Compressor performance, absolutely!

Mark R. Titchener
Dept of Computer Science
The University of Auckland, NZ.
email: mark@tcode.auckland.ac.nz

"... any attempt to derive an allegedly absolute measure for the complexity of an individual sequence will raise understandable objections." [Lempel and Ziv, 1976]

Abstract

In this paper we use results from non-linear/symbolic dynamics, to aid in measuring the performance of a range of popular compressors. We use the logistic map as a well known, 'calibrated' source of information, whose entropy may be controlled by a single parameter r, and which by Pesin's identity (1977), is given by its positive Lyapunov exponent. Using a grammar-based information measure, we derive sample strings from corresponding values of r selected to give entropies that match target normalised values in the range $0.1 - 1.0$ bits/bit. We evaluate popular compressors including `ppmz`, `bzip`, `gzip`, `compress`, `Huffman` (with/without adaption) and `Shannon Fano`, bench-marking their performance against that expected for these selected source samples. This reveals a ranking of the compressors in relation to Shannon's bound, that is ultimately dependent on the source entropy.

I. Introduction

Compression techniques are as much of interest today as they have ever been, this, in spite of the dramatic increases in available bandwidth and similarly spectacular decreases in the costs of communication. Aside from the obvious practical benefits, compression epitomises all that is understood of source coding within the context of Shannon's coding limits [1]. Compression encapsulates the hard reality, i.e, the physical limits of classical IT, in much the same way that heat engines do for thermodynamics. The connection between information theory and thermodynamics is of course more than just superficial since both statistical thermodynamics and classical information theory draw heavily on probability theory. Shannon has in some sense cemented the relationship by borrowing from thermodynamics, the term entropy for his measure of uncertainty.

The historical evolution of the two theories also follow similar patterns. Just as the laws of thermodynamics evolved out of practical endevours, the boring of cannon and subsequently the development of the heat engine, so also classical information theory has arisen from a quest for better data and speech communications techniques, where questions of channel capacity and "amount of information" are pertinent. And as with the development of the heat engine, most progress in relation to compression has come from ad hoc innovations, experimental discoveries, and the ingenious titivation of known techniques, rather than from application of the theory. Indeed the practical applications have tended to stimulate development of the respective theories, rather than the other way round, though the theories shed insight and understanding on the physical limits and constraints that apply respectively.

Invariably, new compression refinements are today still tested, not against the theory and its well defined bounds, but against other existing compressors. Arbitrary collections of files serve as test material for comparative trials, though the information properties of such

files are ultimately known only indirectly by way of the performance outcomes.

Few formal tools or methods are to be found for evaluating compressors against the rather explicit and absolute entropy scale and information bounds of Shannon's theory. The lack of progress on this front may be attributed in part to a quite commonly held view, that it is *not* possible, nor meaningful to discuss the information content of a finite string since this is not defined by the theory. Instead Shannon's definition of entropy is for a source rather than for individual finite messages taken in isolation from the source ensemble.

On the other hand, from the asymptotic equipartition (AEP) theorem [2] we see that for a stationary source having entropy H_s there exists a class of strings typical for the source, strings whose pattern properties are characteristic of H_s. One may indeed use these properties to estimate the source symbol probabilities, but then one faces certain practical limitations, as noted by Kolmogorov [3]; *"A realistic interpretation of probability results is always statistical, and error estimates (occurring in the application of probability results to finite objects) are considerably rougher than in the information theory exposition being developed by us."*

The theory [4] that Kolmogorov refers to here has come to be referred to as Algorithmic Information Theory (AIT) [5] and is concerned primarily with individual string objects. Its algorithmic approach ensures AIT is unencumbered by the "roughness" of the statistical approach and its results are in some sense complementary to Shannon's. The *algorithmic entropy* of a string (more often called the *Kolmogorov complexity*), measured within the context of a universal computing machine, is defined as the size of the smallest self-delimiting program sufficient to generate the given finite string. Such a program together with its input data, is implicitly an optimum decompression algorithm and thus the results of AIT are pertinent to the compression area. Importantly, both AIT and Shannon's theory yield asymptotically equivalent results. In practice we are less concerned about the size of the compressor/decompressor but more concerned with the input and output data files.

