

Data Compression and the Limits of Compressibility across Types of Data

Anindya R. Lahiri¹, Rituparna K. Mitra²

ABSTRACT

Storing and transmitting data costs space and bandwidth, and data compression reduces both by representing the same information in fewer bits, but how much a file can be compressed is not the same for all data: some kinds shrink greatly while others barely shrink at all, because compression works by removing redundancy, and different data contain different amounts of it. This study examines data compression and the limits of compressibility across types of data. The compression ratio, the factor by which a file's size is reduced, was examined for several kinds of data, from highly redundant to already dense, so that the wide variation in how much different data can be compressed, and the reason for it, could be seen. The compression ratio varied widely with the kind of data: text, source code, and log files compressed greatly, shrinking to a fraction of their size, because they contain much redundancy, repeated words, patterns, and predictable structure, that compression can remove; whereas already-compressed or encrypted data hardly compressed at all, since they contain little or no redundancy left to remove, and media such as photographs and audio compressed only modestly when compressed without loss, their content being already fairly dense. The amount a file can be losslessly compressed is therefore set by how much redundancy it contains, which differs greatly between kinds of data, and there is a limit beyond which no lossless method can go. For media, much greater compression is possible if some loss of fidelity is accepted, discarding detail the senses barely notice, which is how images and audio are usually compressed. The study shows that compressibility depends on the redundancy in the data and is bounded by it, and it discusses the implications for compressing different kinds of data.

KEYWORDS: *Data compression, Compression ratio, Redundancy, Lossless compression, Lossy compression, Entropy, Information theory, Data storage, File formats*

INTRODUCTION

Storing and transmitting data costs space and bandwidth, both of which are finite and valuable, so reducing the number of bits needed to hold or send a given piece of data is worthwhile throughout information technology. Data compression does this: it represents the same information in fewer bits, so that a compressed file takes less storage and less time to transmit, and is restored to its original form when needed. Compression is therefore used pervasively, in file storage, in communication, and in media (1, 8).

How much a file can be compressed, however, is not the same for all data. Some kinds of file shrink dramatically when compressed, to a small fraction of their original size, while others barely shrink at all, and some cannot be compressed further. This variation is not a quirk of particular methods but reflects something fundamental about the data: compression works by finding and removing redundancy, the predictable or repeated parts of the data that need not be stored in full, and different kinds of data contain very different amounts of redundancy (1).

The measure of how much a file is compressed is the compression ratio, the factor by which its size is reduced. This paper examines data compression and the limits of compressibility across types of data, examining the compression ratio for several kinds of data and the reason it varies. The aim is to show how compressibility depends on the redundancy in the data, and is bounded by it, and to discuss the implications for compressing different kinds of data, including the use of lossy compression for media.

LITERATURE REVIEW

Data compression rests on information theory, which established that the information content of data has a fundamental measure, its entropy, that sets the smallest number of bits into which the data can be losslessly compressed, so that compression can remove redundancy down to this limit but no further (1, 10). The amount a file can be compressed is therefore bounded by its entropy, which reflects how much redundancy it contains (1).

Compression Methods and Their Limits

The literature describes how data is compressed and what bounds it:

- Information theory sets the entropy limit on lossless compression (1, 10).
- Lossless methods, such as Huffman and dictionary coding, remove redundancy exactly (2, 3, 4, 5).
- Lossy methods, such as for images, discard imperceptible detail for greater compression (6).
- Compressibility varies with the redundancy in the data and is bounded by its entropy (1, 8).

A recurring theme is that lossless compression is limited by the data's redundancy, while lossy compression can go further by discarding detail.

The literature develops many lossless compression methods, which remove redundancy without losing any information, including codes that represent frequent symbols in fewer bits, dictionary methods that replace repeated sequences with short references, and arithmetic and transform-based methods, all bounded by the entropy of the data (2, 3, 4, 5, 7). It also develops lossy methods, used chiefly for media such as images and audio, which achieve far greater compression by discarding detail that the human senses barely perceive, accepting a small loss of fidelity for a large saving in size (6). The methods, their performance on different data, and the theory bounding them are treated comprehensively in the standard references (8, 9, 10).

RESEARCH GAP

Although compression and its limits are well understood, a concise demonstration that shows how the compression ratio varies across kinds of data, relating the variation to their redundancy, remains useful as a clear illustration. There is therefore value in a study that displays compression ratios by data type and relates them to the redundancy and entropy of the data.

