:*

$$\begin{aligned}
\sigma^2_{2,\alpha} &= \mathrm{Var}_{P_\alpha}\Big(\log_e \frac{P_\alpha(X)}{P(X)}\Big),\\
\rho_{2,\alpha} &= \mathbb{E}_{P_\alpha}\Big|\log_e \frac{P_\alpha(X)}{P(X)} - \mathbb{E}_{P_\alpha}\Big(\log_e \frac{P_\alpha(X)}{P(X)}\Big)\Big|^3,\\
\sigma^2_{3,\alpha} &= \mathrm{Var}_{P_\alpha}(\log_e P(X)),\\
\rho_{3,\alpha} &= \mathbb{E}_{P_\alpha}|\log_e P(X) - \mathbb{E}_{P_\alpha}(\log_e P(X))|^3.
\end{aligned}$$

Then, for all $\alpha \in (0,1)$:

$$\begin{aligned}
\sigma^2_{2,\alpha} &= (1-\alpha)^2\sigma^2_{3,\alpha},\\
\rho_{2,\alpha} &= (1-\alpha)^3\rho_{3,\alpha}.
\end{aligned}$$

Lemma 4. *For all $\alpha \in (0,1)$ we have the following expressions for the derivatives of $D(P_\alpha\|P)$ and $\sigma^2_{3,\alpha}$:*

$$\begin{aligned}
\frac{dD(P_\alpha\|P)}{d\alpha} &= (\alpha-1)\sigma^2_{3,\alpha}(\log e),\\
\frac{d^2D(P_\alpha\|P)}{d\alpha^2} &= (\log e)\sigma^2_{3,\alpha} + (\log e)(\alpha-1)\frac{d\sigma^2_{3,\alpha}}{d\alpha},\\
\frac{d\sigma^2_{3,\alpha}}{d\alpha} &= \mathbb{E}_{P_\alpha}\Big[\Big(\log_e P(X) - \mathbb{E}_{P_\alpha}[\log_e P(X)]\Big)^3\Big].
\end{aligned}$$

Lemma 5. *$D(P_\alpha\|P)$ is a continuous, strictly decreasing function of α, for $\alpha \in (0,1)$.*

Lemma 6. *For all $\alpha \in (0,1)$ we have the following expressions for the derivatives of $H(P_\alpha)$:*

$$\begin{aligned}
\frac{dH(P_\alpha)}{d\alpha} &= -(\log e)\alpha\sigma^2_{3,\alpha},\\
\frac{d^2H(P_\alpha)}{d\alpha^2} &= -(\log e)\Big[\sigma^2_{3,\alpha} + \alpha\frac{d\sigma^2_{3,\alpha}}{d\alpha}\Big].
\end{aligned}$$

Lemma 7. *For all $\alpha \in (0,1)$, and any pair of p.m.f.s P, Q such that P has full support on A, we have:*

$$\begin{aligned}
&\alpha[D(Q\|P) - D(P_\alpha\|P)]\\
&\quad = D(Q\|P_\alpha) + (1-\alpha)[H(Q) - H(P_\alpha)].
\end{aligned}$$

III. The number of low-entropy types

In this section we state a fine combinatorial estimate on the number of stings whose empirical entropy is below a given threshold. This is important in the proof of the universality result we state in Section IV, and it may also be of independent interest.

Let A be a fixed finite alphabet of size $|A| = m \ge 2$. We write $\mathcal{P}$ for the space of p.m.f.s P on A and denote by $\mathcal{P}_n \subset \mathcal{P}$ the subset of all n-types on A; recall that $P \in \mathcal{P}$ is an n-type if, for each $a \in A$, $P(a)$ is of the form k/n for an

integer $0 \le k \le n$. For a string $x^n \in A^n$, we write $\hat{P}_{x^n}$ for the *type* of x^n, i.e., for the empirical p.m.f. it induces on A:

$$\hat{P}_{x^n}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{x_i=a\}}, \quad a \in A.$$

Clearly, $\hat{P}_{x^n}$ is always an n-type.

The first main result of this paper, stated next, gives an accurate estimate of the number of strings x^n whose types $\hat{P}_{x^n}$ have entropy no greater than a given threshold. Recall that the notation $a_n = \Theta(b_n)$ for two nonnegative sequences a_n and b_n means that there are finite nonzero constants c, c' such that $cb_n \le a_n \le c'b_n$ for all n.