A key outcome of AIT is the proof that the *algorithmic information content* of a given finite string is uncomputable, except to within an unknown finite constant [6]. This result is often sported as corroboration of the view that the information content of a finite string object may not in of itself be measured. To avoid becoming bogged down in these philosophical considerations, we simply point out that knowing approximately, even quite precisely the amount of information contained in a finite string, is quite a distinct proposition from knowing exactly how to create, regenerate, or decompress optimally the string. There is indeed *no contention* and a number of techniques for computing the entropy of individual strings have been identified within the rapidly evolving area of physics known as Non-linear Dynamics or Deterministic Chaos [7].

We know that under ideal circumstances the compression factor, C, that will be achieved by a lossless compressor, for a source whose normalised entropy is H_s, $0 \leq H_s \leq 1$, is given simply by:

$$C = \frac{1}{H_S}.$$

It is evident from Shannon's theory that the compression factor C in the limit represents a hard bound, dependent only on the source, a bound which we might expect to approach in practice by suitably clever coding and source modeling techniques. Improvements in compression have noticeably tapered off in recent times, suggesting that we are near the limiting compression factor C. How close to Shannon's bound is not clear. In this paper we exploit results from non-linear dynamics to attempt to answer this question.

In 1958 Kolmogorov noted that Shannon's entropy could be applied meaningfully to

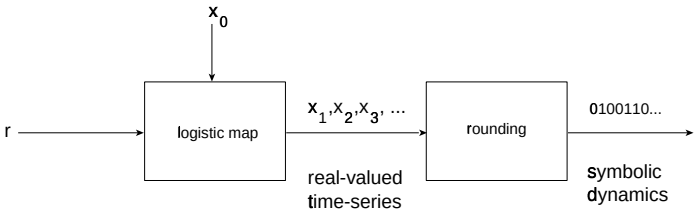

Fig. 1. Diagram illustrating the use of the logistic map as a information source

dynamical systems and then in 1977, Pesin proved that for certain classes of nonlinear dynamical system, including one dimensional maps like the logistic map, the entropy is given by the sum of the positive Lyapunov exponents. A number of techniques have evolved for computing the Lyapunov exponents of simple systems from time series observed of the dynamics. More recently, in [8] [9], a grammar-based information measure [10] was shown to be useful in computing the entropy (positive Lyapunov exponent) for the logistic map from finite samples of the symbolically coded dynamics.

Using the logistic map as an information source, a range of files have been generated with entropies uniformly spaced over the range $0.1 - 1.0$ bits/bit. These are used as source files compressed by a selection of popular compressors, and by measuring the information content of these files before and after compression, it has been possible to establish with good accuracy, each compressor's performance in relation to the Shannon bound.

II. The logistic map

The logistic map is the simplest dynamical system known to exhibit deterministic chaos. In general terms one may visualise the time evolution of the state of such a system as a trajectory in phase space. The non-linearities in the system mean that a pair of initially close neighbour states can rapidly evolve to widely separated states. This sensitive dependence on initial conditions leading to exponential divergence of the trajectories may be quantified by way of Lyapunov exponents, essentially a measure of the global average tendency for the divergence of initially close trajectories. In a D dimensional system there will be D such exponents, λ_i, $i = 1, \ldots D$ with negative exponents indicative of stable periodic orbits, and positive exponents a prerequisite for deterministic chaos. For the logistic map, there is only a single Lyapunov exponent, that is dependent only the system control parameter r. This may be readily computed from the state time-series (c.f. Figure 2).

In 1958 Kolmogorov noted that by partitioning the phase space, and labeling the resultant D-dimensional hypercubes (Markov cells) symbolically, and associating with these the visitation probabilities, conditioned on the visitations to prior cells, one could apply directly Shannon's formula to obtain an entropy of the dynamical system from a symbolic encoding of the trajectory. The *Kolmogorov-Sinai* entropy is thus defined to be the supremum of the Shannon entropy computed over all possible partitions, finite or infinite. In 1977 Pesin proved that for certain classes of system the KS-entropy is given by the sum of the positive Lyapunov exponents. For some systems, such as the logistic map, certain coarse grained partitions known as *generating partitions*, give rise to the supremum. Thus the trajectory may be symbolically encoded by way of a finite alphabet sequence without loss of information. The entropy of the symbolic dynamics for the logistic map is then a function of a single system control parameter, r.

