Lossless Data Compression with Side Information: Nonasymptotics and Dispersion

Lampros Gavalakis and Ioannis Kontoyiannis
University of Cambridge
lg560@cam.ac.uk
i.kontoyiannis@eng.cam.ac.uk

Abstract—**The problem of lossless data compression with side information available to both the encoder and the decoder is considered. The finite-blocklength fundamental limits of the best achievable performance are defined, in two different versions of the problem: *Reference-based compression*, when a single side information string is used repeatedly in compressing different source messages, and *pair-based compression*, where a different side information string is used for each source message. General achievability and converse theorems are established. Nonasymptotic normal approximation expansions are proved for the optimal rate with memoryless sources, in both the reference-based and pair-based settings. These are stated in terms of explicit, finite-blocklength bounds, that are tight up to third-order terms. Extensions that go significantly beyond the class of memoryless sources are obtained. The relevant source dispersion is identified and its relationship with the conditional varentropy rate is established. Interestingly, the dispersion is different in reference-based and pair-based compression, and it is proved that the reference-based dispersion is in general smaller.**

I. Introduction

It has long been recognised in information theory [5] that the presence of correlated side information can dramatically improve compression performance.

Reference-based compression. A particularly important and timely application of compression with side information is to the problem of storing the vast amounts of genomic data currently being generated by modern DNA sequencing technology [6, 11]. In a typical scenario, the genome X of a new individual that needs to be stored is compressed using a reference genome Y as side information. Since most of the time X will only be a minor variation of Y, the potential compression gains are large. An important aspect of this scenario is that the same side information – in this case the reference genome Y – is used in the compression of many new sequences $X^{(1)}, X^{(2)}, \ldots$. We call this the *reference-based* version of the compression problem.

Pair-based compression. Another important application of compression with side information is to the problem of file synchronization [16], where updated computer files need to be stored along with their earlier versions, and the related problem of software updates [14], where remote users need to be provided with newer versions of possibly large software suites. Unlike genomic compression, in these cases a different side information sequence Y (namely, the older version of the specific file or of the particular software) is used every time a new piece of data X is compressed. We refer to this as the *pair-based* version of the compression problem, since each time a different (X, Y) pair is considered.

A number practical algorithms for compression with side information have been developed over the past 25 years, including the following. The most common approach is based on generalisations of the celebrated family of Lempel-Ziv compression methods [8, 13, 17]. Turbo codes were used in [1], and a generalization of the context-tree weighting algorithm was developed in [3]. For applications in image and video compression, see, e.g., [2, 12].

A. Outline of main results

All our results are stated here without their proofs; those can be found in the full version of this paper, available online [7].

In Section II we give precise definitions for the finite-blocklength fundamental limits of reference-based and pair-based compression, and we identify the theoretically optimal one-to-one compressor in each case. Throughout, we consider an arbitrary source-side information pair $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_n, Y_n)\ ;\ n \geq 1\}$, with values in the finite alphabets $\mathcal{X}, \mathcal{Y}$, respectively, where $\boldsymbol{X}$ is the source to be compressed and $\boldsymbol{Y}$ is the side information process. In Theorem 2.3 we show that, for any blocklength n, requiring the compressor to be prefix-free imposes a penalty of no more than $1/n$ bits per symbol on the optimal rate.

In Section III we state and prove four general, single-shot, achievability and converse results, for the compression of arbitrary sources with arbitrarily distributed side information.

Sections IV and V contain our main results, giving non-asymptotic, normal-approximation expansions to the optimal reference-based and pair-based rates. For the sake of clarity, we first describe the pair-based results of Section V.

Let $R^*(n, \epsilon)$ be the best pair-based compression rate that can be achieved at blocklength n, with excess rate probability no greater than ϵ. For a memoryless source-side information pair $(\boldsymbol{X}, \boldsymbol{Y})$ in Theorems 5.1 and 5.2 we show that there are explicit constants $C, C' > 0$ and $n_0 \geq 1$, such that, for all $n \geq n_0$,

$$R^*(n, \epsilon) = H(X|Y) + \frac{1}{\sqrt{n}} \sigma(X|Y) Q^{-1}(\epsilon) - \frac{\log n}{2n} + \frac{\Delta}{n}, \quad (1)$$

where $-C \leq \Delta \leq C'$.

[0] L.G. was supported in part by EPSRC grant number RG94782. I.K. was supported in part by a grant from the Hellenic Foundation for Research and Innovation.

