If Some Strings Shrink, Others Must Grow, and What Makes Room for Compression Is Not the String but the Bias of the Source

— Against 2^n strings of n bits there are $2^n - 1$ shorter strings, always one too few — the separator is whether one faces the set of all possible strings taken equally or the strings a biased source emits — [Paper 751]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— *From thinking to syncing.* —

Abstract

Compressors shrink most files. Yet no lossless compressor can shrink every string.Why strings cannot be shrunk, and why they still shrink, are separated by counting and by entropy.No new theorem or law is claimed. First, shorter strings are always one too few——against 4294967296 strings of 32 bits there are 4294967295 shorter ones (Section 2). **Second, so an unshrinkable string always exists——if some strings shrink, others must grow** (Section 2). **Third, and this is the core. Taking all strings equally, few can shrink much——among 20-bit strings, those shrinkable by a byte are below 0.007812 of all** (Section 2). **Fourth, the share falls exponentially with the amount saved——the share shrinkable by k bits never exceeds 2^{1-k}** (Section 2). **Fifth, compression still works because the source is biased ——0.468996 bits per symbol at probability 0.90, 0.080793 at 0.99** (Section 3). **Sixth, the separator is whether one faces the set of all possible strings taken equally or the strings a biased source emits——"compresses well" has no fixed meaning until one says which set of strings is faced** (Section 4). **Keywords**: lossless compression, pigeonhole principle, counting argument, limits of compression, entropy, bias of the source.

> **Scope of this paper (scope note)**: **No new theorem or law is claimed——the counting limit on lossless compression, that most strings cannot be shrunk much, and that the entropy of a memoryless source is the lower limit of compression are all known (Cover and Thomas 2006, Sipser 2012).Compressors are not treated——the performance and methods of particular compressors are not treated.Length delimiting is ignored——the count is an upper bound allowing strings of every shorter length, without the cost of marking where a string ends.The source is limited——bias is examined only in memoryless two-symbol sources.Uncomputability is not treated——that the shortest description of a single string cannot be computed belongs to Paper 252 and is only cited.No real data are claimed——particular files are not treated.Relation to earlier papers**: **Paper 252 showed that information is four quantities, needing different things——its scope note stated that a compressor's failure to shrink is no evidence that a string cannot be written briefly, and that no lower bound can be given. There no lower bound came for a single string. Here, for a set of strings, counting gives one: it cannot say which strings, but it can say that few shrink much. A lower bound for one string, and a lower bound for a set.**What is added is **counting at three lengths that shorter strings are one too few**, **setting the bound on the share shrinkable by k bits beside 2^{1-k} and showing it below 0.007812 for a byte**, **tabulating entropy by source bias**, and **placing the separator on all strings taken equally against the strings of a biased source**.

1. Notation and the Question

Setting: **lossless compression maps each string to another string from which it can be recovered. Mapping two strings to the same one would make recovery impossible, so the map must be one-to-one. Strings of length n and shorter strings are counted.**

$$\#\{\text{length } n\} = 2^n, \qquad \#\{\text{length below } n\} = 2^n - 1, \qquad \frac{\#\{\text{length at most } n-k\}}{2^n} = \frac{2^{n-k+1}-1}{2^n} < 2^{1-k}$$

For a memoryless two-symbol source emitting 1 with probability p, the entropy per symbol is $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$.

Question: **is there a lossless compressor that shrinks every string? If not, why does compression work?**

2. Counting Forbids Compression

Computation 2.1: **at three lengths, strings of that length and shorter strings were counted.**

length	strings of that length	shorter strings
8 bits	256	255
16 bits	65536	65535
32 bits	4294967296	4294967295

> **Observation 2.2**
>
> **shorter strings are always one too few.So all strings of length n cannot be mapped one-to-one into shorter strings, and every lossless compressor leaves some string unshrunk——if some strings shrink, others must grow to make room.**

Computation 2.3: **among 20-bit strings, the upper bound on the share shrinkable by k bits or more was derived.**

k	bound on the share	2^{1-k}
1	9.999990×10^{-1}	1.000000
2	4.999990×10^{-1}	5.000000×10^{-1}
4	1.249990×10^{-1}	1.250000×10^{-1}
8	7.811546×10^{-3}	7.812500×10^{-3}
16	2.956390×10^{-5}	3.051758×10^{-5}