OBJECTIVES

The objectives of this study are:

- To examine the compression ratio for several kinds of data, from highly redundant to

already dense.

- To show how widely the compression ratio varies with the kind of data.
- To relate the variation to the redundancy each kind of data contains and the entropy limit.
- To relate the result to compressing different kinds of data, including the use of lossy compression.

METHODOLOGY

Compression Ratio and Data Type

The compression ratio, the factor by which a file's size is reduced, was examined for several kinds of data, from highly redundant kinds such as text and logs to already dense kinds such as compressed or encrypted data, using lossless compression (8). The elements are summarised in Table 1.

Table 1: Elements of data compression

Element	Description
Compression ratio	Factor by which the size is reduced
Redundancy	Predictable or repeated parts that can be removed
Lossless	Compression that loses no information
Lossy	Compression that discards imperceptible detail

Across Kinds of Data

The compression ratio was compared across the kinds of data, and the variation related to how much redundancy each contains, since lossless compression removes redundancy down to the entropy limit but no further (1, 10). The far greater compression possible for media when some loss is accepted was also considered (6).

RESULTS AND FINDINGS

The compression ratio varied widely with the kind of data, as shown in Figure 1, which gives the compression ratio for several kinds of data. Text, source code, and log files compressed

greatly, shrinking to a fraction of their size, because they contain much redundancy, repeated words, patterns, and predictable structure, that compression can remove (1, 2).

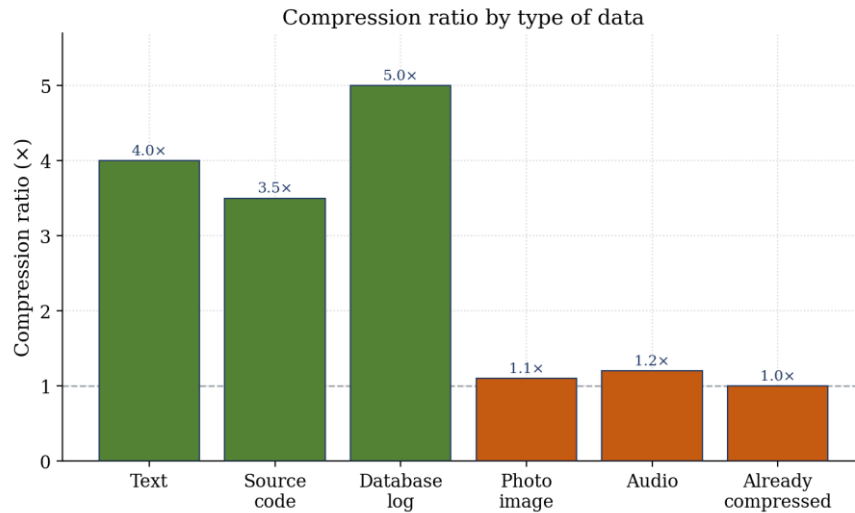


Figure 1: Compression ratio by type of data

Already-compressed or encrypted data, by contrast, hardly compressed at all, since they contain little or no redundancy left to remove, and media such as photographs and audio compressed only modestly when compressed without loss, their content being already fairly dense (1). The amount a file could be losslessly compressed was therefore set by how much redundancy it contained, which differed greatly between kinds of data, with a limit beyond which no lossless method could go (10). For media, much greater compression was possible if some loss of fidelity was accepted (6). The findings are summarised in Table 2.

Table 2: Compressibility by kind of data

Kind of data	Redundancy	Lossless compression
Text, code, logs	High	Compresses greatly
Photographs, audio	Moderate	Compresses modestly
Compressed, encrypted	Little or none	Hardly compresses
Limit	Set by entropy	Cannot exceed

DISCUSSION

The results show that compressibility depends on the redundancy in the data and is bounded by it, and the reason lies in what compression does. Compression saves space by removing redundancy: by not storing in full the parts of the data that are predictable or repeated, encoding them instead more briefly, since they can be reconstructed. Data rich in redundancy, such as text with its repeated words and regular structure, has much that can be removed in this way, so it compresses greatly; data with little redundancy has little to remove, so it compresses little, which is why the compression ratio varies so widely with the kind of data (1).