Here, Q denotes the standard Gaussian tail function, $Q(z) = 1 - \Phi(z)$, $z \in \mathbb{R}$, $H(X|Y) = H(X_1|Y_1)$ is the conditional entropy of X_1 given Y_1, and,

$$\sigma^2(X|Y) = \text{Var}\big(-\log P(X_1|Y_1)\big),$$

is the *conditional varentropy* of X_1 given Y_1. This generalizes the *minimal coding variance* of [9]. Throughout, $\log = \log_2$ and all information-theoretic quantities are expressed in bits.

The bounds in (1) generalize the corresponding no-side-information results in Theorems 17 and 18 of [10]. Our proofs rely on the general coding theorems of Section II combined with appropriate versions of the classical Berry-Esséen bounds. An important difference with [10] is that the approximation used in the proof of the upper bound in [10, Theorem 17] does not admit a natural analog in the case of compression with side information. Instead, we use the tight approximation to the description lengths of the optimal compressor given in Theorem 3.4. Results analogous to (1) are also established in a slightly weaker form for the case of Markov sources [7].

In Section IV we consider the the reference-based setting. For any source-side information pair $(\boldsymbol{X}, \boldsymbol{Y})$, we write x_i^j and X_1^j for the string $(x_i, x_{i+1}, \ldots, x_j)$ with values in $\mathcal{X}$ and the block of random variables $(X_i, X_{i+1}, \ldots, X_j)$, respectively; and similarly for y_i^j and Y_i^j.

Given a fixed side information string $y_1^n \in \mathcal{Y}^n$, let $R^*(n, \epsilon|y_1^n)$ denote the best pair-based compression rate that can be achieved at blocklength n, conditional on $Y_1^n = y_1^n$, with excess-rate probability no greater than ϵ. Suppose that the distribution of $\boldsymbol{Y}$ is arbitrary, and that the source $\boldsymbol{X}$ is conditionally i.i.d. (independent and identically distributed) given $\boldsymbol{Y}$. In Theorems 4.4 and 4.5 we prove that there are explicit finite constants $C(y_1^n), C'(y_1^n) > 0$ and $n_0(y_1^n) \geq 1$, such that, for all $n \geq n_0(y_1^n)$, we have,

$$\begin{aligned} &R^*(n, \epsilon|y_1^n) \\ &\quad = H_n(X|y_1^n) + \frac{1}{\sqrt{n}}\sigma_n(y_1^n)Q^{-1}(\epsilon) - \frac{\log n}{2n} + \frac{\Delta(y_1^n)}{n}, \end{aligned} \quad (2)$$

where now the first-order rate is given by,

$$H_n(X|y_1^n) = \frac{1}{n}\sum_{i=1}^n H(X|Y = y_i),$$

the variance $\sigma_n^2(y_1^n)$ is,

$$\sigma_n^2(y_1^n) = \frac{1}{n}\sum_{i=1}^n \text{Var}\big(-\log P(X|y_i)\big|Y = y_i\big),$$

and $-C'(y_1^n) \leq \Delta(y_1^n) \leq C(y_1^n)$. A numerical example illustrating the accuracy of the normal approximation in (2) is shown in Figure 1.

Note the elegant analogy between (1) and (2). Indeed, there is further asymptotic solidarity in the normal approximation of the two cases. If $\boldsymbol{Y}$ is ergodic, then for a random side information string Y_1^n we have, $H_n(X|Y_1^n) \to H(X|Y)$, as $n \to \infty$, with probability 1, but the corresponding variances are different: With probability 1,

