On Efficient Entropy Approximation via Lempel-Ziv Compression

Mickey Brautbar * Alex Samorodnitsky *

Abstract

We observe a classical data compression algorithm due to Lempel and Ziv, well-known to achieve asymptotically optimal compression on a wide family of sources (stationary and ergodic), to perform reasonably well even on short inputs, provided the source is memoryless. More precisely, given a discrete memoryless source with large alphabet and entropy bounded away from zero, and a source sequence whose length is a fraction of the alphabet size, the length of the compressed sequence approximates the entropy of the source, up to a constant *multiplicative* factor.

1 Introduction

The universal compression scheme of Lempel and Ziv [10], also known as the 'LZ78' algorithm, is well known to compress any stationary and ergodic source down to the entropy rate of the source per source symbol [9, 4], provided the input source sequence is sufficiently long. However, the rate of convergence of this algorithm, as a function of input length, can be slow [8, 6].

In this short note we assume the source to be memoryless and have entropy bounded away from zero, and observe that, even for a source sequence whose length is a fraction of the source alphabet size, the length of the compressed sequence provides an estimate of the entropy of the source, up to a constant multiplicative factor.

A word on our motivation. Most of the research in information theory concentrated on additive approximation of entropy [5, 1]. This approximation is tighter, but impossible for small input length [2, 7]. Multiplicative approximation was introduced and investigated in [2] from the point of view of sub-linear algorithms, and further studied in [3]. The question of multiplicative approximation achieved by classical compression schemes such as LZ78 was not, to the best of our knowledge, addressed before.

We will try to adhere to the terminology and notation of [4]. We also refer to [4] for wider background and the description and analysis of the compression scheme. Let $\mathcal{X} = (X_i)_{i=1}^{\infty}$ be a discrete memoryless source on alphabet of size q. The alphabet size is assumed to be large. Given a source sequence $(x_1...x_n)$, let $l(x_1...x_n)$ denote the length of the compressed sequence (in bits).

*School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel.

Theorem 1.1: *Assume $H(\mathcal{X}) \geq 1$. There exists a constant c_{lz} such that, for any $0 < \alpha < 1$, and $n = q^{\alpha}$, holds*

$$\Big(1 - o(1)\Big)\frac{\alpha}{(1+\alpha)} \cdot \frac{1}{n} l\left(X_1 ... X_n\right) - c_{lz} - o(1) \leq H(\mathcal{X}) \leq \frac{1}{n} l\left(X_1 ... X_n\right) + o(1)$$

with probability $1 - o(1)$.

The asymptotic notation is w.r.t. $q \to \infty$. The constants hidden in the asymptotic notation are easily computable and small. We may take $c_{lz} = 2$.

Example 1.2: Taking $\alpha = \frac{1}{2}$ in the theorem shows LZ78 to approximate the entropy within a multiplicative factor of 3 on a sample of size $\sqrt{q}$. ∎

2 Proof of the theorem

Recall that Lempel-Ziv encoding parses the input sequence $(x_1 ... x_n)$ into c distinct phrases, and the length of the encoding is $l\left(x_1 ... x_n\right) = c(\log c + \log q)$.

Let p denote the distribution of $X = X_1$. That is $H(\mathcal{X}) = H(X) = H(p)$. The analysis of LZ78 in [4], Chapter 12, implies that for any source sequence $(x_1 ... x_n)$, parsed into c phrases, holds

$$-\frac{1}{n}\sum_{i=1}^{n} \log p\left(x_i\right) \geq \frac{c \log c}{n} - c_{lz}$$

Consider the random variable $Y = -\log(X)$. The expectation of Y is $H(p)$. A simple constrained optimization argument shows $\mathbb{E}Y^2 \leq \log^2(q)$. Let us assume here and (implicitly) below that q is sufficiently large, so that, in particular, $n = q^{\alpha} \geq \log^3 q$. Then, by Chebyshev's inequality, $-\frac{1}{n}\sum_{i=1}^{n} \log p\left(x_i\right)$ is $(1/\log q)$ close to $H(p)$ with probability at least $1 - 1/\log q$. Therefore

$$Pr\Big\{H(\mathcal{X}) \geq \frac{c \log c}{n} - c_{lz} - o(1)\Big\} \geq 1 - 1/\log q \tag{1}$$

For the other direction, we have the following lemma.

Lemma 2.1: *With probability at least $1 - 2/\log q$ holds*

$$H(\mathcal{X}) \leq \frac{1}{n} l\left(X_1 ... X_n\right) + o(1) \tag{2}$$

Proof: Let x^n denote a source sequence $(x_1 ... x_n)$. Let $\delta = 1/\log^2 q + (\log\log q)/n - 2/n$. Consider the following two events.

$$A = \Big\{x^n : \ -\frac{1}{n}\log p\left(x^n\right) - H(\mathcal{X}) < -\frac{1}{\log q}\Big\} \quad \text{and} \quad B = \Big\{x^n : \ \frac{1}{n} l\left(x^n\right) < H(\mathcal{X}) - \delta\Big\}$$

We will show $Pr\{B\} < 2/\log q$. This will suffice to prove the lemma, since for all x^n outside B holds $(1/n)\cdot l\left(x^n\right) \geq H(\mathcal{X}) - \delta \geq H(\mathcal{X}) - o(1)$.