The near-incompressibility of already-compressed or encrypted data is the clearest illustration of this principle, and of its limit. Compressed data has already had its redundancy removed, and encrypted data is deliberately made to look random; both therefore contain almost no redundancy left for a further pass of compression to find, so attempting to compress them again achieves nothing, and may even enlarge them slightly. This reflects the fundamental limit set by information theory: data can be losslessly compressed down to its entropy, the true measure of its information content, but no further, so once the redundancy is gone, no method can shrink the data more (1, 10). There is thus a hard floor to lossless compression, set by the data itself, not by the cleverness of the method.

The modest lossless compression of media, and the way round it, carries the main practical lesson for compressing different data. Photographs and audio, when compressed without loss, shrink only modestly, because their raw content is already fairly dense in information; yet such media are in practice compressed far more than this, because much of their detail can be discarded without the human eye or ear noticing, and lossy compression exploits this, throwing away imperceptible detail to achieve large reductions in size at a small, often unnoticeable, cost in fidelity (6). This is why images and audio are usually stored in lossy formats, and why the choice of compression depends on the data and the use: lossless compression where every bit must be preserved, as for text and data files, and lossy compression where some loss is acceptable in exchange for much smaller files, as for media (8, 9). The practical conclusion is that the compression to use depends on the data's redundancy and on whether loss can be tolerated, lossless within the entropy limit and lossy beyond it for media, the main caveats being that the achievable ratio depends on the data and method and that the figure illustrates

the variation rather than the compression of specific files.

LIMITATIONS

The following limitations apply to this study:

- The achievable compression ratio depends on the particular data and the compression method.
- The figure illustrates the variation across kinds of data rather than the compression of specific files.
- Lossless and lossy compression were compared in principle; their detailed performance varies.
- The entropy of real data is hard to determine exactly and depends on how it is modelled.

FUTURE SCOPE

Future work could extend the present study as follows:

- Measurement of compression ratios for specific files and methods.
- Comparison of lossless compression algorithms on different kinds of data.
- Study of lossy compression and the trade-off between size and fidelity for media.
- Estimation of the entropy of data to bound its compressibility.
- Examination of compression in storage and transmission systems in practice.

CONCLUSION

Data compression and the limits of compressibility across types of data were examined by examining the compression ratio for several kinds of data. The compression ratio varied widely: text, code, and logs compressed greatly, having much redundancy to remove; already-compressed and encrypted data hardly compressed, having little; and media compressed only modestly without loss, the lossless compression of any data being bounded by its entropy. The study shows that compressibility depends on the redundancy in the data and is bounded by it, so that the compression to use depends on the data's redundancy and on whether loss can be tolerated, lossless within the entropy limit and lossy beyond it for media, with the caveats that the achievable ratio depends on the data and method and that the figure illustrates the variation rather than the compression of specific files.

REFERENCES

1. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.
2. Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9), 1098–1101.
3. Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3), 337–343.
4. Welch, T. A. (1984). A technique for high-performance data compression. *IEEE Computer*, 17(6), 8–19.
5. Witten, I. H., Neal, R. M., & Cleary, J. G. (1987). Arithmetic coding for data compression. *Communications of the ACM*, 30(6), 520–540.
6. Wallace, G. K. (1992). The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1), xviii–xxxiv.
7. Burrows, M., & Wheeler, D. J. (1994). *A block-sorting lossless data compression algorithm* (Technical Report 124). Digital Systems Research Center.
8. Salomon, D. (2007). *Data compression: The complete reference* (4th ed.). Springer.
9. Sayood, K. (2017). *Introduction to data compression* (5th ed.). Morgan Kaufmann.
10. Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons.

***Author for Correspondence**

Anindya R. Lahiri

E-mail: anindya.lahiri88@yahoo.com

¹Associate Professor, Department of Information Technology, Bidhan Nagar Institute of Technology, Krishnanagar, West Bengal

²Assistant Professor, Department of Computer Science and Engineering, Bidhan Nagar Institute of Technology, Krishnanagar, West Bengal

Received Date: January 20, 2026

Accepted Date: February 03, 2026

Published Date: February 22, 2026

Citation: Anindya R. Lahiri, Rituparna K. Mitra. Data Compression and the Limits of Compressibility across Types of Data. Journal of Research in Information Technology and Its Applications. 2026; 11(1): 10-18p.