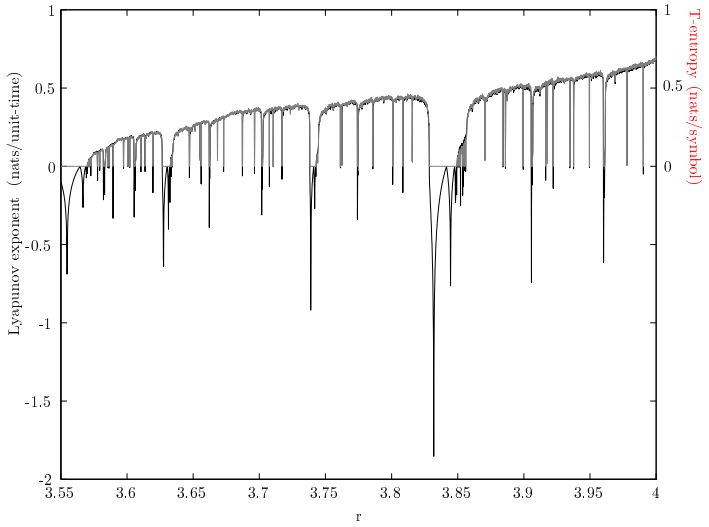

Fig. 2. Plot of the Lyapunov exponent together with the T-entropy computed at increments $\Delta r = 0.001$, from strings of $1,000,000$ symbols derived using a uniformly striped partition with 64 symbols. Similar results are obtained from the generating bi-partition. Thus the T-entropy may be used to reliably measure the entropy of this source from sample strings. The average deviation is ¡ 3% RMS, over the range.

The system state x_{t+1}, at time $(t+1)$, is given in terms of its previous state x_t, at time t by:

$$x_{t+1} = r x_t (1 - x_t).$$

Limiting r to the range $0 \leq r \leq 4.0$ and the initial state x_0 by $0 < x_0 \leq 1.0$, the real-valued time series given by the system states $x_1, x_2, \ldots x_t \ldots$ lies within the interval $(0, 1]$. The state space may be partitioned at c, $0 < c < 1.0$, and encoded symbolically to give a binary string whose elements are: $s_t = 0$ when $x_t < c$, and $s_t = 1$ when $x_t \geq c$. The bipartition $c = 0.5$ is a known generating partition for the logistic map, so the KS-entropy of the binary coded source sequence is given by the corresponding positive Lyapunov exponent, and is 0 elsewhere.

For $r \leq r_\infty (\approx 3.56)$ the KS-entropy $H_{KS} = 0$. For $r > 3.56$ the normalised entropy ranges through from 0, to $ln(2)$ ($ln(2)$ corresponds to the normalised Shannon entropy 1.0) which occurs at $r = 4.0$ [7]. The precise relationship between r and the system entropy is a complex one. It will be evident that the system's behavior is as sensitive to small deviations in r as it is to the initial state. This means that the entropy at $r + \Delta r$ may deviate dramatically from that at r, for $\Delta r << r$.

Figure 2 shows the Lyapunov exponent (courtesy R. Steuer), graphed against r for increments $\Delta r = 0.001$ over the range $[3.55, 4.0]$. The general trend is for H_{KS} to increase with r, but punctuated by sharp troughs and a marked unevenness. To find a value of r that yields a particular entropy requires a hit-and-miss search process within a neighbourhood of an initial estimate of r. In Figure 2 we have also superposed a plot of T-entropy computed from finite strings of $1,000,000$ symbols long sampled from the coded source. From a variety of experiments it is evident that for strings of even modest size, as little as 1000 bits, the normalised T-entropy values are highly indicative of the corresponding Shannon entropy. Thus the grammar-based T-entropy measure provides us with a method for quite