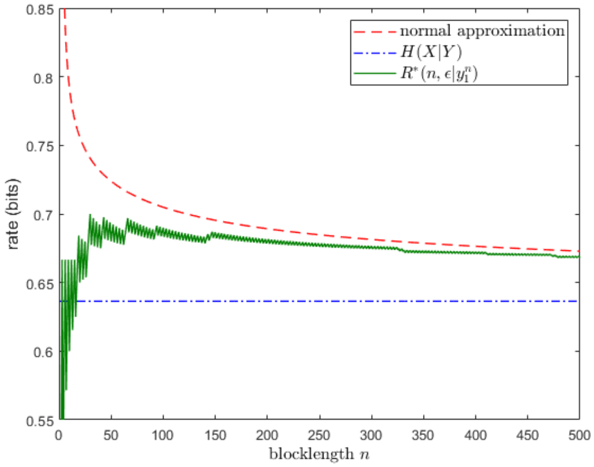

Fig. 1: Normal approximation to the reference-based optimal rate $R^*(n, \epsilon|y_1^n)$ for a memoryless side information process $\{Y_n\}$ with Bern(1/3) distribution. The source $\{X_n\}$ has $X|Y = 0 \sim$ Bern(0.1) and $X|Y = 1 \sim$ Bern(0.6). The conditional entropy rate $H(X|Y) \approx 0.636$, whereas the entropy rate of the source is $H(X) \approx 0.837$. The side information sequence is taken to be $y_1^n = 001001001\cdots$. The graph shows $R^*(\epsilon, n|y_1^n)$ itself, with $\epsilon = 0.1$, for blocklengths $1 \leq n \leq 500$, together with the normal approximation to $R^*(\epsilon, n|y_1^n)$ given by the first three terms in the right-hand side of (2).

$$\sigma_n^2(Y_1^n) \to E\Big[\text{Var}\big(-\log P(X|Y)|Y\big)\Big], \qquad \text{as } n \to \infty,$$

which is shown in Proposition 4.3 to be strictly smaller than $\sigma^2(X|Y)$ in general. This admits the intuitively satisfying interpretation that, in reference-based compression, where a single side information string is used to compress multiple source messages, the optimal rate has smaller variability.

Although the approximation bounds in (2) do not directly relate the optimal finite-n rate $R^*(n, \epsilon|y_1^n)$ to the optimal asymptotic rate $H(X|Y)$, in Theorem 4.6 we give an explicit, finite-n concentration bound, showing that, for a random side information string Y_1^n, the probability that $R^*(n, \epsilon|Y_1^n)$ exceeds $H(X|Y)$ by any $\delta > 0$, is exponentially small in n.

Finally, in Section VI we examine the pair-based dispersion $D(\boldsymbol{X}|\boldsymbol{Y})$, defined as the limiting normalised variance of the optimal description lengths of $\boldsymbol{X}$ given $\boldsymbol{Y}$, and the reference-based dispersion $D(\boldsymbol{X}|\boldsymbol{y})$ similarly defined for a given side information sequence $\boldsymbol{y} = y_1^\infty$. Theorem 6.2 states that, under general conditions, the pair-based dispersion $D(\boldsymbol{X}|\boldsymbol{Y})$ is equal to the conditional varentropy rate $\sigma^2(\boldsymbol{X}|\boldsymbol{Y})$, defined in (3), and relates $D(\boldsymbol{X}|\boldsymbol{Y})$ to the behaviour of the pair-based optimal rate $R^*(n, \epsilon)$ as $n \to \infty$ and $\epsilon \to 0$. Analogous results for the reference-based dispersion $D(\boldsymbol{X}|\boldsymbol{y})$ are established in Theorem 6.3.

B. Related work

The is a direct relationship between the present development and current work on the Slepian-Wolf (SW) problem. Tan and Kosut [15] give a second-order multidimensional normal approximation to the SW region for memoryless sources. Chen *et al.* [4] refined the results of [15] by establishing inner and outer asymptotic bounds for the SW region, which are tight up to and including third-order terms. Since, by definition, any SW code is also a pair-based code for our setting, the achievability result from [4] implies a slightly weaker form of our Theorem 5.1, with an asymptotic $O(1/n)$ term in place of the explicit C/n in (6). It is interesting that this level of accuracy can be derived both by random coding as in [4] and by deterministic methods as in Theorem 5.1.

II. Fundamental Limits

Let $(\boldsymbol{X},\boldsymbol{Y})=\{(X_n,Y_n)\ ;\ n\geq 1\}$ be an arbitrary source-side information pair with finite alphabets $\mathcal{X},\mathcal{Y}$, respectively. Given a source string $x_1^n\in\mathcal{X}^n$ and assuming $y_1^n\in\mathcal{Y}^n$ is available to both the encoder and decoder, a *fixed-to-variable one-to-one compressor with side information*, of blocklength n, is a collection of functions f_n, where each $f_n(x_1^n|y_1^n)$ takes a value in the set of all finite-length binary strings, $\{0,1\}^*$. For each $y_1^n\in\mathcal{Y}^n$, $f_n(\cdot|y_1^n)$ is assumed to be an injective function from $\mathcal{X}^n$ to $\{0,1\}^*$, so that the compressed string $f_n(x_1^n|y_1^n)$ is always uniquely and correctly decodable. The associated description lengths of $\{f_n\}$ are,

$$\ell(f_n(x_1^n|y_1^n))=\text{length of } f_n(x_1^n|y_1^n),\qquad \text{bits},$$

where $\ell(s)$ denotes the length, in bits, of a binary string s.