Theorem 8 (Low-entropy types)**.** *Let Q be a non-uniform p.m.f. of full support on A, and for each $n \ge 1$ consider the set*

$$B_n = \big\{x^n \in A^n : H(\hat{P}_{x^n}) \le H(Q)\big\}. \tag{9}$$

Then the cardinality $|B_n|$ of B_n satisfies:

$$|B_n| = \Theta\Big(n^{\frac{m-3}{2}} 2^{nH(Q)}\Big), \tag{10}$$

where the implied constants depend on Q and m.

Theorem 8 follows immediately from the lower bound in Theorem 11 combined with the upper bound in Theorem 13, established in Section III-B. These results also provide additional information on how the implied constants in (10) depend on the problem parameters. In fact, the upper bound in Theorem 13 is shown to hold uniformly in a neighbourhood of Q.

A. Preliminaries

For an n-type $P \in \mathcal{P}_n$, we write

$$\mathcal{T}(P) = \{x^n \in A^n : \hat{P}_{x^n} = P\}$$

for the *type class* of P, and we denote the support of any $P \in \mathcal{P}$ as $\mathrm{supp}(P)$.

The following useful notation will be used throughout this section. For two expressions a and b that may depend on n, Q, and possibly on other parameters, we write $a \gtrsim_Q b$ to indicate that there exists a constant $K = K(Q)$ that depends on Q such that $a \ge K(Q) \cdot b$. Similarly, $a \approx b$ means $a \lesssim b$ and $b \lesssim a$. When no subscripts are given in $\gtrsim, \lesssim$ and $\approx$, the implied constants are absolute. The big-O and small-o notation is used as usual, and again we use subscripts to denote the dependence of the implied constants on the problem parameters. For example, for a positive function g, $f(n) = O_Q(g(n))$ means that there exists $K = K(Q)$ such that $|f(n)| \le K(Q)g(n)$ for all n large enough.

The derivations of Theorems 11 and 13 involve geometric considerations regarding certain subsets of $\mathcal{P}$. We will use the following terminology. A *convex body* is a convex, open, non-empty, and bounded subset of $\mathbb{R}^d$.

For a convex body C, we write $\mathrm{Vol}(C)$ and $\mathrm{Sur}(C)$ for its volume and surface area, respectively. The Minkowski sum of two subsets C_1, C_2 or $\mathbb{R}^d$ is $C_1 + C_2 = \{x + y : x \in C_1, y \in C_2\}$. Writing $B_2(\epsilon)$ for the closed Euclidean ball of radius $\epsilon > 0$ centered at the origin, the Minkowski-Steiner formula [10] says that the surface area of a convex body can be expressed as

$$\mathrm{Sur}(C) = \lim_{\epsilon \downarrow 0} \frac{\mathrm{Vol}(C + B_2(\epsilon)) - \mathrm{Vol}(C)}{\epsilon}.$$

Moreover, the volume $\mathrm{Vol}(C + B_2(\epsilon))$ is locally a polynomial of degree d in ϵ.

A *full-rank lattice* Γ in $\mathbb{R}^d$ is a discrete additive subgroup Γ of $\mathbb{R}^d$ generated by d linearly independent elements $v_1, \ldots, v_d$ of $\mathbb{R}^d$. The lemma below is a standard result on counting lattice points in a convex body; see [29, Lemma 3.22].

Lemma 9. *Let $\Gamma \subset \mathbb{R}^d$ be a full-rank lattice generated by $\{v_1, \ldots, v_d\}$, and let $B \subset \mathbb{R}^d$ be a convex body. Then for $R > 0$,*

$$|(R \cdot B) \cap \Gamma| = (R^d + O_{\Gamma,B,d}(R^{d-1}))\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(v_1 \wedge \ldots \wedge v_d)},$$

where $v_1 \wedge \ldots \wedge v_d$ denotes the parallelepiped with vertices $v_1, \ldots, v_d$, and the $O_{\Gamma,B,d}(R^{d-1})$ term is bounded in absolute value by KR^{d-1}, for some constant K that depends on Γ, B and d.

