

Efficient Compression of Long Arbitrary Sequences with No Reference at the Encoder

Conference Paper

Author(s):

Cassuto, Yuval; Ziv, Jacob

Publication date:

2020-02-26

Permanent link:

<https://doi.org/10.3929/ethz-b-000402703>

Rights / license:

In Copyright - Non-Commercial Use Permitted

Efficient Compression of Long Arbitrary Sequences with No Reference at the Encoder

Yuval Cassuto and Jacob Ziv

Viterbi Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa Israel 32000

Email: {ycassuto, jz}@ee.technion.ac.il

Abstract—In a distributed information application an encoder compresses an arbitrary vector while a similar reference vector is available to the decoder as side information. For the Hamming-distance similarity measure, and when guaranteed perfect reconstruction is required, we present two contributions to the solution of this problem. One potential application of the results is the compression of DNA sequences, where similar (but not identical) reference vectors are shared among senders and receivers.

I. INTRODUCTION

This paper¹ continues the line of work on guaranteed-success compression with Hamming-bounded side information [1]. In the first part of the paper (Section II), we study the case where the encoder as usual does not know the decoder's reference vector z , but it does have a set Z of vectors that contains z (among many other vectors). Our results in this part show that if the vectors in Z have a certain well-defined “clustering” property, then it is possible to reduce the compression rate below the best known. This can be achieved without any probabilistic assumptions on the set Z , and without directly enforcing a bound on its size. Our results in this part are for guaranteed-decoding average compression rate, where the average is taken over the random hash function used, and *not* over the input y (which has no probability distribution). For the same model our results also include a lower bound on compression rate for any scheme that uses random hashing. In the second part of the paper (Section III), we return to the classical model of [1] (no Z in the encoder), and propose coding schemes with low complexity of encoding and decoding. For guaranteed decoding of length- n vectors with a constant fractional distance bound p , existing schemes require decoding complexity that is exponential in n due to the complexity of decoding an error-correcting code. Our proposed schemes have $O(n\sqrt{n})$ decoding complexity, which is low enough for practical implementation even for long input sequences. For low distance fractions p , our scheme has low compression rates, although not as low as the prior schemes that do not consider the decoding complexity. We use codes with structure similar to *generalized concatenation* (GC) codes [2].

¹A full version of this paper is currently under review for the IEEE Transactions on Information Theory.

II. STRUCTURED SIDE INFORMATION

Let $Z = \{z_1, \dots, z_M\}$ be a set of vectors, where each vector z_i is a binary vector of length n . The set Z is known to the encoder, and it contains the reference vector z available at the decoder (but the encoder does not know which one it is). The structure of Z is defined through the *p-spread parameter*: $p'(Z, p) \triangleq \frac{D_p(Z)}{2n}$, where $D_p(Z)$ is the maximal distance between a pair of vectors in Z whose distance is at most $2pn$. Given those definitions, we have an achievability result

Theorem 1. *Let Z be a set of reference vectors with p -spread parameter p' . Then there exists a coding scheme where for any input vector y ,*

$$|ENC(y)| \leq n[H(p) + H(p') + \epsilon], \quad (1)$$

as $n \rightarrow \infty$ and on average over the random hash functions.

$H(\cdot)$ is the entropy function. We also have the converse

Theorem 2. *Given the parameters p and p' , any compression scheme that encodes y as $u(y)$, where $u: \{0, 1\}^n \rightarrow \{0, 1\}^m$ is a random hash function, requires asymptotically for any y , on average over the random hash functions*

$$|ENC(y)| \geq n[H(p') + p]. \quad (2)$$

III. UNSTRUCTURED SIDE INFORMATION

For the case of unstructured side information that only assumes that the Hamming distance between y and z is at most pn , we propose a deterministic (guaranteed success) compression scheme built on a GC code construction with the following compression rate

Theorem 3. *For any constant integer l the compression rate of the GC-based construction is*

$$H\left(3p + \frac{1}{l}\left(\frac{1}{2} - 3p\right)\right) + \sum_{i=2}^l \left[H\left(3p + \frac{i}{l}\left(\frac{1}{2} - 3p\right)\right) - H\left(3p + \frac{i-1}{l}\left(\frac{1}{2} - 3p\right)\right) \right] \cdot \frac{3p}{3p + \frac{i-1}{l}\left(\frac{1}{2} - 3p\right)}. \quad (3)$$

REFERENCES

- [1] A. Orlitsky and K. Viswanathan, “One-way communication and error-correcting codes,” *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1781–1788, 2003.
- [2] E. L. Blokh and V. V. Zvyablov, *Generalized Concatenated Codes*. Moscow, Sviaz’ (in Russian), 1976.