The following fundamental limits describe the best achievable performance among one-to-one compressors with side information.

Definition 2.1 (Reference-based optimal rate $R^*(n,\epsilon|y_1^n)$): For any blocklength n, any fixed side information string $y_1^n\in\mathcal{Y}^n$, and any $\epsilon\in[0,1)$, let $R^*(n,\epsilon|y_1^n)$ denote the smallest compression rate that can be achieved with excess-rate probability no larger than ϵ. Formally, $R^*(n,\epsilon|y_1^n)$ is the infimum among all $R>0$ such that,

$$\min_{f_n(\cdot|y_1^n)}\mathbb{P}\left[\ell(f_n(X_1^n|y_1^n))>nR|Y_1^n=y_1^n\right]\leq\epsilon,$$

where the minimum is over all one-to-one compressors $f_n(\cdot|y_1^n):\mathcal{X}^n\to\{0,1\}^*$.

Definition 2.2 (Pair-based optimal rate $R^*(n,\epsilon)$): For any blocklength n and any $\epsilon\in[0,1)$, let $R^*(n,\epsilon)$ denote the smallest compression rate that can be achieved with excess-rate probability no larger than ϵ. Formally, $R^*(n,\epsilon)$ is the infimum among all $R>0$ such that,

$$\min_{f_n}\mathbb{P}\left[\ell(f_n(X_1^n|Y_1^n))>nR\right]\leq\epsilon,$$

where the minimum is over all one-to-one compressors f_n with side information.

The optimal compressor f_n^*. It is easy to see from Definitions 2.1 and 2.2 that, in both cases, the minimum is achieved by the same simple compressor f_n^*: For each side information string y_1^n, $f_n^*(\cdot|y_1^n)$ is the optimal compressor for the distribution $\mathbb{P}(X_1^n=\cdot|Y_1^n=y_1^n)$, namely, the compressor that orders the strings x_1^n in order of decreasing probability $\mathbb{P}(X_1^n=x_1^n|Y_1^n=y_1^n)$, and assigns them codewords from $\{0,1\}^*$ in lexicographic order; cf. Property 1 in [10].

Prefix-free compressors. Let $R_p^*(n,\epsilon|y_1^n)$ and $R_p^*(\epsilon,n)$ be the corresponding fundamental limits as those in Definitions 2.1 and 2.2, when the compressors are required to be prefix-free. As it turns out, the prefix-free condition imposes a penalty of at most $1/n$ on the rate.

Theorem 2.3: For all $n\geq 1$ and any $0\leq\epsilon<1$:

$$R^*(n,\epsilon)\leq R_p^*(n,\epsilon)\leq R^*(n,\epsilon)+\frac{1}{n},$$
$$R^*(n,\epsilon|y_1^n)\leq R_p^*(n,\epsilon|y_1^n)\leq R^*(n,\epsilon|y_1^n)+\frac{1}{n}.$$

III. Coding Theorems for Arbitrary Sources

Consider two arbitrary discrete random variables (X,Y), with joint (PMF) $P_{X,Y}$, taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. For the sake of simplicity we may assume, without loss of generality, that the source alphabet $\mathcal{X}$ is the set of natural numbers $\mathcal{X}=\mathbb{N}$, and that, for each $y\in\mathcal{Y}$, the values of X are ordered with nonincreasing conditional probabilities given y, so that $\mathbb{P}(X=x|Y=y)$ is nonincreasing in x, for each $y\in\mathcal{Y}$.

Let $f^*=f_1^*$ be the optimal compressor described in the last section, and write P_X and $P_{X|Y}$ for the PMF of X and the conditional PMF of X given Y, respectively. The ordering of the values of X implies that, for all $x\in\mathcal{X},y\in\mathcal{Y}$,

$$\ell(f^*(x|y))=\lfloor\log x\rfloor.$$

The following is a general achievability result that applies to both the reference-based and the pair-based versions of the compression problem:

Theorem 3.1: For all $x\in\mathcal{X}$, $y\in\mathcal{Y}$,

$$\ell(f^*(x|y))\leq-\log P_{X|Y}(x|y),$$

and for any $z\geq 0$,

$$\mathbb{P}\big[\ell(f^*(X|Y))\geq z\big]\leq\mathbb{P}\big[-\log P_{X|Y}(X|Y)\geq z\big].$$

The next two theorems give general converse results for the pair-based compression problem:

Theorem 3.2: For any integer $k\geq 0$ and any $\tau>0$:

$$\begin{aligned}&\mathbb{P}\left[\ell(f^*(X|Y)\geq k)\right]\\&\geq\sup_{\tau>0}\left\{\mathbb{P}\left[-\log P_{X|Y}(X|Y)\geq k+\tau\right]-2^{-\tau}\right\}.\end{aligned}$$

Theorem 3.3: For any compressor f and any $\tau>0$:

$$\mathbb{P}\big[\ell(f(X|Y))\leq-\log P_{X|Y}(X|Y)-\tau\big]\leq 2^{-\tau}(\lfloor\log|\mathcal{X}|\rfloor+1).$$

Our next result is one of the main tools in the proofs of the achievability results in the normal approximation bounds for $R^*(n,\epsilon|y_1^n)$ and $R^*(n,\epsilon)$.

Theorem 3.4: For all x,y, $\ell(f^*(x|y))$ is bounded below by,

$$\log\left(\mathbb{E}\left[\frac{1}{P_{X|Y}(X|y)}\mathbb{I}_{\{P_{X|Y}(X|y)>P_{X|Y}(x|y)\}}\middle|Y=y\right]\right)-1,$$

and bounded above by,

$$\log\left(\mathbb{E}\left[\frac{1}{P_{X|Y}(X|y)}\mathbb{I}_{\{P_{X|Y}(X|y)\geq P_{X|Y}(x|y)\}}\middle|Y=y\right]\right),$$

where $\mathbb{I}_A$ denotes the indicator function of an event A, with $\mathbb{I}_A=1$ when A occurs and $\mathbb{I}_A=0$ otherwise.

IV. Normal Approximation for $R^*(n,\epsilon|y_1^n)$

A. Preliminaries: General sources

Recall that, for any source-side information pair $(\boldsymbol{X},\boldsymbol{Y})$, the *conditional entropy rate* $H(\boldsymbol{X}|\boldsymbol{Y})$ is defined as:

$$H(\boldsymbol{X}|\boldsymbol{Y}) = \limsup_{n\to\infty} \frac{1}{n} H(X_1^n|Y_1^n), \qquad \text{bits/symbol.}$$

If $(\boldsymbol{X},\boldsymbol{Y})$ are jointly stationary, then the above $\limsup$ is in fact a limit [5].

Definition 4.1: For a source-side information pair $(\boldsymbol{X},\boldsymbol{Y})$, the *conditional varentropy rate* is:

$$\sigma^2(\boldsymbol{X}|\boldsymbol{Y}) = \limsup_{n\to\infty} \frac{1}{n} \text{Var}(-\log P(X_1^n|Y_1^n)). \tag{3}$$

Lemma 4.2: For a broad class of jointly stationary and ergodic source-side information pairs $(\boldsymbol{X},\boldsymbol{Y})$ that include: (a) Markov chains $(\boldsymbol{X},\boldsymbol{Y})$ with all positive transition probabilities, and (b) irreducible and aperiodic, kth order Markov chains $(\boldsymbol{X},\boldsymbol{Y})$ such that $\boldsymbol{Y}$ is also a kth order Markov chain, the $\limsup$ in (3) is a limit, and:

$$\sigma^2(\boldsymbol{X}|\boldsymbol{Y}) = \lim_{n\to\infty} \frac{1}{n} \text{Var}\left(-\log\Big(\frac{P(X_1^n, Y_1^n | X_{-\infty}^0, Y_{-\infty}^0)}{P(Y_1^n|Y_{-\infty}^0)} \Big) \right).$$

A characterization of the case when the conditional varentropy rate is zero for memoryless sources and Markov chains is given in [7].

B. Preliminaries: Conditionally-i.i.d. sources

Suppose the source and side information, $(\boldsymbol{X},\boldsymbol{Y})$, consist of independent and identically distributed (i.i.d.) pairs $\{(X_n,Y_n)\}$, or, more generally, that $(\boldsymbol{X},\boldsymbol{Y})$ is a *conditionally-i.i.d. source-side information pair*, i.e., that the distribution of $\boldsymbol{Y}$ is arbitrary, and for each n, given $Y_1^n = y_1^n$, the random variables X_1^n are conditionally i.i.d.,

$$\mathbb{P}(X_1^n = x_1^n | Y_1^n = y_1^n) = \prod_{i=1}^n P_{X|Y}(x_i|y_i),\ x_1^n \in \mathcal{X}, y_1^n \in \mathcal{Y}.$$

For any $y \in \mathcal{Y}$, we write, $H(X|y)$ for the entropy of the conditional distribution of X given $Y = y$, namely, $-\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log P_{X|Y}(x|y)$, and,

$$V(y) = \text{Var}[-\log P_{X|Y}(X|y)|Y = y]. \tag{4}$$

We also write $\hat{H}_X(Y)$ for the random variable,

$$\hat{H}_X(Y) = -\sum_{x\in\mathcal{X}} P_{X|Y}(x|Y)\log P_{X|Y}(x|Y).$$

Recall the definitions of $H_n(X|y_1^n)$ and $\sigma_n^2(y_1^n)$ in the Introduction.