The next lemma is a standard result, which can be obtained by using the Robbins refinement of the Stirling approximation in the multinomial coefficient; see, e.g., [6].

Lemma 10 (Size of type classes)**.** *Let P be an n-type with full support on an alphabet of size k. Then*

$$|\mathcal{T}(P)| \approx_k \frac{2^{nH(P)}}{n^{(k-1)/2}} \prod_{a \in A} \frac{1}{\sqrt{P(a)}},$$

where the implied constants depend only on k.

B. Upper and lower bounds

First we establish a lower bound on the cardinality $|B_n|$ of the set B_n in (9).

Theorem 11. *Let Q be a non-uniform p.m.f. with full support on A. Then the set B_n in* (9) *satisfies,*

$$\big|\big\{x^n : H(\hat{P}_{x^n}) \le H(Q)\big\}\big| \gtrsim_{Q,m} n^{\frac{(m-3)}{2}} 2^{nH(Q)},$$

for all $n \ge N(Q, m)$, for an integer $N(Q, m)$ that depends only on Q and m.

A key element of the proof of Theorem 11 is the following estimate.

Lemma 12. *Under the assumptions of Theorem 11, we have*

$$\Big|\Big\{P \in \mathcal{P}_n : H(P) \in \Big[H(Q) - \frac{1}{n}, H(Q)\Big]\Big\}\Big| \gtrsim_{Q,m} n^{m-2} + n^{m-3},$$

for all $n \ge N'(Q, m)$, for an integer $N'(Q, m)$ that depends only on Q and m.

Next we establish a matching upper bound. In fact, we prove a stronger result, where the polynomial prefactor, the implied

constants, and the value of n beyond which the bound holds are the same for all Q' in a neighbourhood of Q.

Theorem 13. *Let Q be a non-uniform p.m.f. of full support on A. Suppose $\{Q_n\}_{n\geq 1}$ is a sequence of p.m.f.s on A, such that $Q_n \to Q$ as $n \to \infty$. Then*

$$\big|\{x^n \in A^n : \ H(\hat{P}_{x^n}) \leq H(Q_n)\}\ \big| \lesssim_{Q,m} \ n^{\frac{m-3}{2}}\, 2^{nH(Q_n)},$$

for all $n \geq n_0(Q,m)$, where $n_0(Q,m)$ depends only on Q and m.

IV. Universal compression

The second main result in this paper is the following.

Theorem 14 (Universal achievability)**.** *Consider all memoryless sources $\boldsymbol{X}$ on a finite alphabet A of size $|A| = m \geq 2$. There exists a sequence of universal, one-to-one, variable-rate compressors $\{\phi_n^*\}$ such that, on any source with distribution P with full support on A, and for any $\delta \in (0, D(U\|P))$, the compressors $\{\phi_n^*\}$ simultaneously achieve excess-rate probability,*

$$\mathbb{P}\big(\ell(\phi_n^*(X^n)) \geq nR^{\mathsf{u}}(n,\delta,P)\big) \leq 2^{-n\delta} \quad \textit{eventually}, \quad (11)$$

at a rate $R^{\mathsf{u}}(n,\delta,P)$ that satisfies, as $n \to \infty$,

$$nR^{\mathsf{u}}(n,\delta,P) \leq nH(P_{\alpha^*}) + \Big(\frac{m-2}{2} - \frac{1}{2(1-\alpha^*)}\Big)\log n + O(1), \quad (12)$$

where α^* is the unique $\alpha \in (0,1)$ for which $\delta = D(P_{\alpha^*}\|P)$, the p.m.f. P_α is defined in (8)*, and the implied constant in the $O(1)$ term depends on m, P and δ.*

Remark. An examination of the proof of Theorem 14 shows that the excess-rate probability bound in (11) holds for all n greater than some $M_0(P,\delta)$, which can be explicitly identified in terms of P and δ in a similar way to how the corresponding expressions were derived in the case of a known source distribution in [31]. But in the present context, the exact value of $M_0(P,\delta)$ is less relevant, since our bound on the rate in (12) is already an asymptotic statement.

Acknowledgment

This work was supported in part by the EPSRC-funded INFORMED-AI project EP/Y028732/1.

References

[1] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, July 2002.

[2] A.R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, October 1998. Information theory: 1948–1998.