> **Observation 2.4**
>
> **this is the core: taking all strings equally, few can shrink much.Strings shrinkable by a byte (8 bits) are below 0.007812 of all strings——which strings shrink is up to the compressor, but how many can shrink is fixed by counting, whatever the compressor.**

Observation 2.5

the share falls exponentially with the amount saved.At every k the bound stays below 2^{1-k}——nearly every string can lose 1 bit, but at most 2.956390×10^{-5} of them can lose 16.

3. Compression Still Works Because the Source Is Biased

Computation 3.1: for a memoryless two-symbol source, the entropy was derived for each probability of a 1.

probability of 1	entropy (bits per symbol)
0.50	1.000000
0.70	0.881291
0.90	0.468996
0.99	0.080793

Observation 3.2

an unbiased source cannot be compressed.At probability 0.50 the entropy is 1.000000, the same as the set of all strings taken equally——the prohibition by counting is exactly the story of the unbiased source.

Observation 3.3

strings from a biased source shrink a great deal.0.468996 bits per symbol at 0.90, 0.080793 at 0.99——a biased source emits almost only the "shrinkable strings" that are rare when all are taken equally. The room for compression is made by the source, not the string.

4. The Separator

Observation 4.1

the separator is whether one faces the set of all possible strings taken equally or the strings a biased source emits.In the first, counting forbids compression; in the second, entropy gives room——"compresses well" has no fixed meaning until one says which set of strings is faced.

Observation 4.2

beside Paper 252.Paper 252 stated that a compressor's failure is no evidence that a string cannot be written briefly, and that no lower bound comes for a single string——here, for a set of strings, counting gives a lower bound. It cannot say which strings, but it can say that those shrinkable by a byte are below 0.007812.

Observation 4.3

what is not said.Particular compressors, the cost of delimiting, sources with memory and uncomputability are not treated——what is shown is that the counting that forbids compression and the bias that permits it are about different sets.

5. Roll Call and Closing

1. **4294967296 against 4294967295** (Computation 2.1): **one too few.**
2. **an unshrinkable string exists** (Observation 2.2): **others must grow.**
3. **7.811546×10^{-3} at 8 bits** (Computation 2.3): **the counting bound.**
4. **below 0.007812** (Observation 2.4): **the core.**
5. **never above 2^{1-k}** (Observation 2.5): **falling exponentially.**
6. **1.000000 with no bias** (Computation 3.1): **incompressible.**
7. **0.881291 at 0.70** (Computation 3.1): **a little room.**
8. **0.468996 at 0.90** (Computation 3.1): **below half.**
9. **0.080793 at 0.99** (Computation 3.1): **the source makes room.**
10. **all strings, or a biased source's strings** (Observation 4.1): **the separator; Paper 252.**

Closing: **If some strings shrink, others must grow, and what makes room for compression is not the string but the bias of the source.Against 2^n strings of n bits there are $2^n - 1$ shorter ones, so even at 32 bits it is 4294967296 against 4294967295, one too few——every lossless compressor leaves some string unshrunk.Among 20-bit strings, those shrinkable by 8 bits number at most 7.811546×10^{-3} of all, below 0.007812, and the share shrinkable by k bits never exceeds 2^{1-k}——how many can shrink is fixed by counting, whatever the compressor.Compression still works because sources are biased: 0.468996 bits per symbol at probability 0.90, 0.080793 at 0.99.The separator is whether one faces the set of all possible strings taken equally or the strings a biased source emits.Placed among the earlier papers——Paper 252 stated that no lower bound comes for a single string; this gives one for a set by counting.To be honest——particular compressors, the cost of delimiting and sources with memory are not treated.**

References

1. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.
2. Sipser, M. (2012). *Introduction to the Theory of Computation* (3rd ed.). Cengage Learning.
3. Yamagishi, Y. (2026). *Information Is Four Quantities, and They Need Different Things* [Paper 252]. Zenodo. 10.5281/zenodo.22161148

— From thinking to syncing. —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)