Proposition 4.3: Suppose $(\boldsymbol{X},\boldsymbol{Y})$ is an i.i.d. source-side information pair, with each $(X_n,Y_n) \sim (X,Y)$. Then, the conditional varentropy $\sigma^2(X|Y) = \text{Var}(-\log P(X|Y))$ can also be expressed:

$$\sigma^2(X|Y) = \mathbb{E}[V(Y)] + \text{Var}[\hat{H}_X(Y)].$$

C. Direct and converse bounds

Theorem 4.4 (Converse for $R^*(n,\epsilon|y_1^n)$): Suppose $(\boldsymbol{X},\boldsymbol{Y})$ is a conditionally-i.i.d. source-side information pair. For any $0 < \epsilon < \frac{1}{2}$, the reference-based optimal compression rate,

$$R^*(n,\epsilon|y_1^n) \geq H_n(X|y_1^n) + \frac{\sigma_n(y_1^n)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log n}{2n} - \frac{1}{n}\eta(y_1^n),$$

for all,

$$n > \frac{(1 + 6m_3\sigma_n^{-3}(y_1^n))^2}{4\big(Q^{-1}(\epsilon)\phi(Q^{-1}(\epsilon))\big)^2},$$

and any side information string y_1^n such that $\sigma_n^2(y_1^n) > 0$, where ϕ is the standard normal density,

$$m_3 = \max_{y\in\mathcal{Y}} \mathbb{E}[|-\log P(X|y) - H(X|y)|^3], \tag{5}$$

$$\text{and}\ \ \eta(y_1^n) = \frac{\sigma_n^3(y_1^n) + 6m_3}{\phi(Q^{-1}(\epsilon))\sigma_n^2(y_1^n)}.$$

By the definitions, Theorem 4.4 obviously also holds for prefix-free codes.

Next we derive an upper bound to $R^*(n,\epsilon|y_1^n)$ that matches the lower bound in Theorem 4.4 up to and including the third-order term. Note that, in view of Theorem 2.3, the result of Theorem 4.5 also holds for prefix-free codes, with $R_p^*(n,\epsilon|y_1^n)$ and $\zeta_n(y_1^n) + 1$ in place of $R^*(n,\epsilon|y_1^n)$ and $\zeta_n(y_1^n)$, respectively.

Theorem 4.5 (Achievability for $R^*(n,\epsilon|y_1^n)$): Let $(\boldsymbol{X},\boldsymbol{Y})$ be a conditionally-i.i.d. source-side information pair. For any $0 < \epsilon \leq \frac{1}{2}$, the reference-based optimal compression rate,

$$R^*(n,\epsilon|y_1^n) \leq H_n(X|y_1^n) + \frac{\sigma_n(y_1^n)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log n}{2n} + \frac{1}{n}\zeta_n(y_1^n),$$

for all $n > 36m_3^2/[\epsilon^2\sigma_n^6(y_1^n)]$, and any side information string y_1^n such that $\sigma_n^2(y_1^n) > 0$, where m_3 is given in (5), and,

$$\begin{aligned} \zeta_n(y_1^n) \ &= \ \frac{6m_3}{\sigma_n^3(y_1^n)\phi\Big(\Phi^{-1}\Big(\Phi(Q^{-1}(\epsilon)) + \frac{6m_3}{\sqrt{n}\sigma_n^3(y_1^n)}\Big)\Big)} \\ &\quad + \log\Big(\frac{\log e}{\sqrt{2\pi\sigma_n^2(y_1^n)}} + \frac{12m_3}{\sigma_n^3(y_1^n)} \Big). \end{aligned}$$

Although the normal approximation description of $R^*(n,\epsilon|y_1^n)$ in the last two theorems is quite detailed and accurate up to and including the third-order term, both the rate $H_n(X|y_1^n)$ and the "dispersion" $\sigma_n^2(y_1^n)$ depend on y_1^n, so it is natural to ask if anything can be said about the behaviour of $R^*(n,\epsilon|y_1^n)$ for "typical" y_1^ns. In Theorem 4.6 we show that if the source-side information pair $(\boldsymbol{X},\boldsymbol{Y})$ is i.i.d., then $R^*(n,\epsilon|Y_1^n) \approx H(X|Y)$ with high probability.