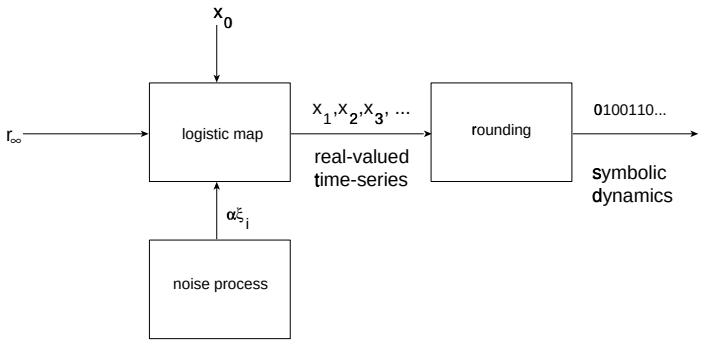

Fig. 3. Logistic map at Fiegenbaum point, with additive noise

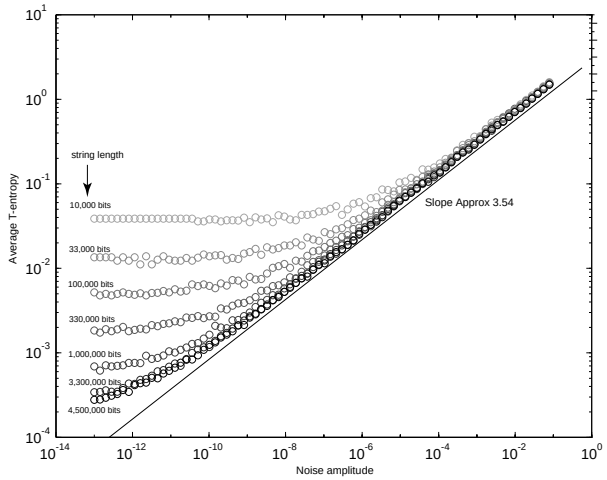

Fig. 4. Illustrating linear sensitivity of T-entropy measure to entropy subject to sample string size.

accurately computing the Shannon entropy directly from the sample strings.

[11] describes a further result involving the logistic map, in which the control parameter is set to the Feigenbaum point $r = r_\infty$ corresponding to the onset of deterministic chaos and noise is injected into the system. It was observed that the entropy of the source was simply proportional to the injected noise amplitude. In duplicating this experiment we find have similarly observed the T-entropy measure to be essentially linearly sensitive for over three orders of entropy magnitude (corresponding to nearly twelve orders of magnitude variation in noise amplitude!). Repeating the experiment for a variety of string lengths leads us to concluded that sample binary strings of 2,000,000 bits are more than adequate to give computed entropies accurate to a few percent over the top decade, i.e., from $0.1 - 1.0$.

A. A corpus of files

Starting with the plotted file of Lyapunov exponents (Figure 1) we have selected initial values of r to obtain sample strings whose normalised entropies covering the range $0.1 - 1.0$ bits/bit at steps of around 0.1. We then tweaked these values or r until the resultant sample files had entropies as close as conveniently possible to the target values. We used the T-entropy measure to enable us to compute these values directly from the symbolic dynamics, and have normalised the computed values on the basis that at $r = 4.0$ the Shannon entropy

	RAW BINARY			PACKED BINARY		
nominal entropy	File Name	T-entropy (nats/bit)	Shannon (bits/bit)	File Name	T-entropy (nats/byte)	Shannon (bytes/byte)
0.1	lgst3.573550	0.04591	.0996	lgst3.573550p	0.35844	.0943
0.2	lgst3.586787	0.09205	.1997	lgst3.586787p	0.71549	.1882
0.3	lgst3.611055	0.13696	.2972	lgst3.611055p	1.1296	.2971
0.4	lgst3.651050	0.18399	.3992	lgst3.651050p	1.46581	.3855
0.5	lgst3.687660	0.23194	.5032	lgst3.687660p	1.90506	.5010
0.6	lgst3.766200	0.27573	.5982	lgst3.766200p	2.23747	.5885
0.7	lgst3.907580	0.32182	.6982	lgst3.907580p	2.58507	.6799
0.8	lgst3.925405	0.36866	.7999	lgst3.925405p	2.95411	.7769
0.9	lgst3.971029	0.41216	.8942	lgst3.971029p	3.25078	.8550
1.0	lgst4.000000	0.46091	1.000	lgst4.000000p	3.80222	1.000

TABLE I
CORPUS OF FILES.

for the logistic map is known to be precisely 1.0.

The result is a set of ten test files each of 2,000,000 bits long (available at `http://tcode.auckland.ac.nz/~corpus/lgst.html`. The ASCII encoded files are named with `lgst` as the prefix indicating the source, followed by the value of r used to derive the file respectively.) Table 1 gives the computed T-entropies values obtained using a UNIX tool `tcalc` (sources also available from the authors home page).