Indeed, assume this is not so. By Chebyshev's inequality, as above, $Pr\{A\} < 1/\log q$. That is, $Pr\{B \setminus A\} \geq 1/\log q$. In particular, the set $B \setminus A$ is not empty.

Consider the effect of LZ compression on the random variable $(X_1...X_n) =: X^n$ conditioned on $B \setminus A$. Let p' denote the conditional distribution. On one hand, clearly,

$$\mathbb{E}\left(\frac{1}{n}l\left(X^n\right) \;\middle|\; B \setminus A\right) \leq H(\mathcal{X}) - \delta$$

On the other hand, Lempel-Ziv encoding is a prefix code on $[q]^n$, and therefore, by ([4], Theorem 5.3.1),

$$\mathbb{E}\left(\frac{1}{n}l\left(X^n\right) \;\middle|\; B \setminus A\right) \geq \frac{1}{n}H(p')$$

We will claim the entropy of p' is larger than $n(H(\mathcal{X}) - \delta)$, providing the contradiction.

For all $x^n \notin A$ holds $(-1/n)\cdot \log p\left(x^n\right) \geq H(\mathcal{X}) - 1/\log q$, that is $p(x) \leq 2^{-n(H(\mathcal{X})-1/\log q)}$. Therefore, for $x^n \in B \setminus A$ holds

$$p'\left(x^n\right) = \frac{1}{Pr\{B \setminus A\}} \cdot p\left(x^n\right) \leq \log q \cdot 2^{-n(H(\mathcal{X})-1/\log q)}$$

The last thing to observe is that the entropy of a distribution of all whose atom weights are at most ϵ, is at least $\log(1/\epsilon) - 1$. Hence $(1/n)\cdot H(p') \geq H(\mathcal{X}) - (1/\log q + (\log\log q)/n + 2/n) > H(\mathcal{X}) - \delta$, completing the proof. ∎

Now, we can complete the proof of the theorem. Assume both (1) and (2) hold, which happens with probability $1 - o(1)$.

The upper bound on $H(\mathcal{X})$ in the theorem is given by (2).

As to the lower bound, we start with a simple calculation which shows (2) implies $\log c \geq \Big(1 - o(1)\Big)\cdot \alpha \log q$.

Indeed, if (2) holds, then $c(\log c + \log q) \geq n(H(\mathcal{X}) - o(1)) \geq \Big(1 - o(1)\Big)\cdot n$ (since, by assumption, $H(\mathcal{X}) \geq 1$).

Since $c \leq n \leq q$, this means $c \geq \Big(1/2 - o(1)\Big)\cdot n/\log q$, that is $\log c \geq \Big(1 - o(1)\Big)\cdot \log n \geq \Big(1 - o(1)\Big)\cdot \alpha \log q$.

Hence,

$$\frac{c\log c}{n} \geq \Big(1 - o(1)\Big)\frac{\alpha}{(1+\alpha)}\cdot\frac{c\log c + c\log q}{n} = \Big(1 - o(1)\Big)\frac{\alpha}{(1+\alpha)}\cdot\frac{1}{n}l\left(X_1...X_n\right),$$

and the lower bound on $H(\mathcal{X})$ in the theorem follows directly from (1).

References

[1] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *RSA: Random Structures and Algorithms*, 19, 2001.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC-02)*, pages 678–687, New York, May 19–21 2002. ACM Press.

[3] M. Brautbar and A. Samorodnitsky. Approximating entropy from sublinear samples. In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 366–375, 2007.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, USA, 1991.

[5] C. Haixiao, S.R. Kulkarni, and S. Verdu. Universal entropy estimation via block sorting. *IEEE Transactions on Information Theory,50*, July,2004.

[6] G. Louchard and W. Szpankowski. On the average redundancy rate of the lempel-ziv code. *IEEE Transactions of Information Theory*, pages 43:2–8, January,1997.

[7] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[8] E. Plotnic, M.J. Weinberger, and J. Ziv. Upper bounds on the probability of sequences emmited by finite-state sources and on the redundancy of the lempel-ziv algorithm. *IEEE Transactions on Information Theory*, pages 38:66–72, 1992.

[9] A. Wyner and J. Ziv. On entropy and data compression. *IEEE Transactions on Information Theory*, 1991.

[10] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions of Information Theory*, pages 24:530–536, September,1978.