Theorem 4.6 (Concentration for $R^*(n,\epsilon|Y_1^n)$): Suppose $(\boldsymbol{X},\boldsymbol{Y})$ is an i.i.d. source-side information pair. Then, for any $0 < \epsilon \leq 1/2$, and any $\delta > 0$, we have,

$$\begin{aligned} &\mathbb{P}[R^*(n,\epsilon|Y_1^n) \geq H(X|Y) + \delta] \\ &\leq \exp\left\{ -\frac{n}{32}\left[\min\Big\{ \frac{\bar{v}}{v^*}, \frac{\delta}{\log|\mathcal{X}|} \Big\}^2 - \frac{32\log_e 2}{n} \right] \right\}, \end{aligned}$$

where $\bar{v} = \mathbb{E}[V(Y)]$ and $v^* = \max_y V(y)$, with V defined in (4), for all n greater than an explicitly identified n_1, depending on δ, ϵ, and the distribution of $(\boldsymbol{X},\boldsymbol{Y})$.

V. Normal Approximation for $R^*(n,\epsilon)$

Theorem 5.1 (Achievability for $R^*(n,\epsilon)$): Let $(\boldsymbol{X},\boldsymbol{Y})$ be an i.i.d. source-side information pair, with conditional varentropy rate $\sigma^2 = \sigma^2(X|Y) > 0$. For any $0 < \epsilon \leq \frac{1}{2}$, the pair-based optimal compression rate satisfies,

$$R^*(n,\epsilon) \leq H(X|Y) + \frac{\sigma(X|Y)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log n}{2n} + \frac{C}{n}, \quad (6)$$

for all,

$$n > \frac{4\sigma^2}{B^2\phi(Q^{-1}(\epsilon))^2} \times \left[\frac{B^2}{2\sqrt{2\pi e}\sigma^2} + \frac{\psi^2}{(1-\frac{1}{2\pi})^2\bar{v}^2}\right]^2,$$

where $\bar{v} = \mathbb{E}[V(Y)]$ and $\psi^2 = \mathrm{Var}(V(Y))$, with,

$$C = \log\Big(\frac{2}{\bar{v}^{1/2}} + \frac{24m_3(2\pi)^{3/2}}{\bar{v}^{3/2}}\Big) + B,$$

m_3 given in (5), and,

$$B = \frac{\mathbb{E}\big[| - \log P(X|Y) - H(X|Y)\big|^3\big]}{\sigma^2\phi(Q^{-1}(\epsilon))}.$$

Theorem 5.2 (Converse for $R^*(n,\epsilon)$): Let $(\boldsymbol{X},\boldsymbol{Y})$ be an i.i.d. source-side information pair, with conditional varentropy rate $\sigma^2 = \sigma^2(X|Y) > 0$. For any $0 < \epsilon < \frac{1}{2}$, the pair-based optimal compression rate satisfies,

$$R^*(n,\epsilon) \geq H(X|Y) + \frac{\sigma(X|Y)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log n}{2n} - \frac{C}{n},$$

for all, $n > C^2/[4(Q^{-1}(\epsilon))^2\sigma^2]$, where,

$$C = \frac{\mathbb{E}[| - \log P(X|Y) - H(X|Y)|^3] + 2\sigma^3}{2\sigma^2\phi(Q^{-1}(\epsilon))}.$$

Once again we observe that, in the case of prefix-free codes, Theorem 5.2 remains valid as stated, and Theorem 5.1 remains true with $C+1$ in place of C.

VI. Dispersion

In analogy with the source dispersion for the problem of lossless compression without side information [10] we define:

Definition 6.1: The *pair-based dispersion* $D(\boldsymbol{X}|\boldsymbol{Y})$ of a source-side information pair $(\boldsymbol{X},\boldsymbol{Y})$ is:

$$D(\boldsymbol{X}|\boldsymbol{Y}) = \limsup_{n\to\infty} \frac{1}{n}\mathrm{Var}\big[\ell(f_n^*(X_1^n|Y_1^n))\big].$$

The *reference-based dispersion* $D(\boldsymbol{X}|\boldsymbol{y})$ of a source $\boldsymbol{X}$ with respect to a side information string $\boldsymbol{y} = y_1^\infty$ is:

$$D(\boldsymbol{X}|\boldsymbol{y}) = \limsup_{n\to\infty} \frac{1}{n}\mathrm{Var}\big[\ell(f_n^*(X_1^n|y_1^n))\big].$$