Since the popular compressors we wish to evaluate, operate at the byte rather than bit level, we have packed these files 8-bits per byte. The corresponding files have now just 250,000 bytes. (The packed files have a further suffix "p" to distinguish them from the binary files.) Packing is equivalent to performing a radix change, base-2 to base-256. Each byte is a distinct character from an alphabet of 256 possible characters. The pattern of bits within each byte is now irrelevant as far as the information content of the new string is concerned. Thus packing effectively 'discards' a small amount of the initial information in the binary string. We observe (c.f. Table 1, columns 4 and 6) that in recomputing the normalised T-entropy values for these packed strings, a slight downward shift occurs in the entropy, consistent with this information loss. It is these adjusted values that we use in evaluating compressor performance.

III. RESULTS

With the entropy of the test file's now established to good accuracy, (we estimate this to be within a couple of percent), we are now positioned to evaluate compressor performance. Graphing the compression factor obtained for each file, as a function of file's entropy certainly gives an indication of performance, but since we have the means to measure information content of the strings, before and after compression we have chosen instead to graph information content against file size for the before compression and after files.

Figure 5 illustrates the kind of graph we get from this approach. On the vertical axis we show the *total* (Shannon) information content (actually computed as the product of the normalised T-entropy times the string size) of a) the initial test file (to the right) and b) the final compressed files for each of the compressors respectively (to the left). Information units are here quoted in *bytes* (appropriate to the 256 character alphabet). On the horizontal axis the file lengths are also in *bytes*. Thus a file having maximum information content and corresponding to Shannon entropy $H_s = 1.0$, lies on a line through the origin having slope 1.0). A string with entropy $H_s' < 1$ will appear as point on a line through the origin, whose

slope corresponds to the average entropy for the file, H'_s, given by the *total information* divided by the *total number of characters* in the file.

A compressed file ideally has maximum entropy, i.e., will correspond to a point on the radial line as shown with slope ($H_S = 1.0$). Our uncompressed file in this example appears as a point (entropy $0.094 \approx 0.1$, see first row, column 7 of the above table) well below the bound and to the right of the graph.

The points plotted corresponding to each compressor and labeled accordingly appear at the left, bounded by the radial line ($H_S = 1.0$) representing the Shannon limit. On the assumption that a compressed file ought to retain at the very least the same quantity of information as the initial file, a lossless algorithm will achieve optimal compression by mapping along the horizontal axis as shown. Clearly if a compressed string has less information than its original uncompressed counterpart, then data must have been removed or discarded. We may further classify a compressor as being *lossy* or *lossless* respectively depending on whether the mapping from the initial point to that for the compressed file lies (a) below or; (b) on or above the horizontal line labeled as the "axis of ideal lossless-compression".

Our definition of *lossy* and *lossless* compression, differs subtly from that which might be generally assumed, though the notion broadly corresponds. To illustrate the distinction, consider a compression algorithm designed to operate efficiently with the texts of Shakespeare plays. Each play is stored in a table that is part of the compressor and decompressor respectively. When compressor parses the input text, it compares it with the contents of the table. If it finds that the input corresponds to one of the entries, it simply transmits a brief index into the table, so that the decompressor can output the appropriate entry. Here the transmitted index string will have by the T-entropy measure, much less information content than the text it replaces. Such a compressor would by our definition be classed as lossy algorithm, when in fact in the configuration described it behaves as a lossless scheme. Our method of presenting the results graphically is thus rather more revealing about the actual behavior of each compressor. More typically we find lossless compressors incur coding inefficiencies and signalling costs which give rise to a lifting of the mapping above the ideal.

If an initial file has high information content relative to its length (i.e. high initial entropy), it already starts close to the Shannon limit ($H_S = 1.0$) at the left of the graph. Statistical compressors invariably map to, or close to, this Shannon limit, but as already mentioned also exhibit a significant amount of "lifting". The degree of "lifting" ultimately determines the compressor's compression factor.

In our example `ppmz`, `compress`, and `gzip` are all observed to map close to the Shannon limit, but of these three `ppmz` clearly achieves the best compression factor simply because the amount of "lifting" is minimal. The amount of extra information being added in the process of compression is smallest. For files with high initial entropy this additional information ultimately translates into an expanded file size (e.g. `compress`). The huffman and adaptive huffman encodings in the present example map closest to the axis of ideal lossless compression, clearly adding much less information than the other compressors, but presumably due to other coding inefficiencies end up falling short of the shannon bound. Clearly further investigation is warranted.