[3] A. Beirami and F. Fekri. Fundamental limits of universal lossless one-to-one compression of parametric sources. In *2014 IEEE Information Theory Workshop (ITW)*, pages 212–216, Hobart, Australia, November 2014.

[4] R.E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inform. Theory*, 20(4):405–417, July 1974.

[5] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, New York, NY, second edition, 2012.

[6] I. Csiszár and J. Körner. *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, Cambridge, U.K., second edition, 2011.

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

[8] I. Csiszár and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, December 2004.

[9] R.L. Dobrushin. Asymptotic bounds of the probability of error for the transmission of messages over a discrete memoryless channel with a symmetric transition probability matrix. *Teor. Veroyatnost. i Primenen*, 7:283–311, 1962.

[10] H. Federer. *Geometric measure theory*. Springer-Verlag, Berlin, Heidelberg, 2014.

[11] P. Grünwald. *The minimum description length principle*. M.I.T. Press, Cambridge, MA, 2007.

[12] T.S. Han. *Information-spectrum methods in information theory*. Springer, Berlin, 2003.

[13] F.K. Jelinek. *Probabilistic information theory: Discrete and memoryless models*. McGraw-Hill, New York, NY, 1968.

[14] I. Kontoyiannis. Second-order analysis of lossless and lossy versions of Lempel-Ziv codes. In *31st Asilomar Conference on Signals, Systems and Computers*, Monterey, CA, November 1997.

[15] I. Kontoyiannis. Second-order noiseless source coding theorems. *IEEE Trans. Inform. Theory*, 43(4):1339–1341, July 1997.

[16] I. Kontoyiannis. Context-tree weighting and Bayesian Context Trees: Asymptotic and non-asymptotic justifications. *IEEE Trans. Inform. Theory*, 70(2):1204–1219, February 2024.

[17] I. Kontoyiannis and S. Verdú. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inform. Theory*, 60(2):777–795, February 2014.

[18] V.Yu. Korolev and I.G. Shevtsova. On the upper bound for the absolute constant in the Berry–Esséen inequality. *Theory Probab. Appl.*, 54(4):638–658, 2010.

[19] O. Kosut and L. Sankar. Asymptotics and non-asymptotics for universal fixed-to-variable source coding. *IEEE Trans. Inform. Theory*, 63(6):3757–3772, June 2017.

[20] V. Lungu and I. Kontoyiannis. Finite-sample expansions for the optimal error probability in asymmetric binary hypothesis testing. *arXiv e-prints*, `2404.09605 [cs.IT]`, April 2024.

[21] V. Lungu and I. Kontoyiannis. The optimal finite-sample error probability in asymmetric binary hypothesis testing. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 831–836, Athens, Greece, July 2024.

[22] V.V. Petrov. *Limit theorems of probability theory*. The Clarendon Press, Oxford University Press, New York, NY, 1995.

[23] Y. Polyanskiy, H.V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory*, 56(5):2307–2359, May 2010.

[24] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, 30(4):629–636, July 1984.

[25] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific, Singapore, 1989.

[26] J. Shtarkov. Coding of discrete sources with unknown statistics. In *Topics in Information Theory* (I. Csiszár and P. Elias, eds.), Coll. Math. Soc. J. Bolyai, no. 16, North-Holland Amsterdam, pages 559–574, 1977.

[27] V. Strassen. Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Information Theory, Statist. Decision Functions, Random Processes (Liblice, 1962)*, pages 689–723. Publ. House Czech. Acad. Sci., Prague, 1964.

[28] V.Y.F. Tan. Asymptotic estimates in information theory with non-vanishing error probabilities. *Foundations and Trends in Communications and Information Theory*, 11(1-2):1–184, September 2014.

[29] T. Tao and V. Vu. *Additive combinatorics*. Cambridge University Press, Cambridge, U.K., 2006.

[30] A. Theocharous, L. Gavalakis, and I. Kontoyiannis. Pragmatic lossless compression: Fundamental limits and universality. *arXiv e-prints*, `2501.10103 [cs.IT]`, November 2025.

[31] A. Theocharous and I. Kontoyiannis. Lossless data compression at pragmatic rates. In *2025 IEEE International Symposium on Information Theory (ISIT)*, Ann Arbor, MI, June 2025.