Theorem 6.2: Suppose that both the pair $(\boldsymbol{X},\boldsymbol{Y})$ and $\boldsymbol{Y}$ itself are irreducible and aperiodic Markov chains, with conditional entropy rate $H(\boldsymbol{X}|\boldsymbol{Y})$ and conditional varentropy rate $\sigma^2(\boldsymbol{X}|\boldsymbol{Y})$. Then, $D(\boldsymbol{X}|\boldsymbol{Y}) = \sigma^2(\boldsymbol{X}|\boldsymbol{Y})$. If, moreover, $\sigma^2(\boldsymbol{X}|\boldsymbol{Y})$ is nonzero, then:

$$D(\boldsymbol{X}|\boldsymbol{Y}) = \lim_{\epsilon\to 0}\lim_{n\to\infty} n\left(\frac{R^*(n,\epsilon) - H(\boldsymbol{X}|\boldsymbol{Y})}{Q^{-1}(\epsilon)}\right)^2.$$

Theorem 6.3: Suppose the side information process $\boldsymbol{Y}$ is stationary and ergodic, and that the pair $(\boldsymbol{X},\boldsymbol{Y})$ is conditionally i.i.d. If $\mathbb{E}[V(Y_1)]$ is nonzero, then, for almost any $\boldsymbol{y}$:

$$\begin{aligned} D(\boldsymbol{X}|\boldsymbol{y}) &= \lim_{n\to\infty} \sigma_n^2(y_1^n) \\ &= \lim_{\epsilon\to 0}\lim_{n\to\infty} n\Big(\frac{R^*(n,\epsilon|y_1^n) - H_n(X|y_1^n)}{Q^{-1}(\epsilon)}\Big)^2. \end{aligned}$$

References

[1] A. Aaron and B. Girod. Compression with side information using turbo codes. In *2002 Data Compression Conference*, pages 252–261, Snowbird, UT, April 2002.

[2] A. Aaron, R. Zhang, and B. Girod. Wyner-Ziv coding of motion video. In *36th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 240–244, Pacific Grove, CA, November 2002.

[3] H. Cai, S.R. Kulkarni, and S. Verdú. An algorithm for universal lossless compression with side information. *IEEE Trans. Inform. Theory*, 52(9):4008–4016, September 2006.

[4] S. Chen, M. Effros, and V. Kostina. Lossless source coding in the point-to-point, multiple access, and random access scenarios. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1692–1696, Paris, France, July 2019.

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

[6] M.H.Y. Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome research*, 21(5):734–740, 2011.

[7] L. Gavalakis and I. Kontoyiannis. Fundamental limits of lossless data compression with side information. *ArXiv e-prints*, `1912.05734 [cs.IT]`, December 2019.

[8] Y. Im and S. Verdú. Fixed-length-parsing universal compression with side information. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2563–2567, Aachen, Germany, June 2017.

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

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

[11] D. Pavlichin, T. Weissman, and G. Mably. The quest to save genomics: Unless researchers solve the looming data compression problem, biomedical science could stagnate. *IEEE Spectrum*, 55(9):27–31, September 2018.

[12] S.S. Pradhan and K. Ramchandran. Enhancing analog image transmission systems using digital side information: A new wavelet-based image coding paradigm. In *2001 Data Compression Conference*, pages 63–72, Snowbird, UT, March 2001.

[13] P. Subrahmanya and T. Berger. A sliding window Lempel-Ziv algorithm for differential layer encoding in progressive transmission. In *1995 IEEE International Symposium on Information Theory (ISIT)*, page 266, Whistler, BC, September 1995.

[14] T. Suel and N. Memon. Algorithms for delta compression and remote file synchronization. In K. Sayood, editor, *Lossless Compression Handbook*. Academic Press, 2002.

[15] V.Y.F. Tan and O. Kosut. The dispersion of Slepian-Wolf coding. In *2012 IEEE International Symposium on Information Theory (ISIT)*, pages 915–919, Cambridge, MA, July 2012.

[16] A. Tridgell and P. Mackerras. The rsync algorithm. Technical report TR-CS-96-05, The Australian National University, Canberra, Australia, June 1996.

[17] T. Uyematsu and S. Kuzuoka. Conditional Lempel-Ziv complexity and its application to source coding theorem with side information. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E86-A(10):2615–2617, October 2003.