In the limit we might anticipate all these points converge, relatively speaking, to the idea Shannon bound. It is beyond the scope of the present paper to explore this dynamic, but it will be obvious that it is a practical proposition to use the techniques introduced here to investigate the rate at which these points converge as a function of file size.

The 10 graphs in Figures 6 and 7 summarize the results obtained for the range of test

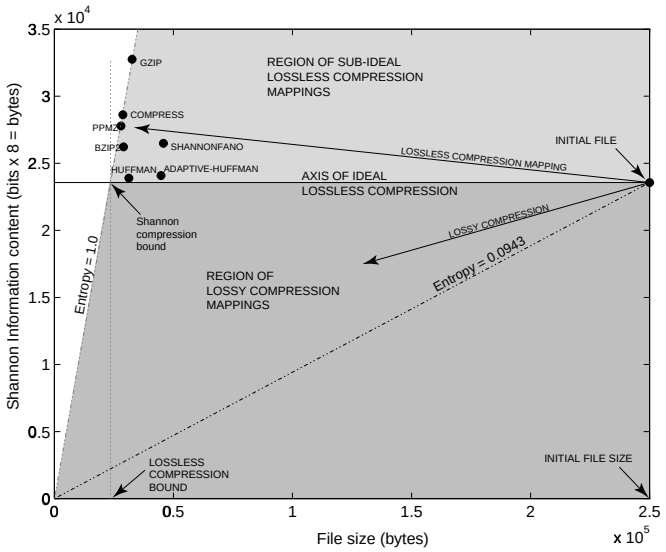

Fig. 5. Graphical depiction of results, explained (see text).

entropies. The performance of these popular compressors is found to be quite variable over this range, giving rise to a ranking of these that alters significantly across the entropy spectrum. The graphs illustrate, to within experimental error (estimated to be a couple of percent), the "absolute" performance of the individual compressors. We assume that the variation of the points evident at or near the shannon bound is simply a result of the limits of accuracy in the measurements. If we assume that the compressed files do in fact fall on the shannon limit, then the collective placement of these points for the compressed files, in relation to the bound shown establishes a bound on the tolerances on the entropy error for the initial uncompressed file. Clearly, more work is possible in both refining the approach taken but also to evaluate the limits to measurement accuracy.

IV. conclusion

In this paper we have used the logistic map as a known information source, to create test files whose entropies are established with reasonable precision by way of the Lyapunov exponents and T-entropy measure. We have thus used these 'calibrated' files to measure the performance of a selection of popular compressors in relation to the Shannon bound. Our approach to assessing performance contrasts with the usual methods which involve comparing the performance of compressors one against another. Used new techniques here we have instead examined compressor performance in relation to the theoretical bound. Graphical depiction of the results provides interesting clues about the coding effects which ultimately limit each of the technologies respectively.

As a general observation it would appear that, particularly for the lower range of entropy values, we are yet some way from the ideal. Noting that the majority of the files in the Calgary Corpus and Canterbury Corpus are texts with entropies in this range [12] it would seem that only limited emphasis can be reasonably assumed when comparing compressors one against another. Much of the variation at and beyond the second significant digit is likely to be 'noise' arising out of the sample variance.

References

[1] C. E. Shannon, "A mathematical theory of communications", *Bell Systems Technical Journal*, vol. 27,

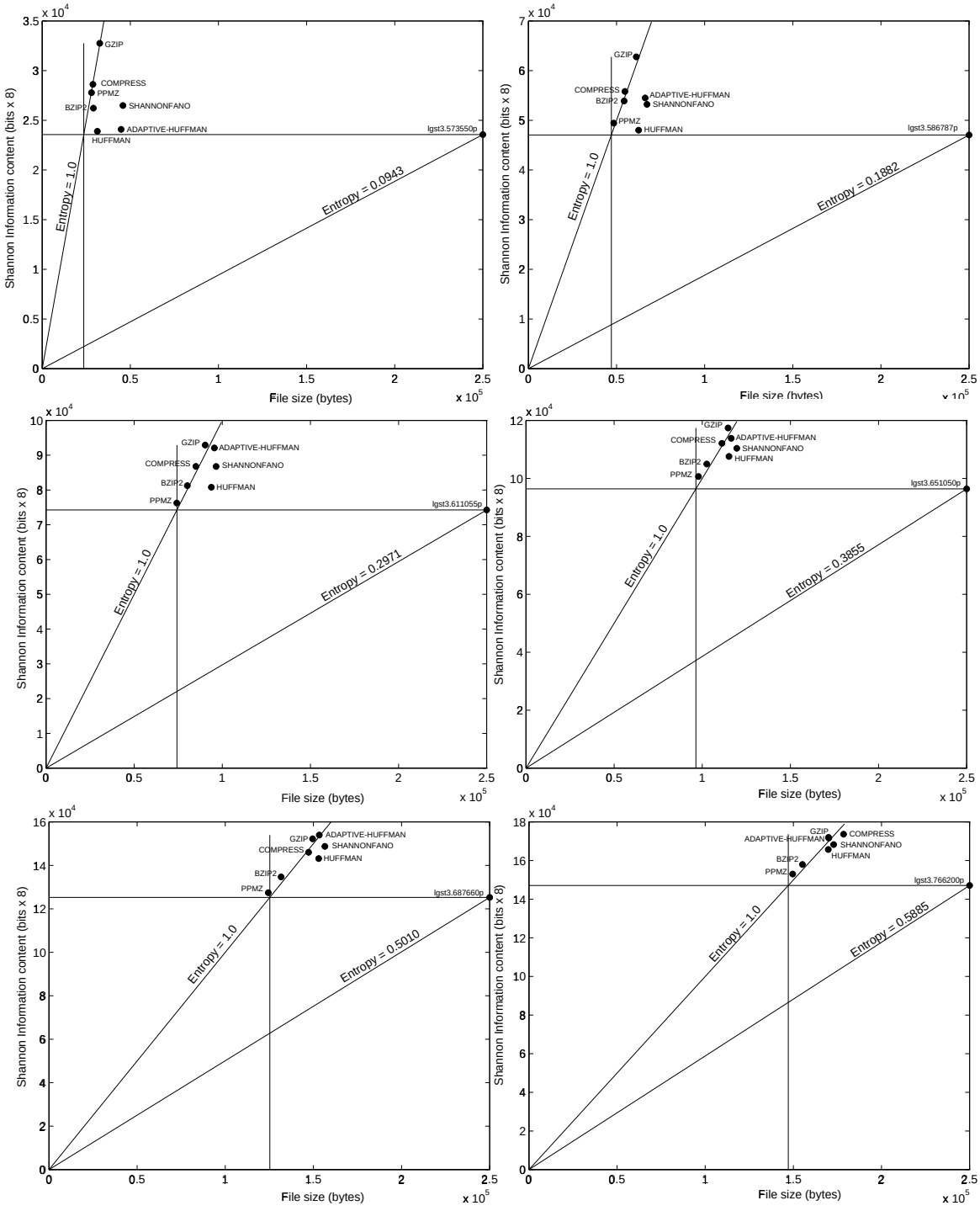

Fig. 6. Depicting results for normalised entropies $0.1 - 0.6$

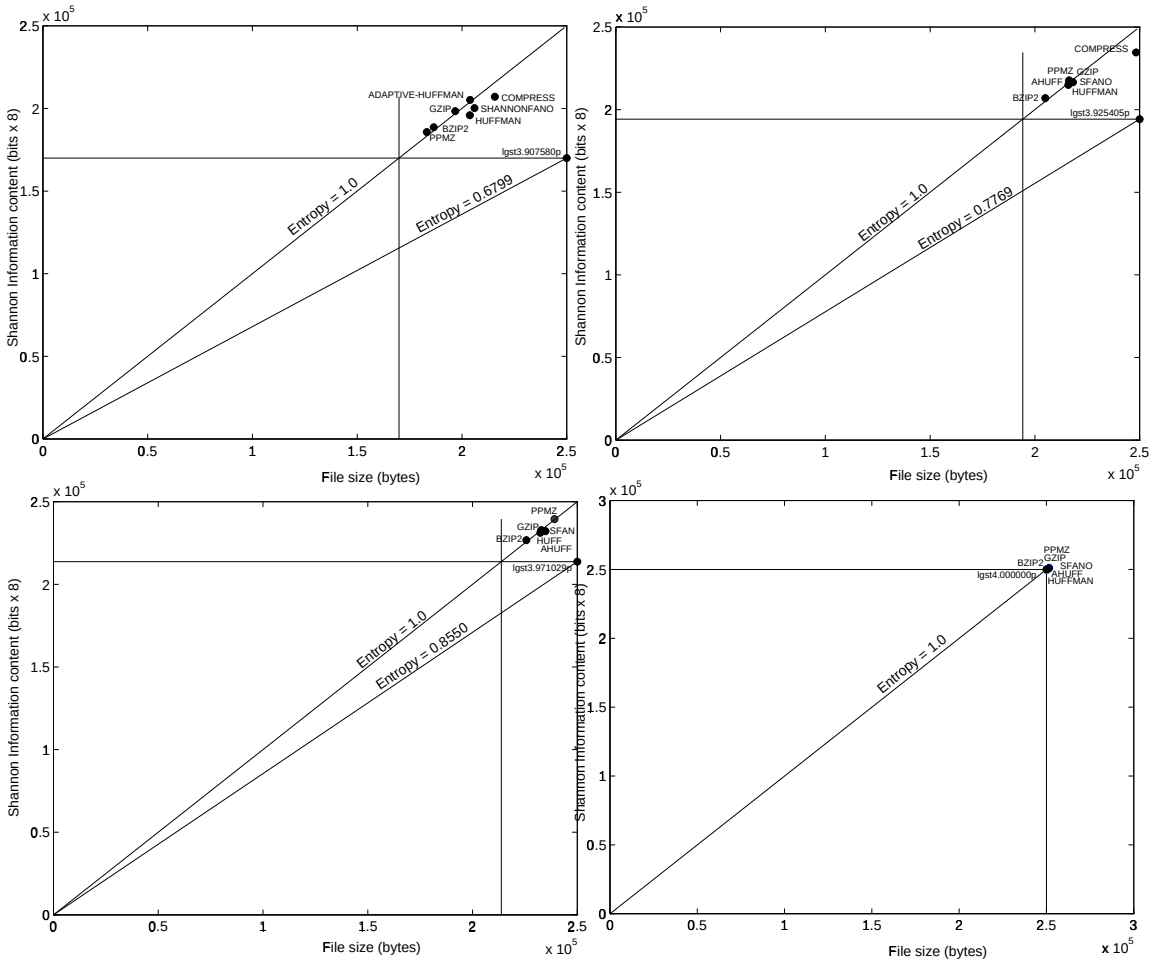

Fig. 7. Depicting results for normalised entropies $0.7 - 1.0$

pp. 379–423, 623–656, July 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.

[3] A. N. Kolmogorov, "Logical basis for information theory and probability theory", *IEEE, Trans IT*, vol. 14, no. 5, pp. 662–664, September 1968.

[4] A. N. Kolmogorov, "Three approaches for definning the concept of "information quantity"", *Problems Inform. Transmission*, vol. 1, pp. 3–11, 1965.

[5] V. A. Uspensky, "Complexity and entropy: an introduction to the theory of kolmogorov complexity", in *Kolmogorov Complexity and Computational Complexity*, Watenabe, Ed. 1987, pp. 86–102, Springer-Verlag, Berlin, Heidelberg.

[6] G. J. Chaitin, "Algorithmic information theory", *IBM J. Res. Develop.*, vol. 21, pp. 350–359,496, 1977.

[7] H. G. Schuster, *Deterministic Chaos: An Introduction*, VCH Verlasgesellschaft mbH, 1989.

[8] R. Steuer W. B. Ebeling and M. R. Titchener, "Partition based entropies of dynamic and stochastic maps", *Stochastics and Dynamics*, vol. 1, no. 1, pp. 45–61, March 2001.

[9] M. R. Titchener and W. B. Ebeling, "Poster: Deterministic chaos and information theory", in *IEEE Data Compression Conference, Snowbird,*, March 27-29 2001, p. 521.

[10] M. R. Titchener, "A measure of information", in *IEEE Data Compression Conference, Snowbird,*, March 28-30 2000, pp. 353–362.

[11] J. P. Crutchfield and N. H. Packard, "Symbolic dynamics of noisey chaos", *Physica*, vol. D, no. 7, pp. 201–223, September 1983.

[12] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, J. C. Lai, and R. L. Mercer, "An estimate of an upper bound for the entropy of English", *Computational Linguistics*, vol. 18, no. 1, pp. 32–40, 1992.