Information Theory and Entropy Coding: Mathematical origins, Structure and Modern Applications

In 1948, Claude Shannon published 'A Mathematical Theory of Communication' and in doing so founded an entirely new discipline. Before Shannon, the concept of 'information' was intuitive and unmeasurable. Afterwards, it was a precisely defined quantity with a unit, a formula, and a fundamental limit. That limit, entropy, is simultaneously the minimum average number of bits required to represent a message without loss and a measure of the irreducible uncertainty in a random source (Lesne, 2014; Cover & Thomas, 2005). This establishes entropy as both a probabilistic and operational quantity.

Entropy coding is the concrete realisation of this bound in compressors: the family of algorithms that convert symbols into bit streams at lengths as close as possible to their theoretical minimum. This essay traces the mathematics from Shannon's entropy formula through the key constructions, such as Huffman, arithmetic, and ANS coding, and examines where theory meets and diverges from practice.

The starting point is a discrete random source X taking values x with probabilities p(x). Shannon defined the entropy of this source as:

$$H(X) = -\Sigma_X\, p(x)\, log_2\, p(x)$$

The unit is bits. H(X) has two complementary interpretations that are in fact theorems, not merely intuitions. First, it is a measure of uncertainty: a uniform distribution over n symbols has entropy $\log_2 n$ bits, while a deterministic source (one symbol with probability 1) has entropy zero. Second, and more powerfully, it is the minimum average code length: no uniquely decodable code for X can have expected length below H(X), and codes achieving expected length arbitrarily close to H(X) exist. This is Shannon's source coding theorem (Cover & Thomas, 2005; Affeldt et al., 2014).

The theorem is made precise by the inequality:

$$H(X) \leq L < H(X)+1$$

where L is the expected code length per symbol of any optimally designed prefix code. This equation captures the gap between the theoretical ideal and what integer codeword lengths permit: $H(X)$ is achievable in the limit but not always with integer lengths for a single symbol. The gap closes as block length grows. Shannon's contribution was not just the first formula, but the proof that $H(X)$ is both the right definition of information and the right limit for compression, a coincidence that took genuine mathematical work to establish (Li, 2023; Lesne, 2014).

The connection between probability and code length is formalised by the Kraft inequality. For any prefix code, one in which no codeword is a prefix of another, guaranteeing unique decodability, the codeword lengths li must satisfy:

$$\Sigma_i \, 2^{(-l_i)} \leq 1$$

Thus, coding reduces to a constrained optimisation problem.
Conversely, any set of positive integers satisfying this equation corresponds to a valid prefix code. This transforms the code design problem into an optimisation: choose integer lengths li as close as possible to $-\log_2 p(x_i)$ while being satisfactory. The ideal length $-\log_2 p(x_i)$ assigns short codes to frequent symbols and long codes to rare ones. This is the exact intuition behind Morse code, but now with a mathematical proof of optimality (Cover & Thomas, 2005).

For long sequences, a striking concentration phenomenon occurs. The Asymptotic Equipartition Property (AEP) states that for an Independent and identically distributed (i.i.d.) source, as block length n grows, almost all probability mass concentrates on a typical set of approximately 2^(nH) sequences, each with probability close to 2^(−nH). Sequences outside this typical set are collectively negligible. This explains why compression can approach entropy in practice. The AEP is the probabilistic backbone of Shannon's coding theorems: it shows that assigning the same code length nH to each typical sequence and ignoring the rest is near-optimal. It also underpins universal coding, large-deviation analysis, and entropy rate estimation for stationary processes (Cover & Thomas, 2005; Lesne, 2014).

For processes with memory like Markov chains or stationary ergodic sources, the entropy rate $H = \lim_n (1/n) H(X_1, \ldots, X_n)$ generalises the single-symbol entropy, characterising the long-run compressibility of the process. This rate determines the fundamental compression limit for any source with temporal dependencies, including natural language and video frames (Cover & Thomas, 2005).

Huffman coding (1952) solves the integer-length approximation to $-\log_2 p(x_i)$ optimally among all prefix codes. The algorithm builds a binary tree greedily: repeatedly merging the two lowest-probability symbols into a parent node until a single root remains. Leaf depths become codeword lengths, and the resulting code is proven optimal. No other prefix code has

a lower expected length. The average length lies within 1 bit of H(X), with equality when all probabilities are negative powers of 2 (Cover & Thomas, 2005; Cheng & Li, 2021). The limitation is its granularity: integer constraints prevent exact matching to $-\log_2 p(x_i)$.

Arithmetic coding resolves this by operating on sequences rather than symbols. The encoder maintains a sub-interval of [0,1), initially the full unit interval. Each new symbol x with probability p(x) narrows the current interval by the factor p(x), selecting a sub-interval proportionally to p(x)'s position in the symbol-by-symbol Huffman table. After n symbols, the interval length equals the joint probability of the sequence; any binary fraction in this interval is a valid codeword. The code length is $-\log_2$ (joint probability), approaching H(X) for large n, the factor-of-n gain over symbol-by-symbol Huffman.

The mathematics of arithmetic coding treats encoding as a map from symbol sequences to points in [0,1), using properties of probability measures to prove that the encoding length per symbol converges to H as $n \to \infty$. Practical implementations use integer arithmetic and finite-precision registers, introducing a small overhead analogous to the literature (Said, 2023). Arithmetic coding is the engine behind JPEG, and VVC, where it is combined with context-adaptive probability models to approach entropy on real images and video data (Schwarz et al., 2021).

Asymmetric Numeral Systems (ANS), developed by Jarek Duda in the 2010s, achieves arithmetic coding's near-entropy efficiency through a finite-state machine rather than interval arithmetic. The encoder maps a (state, symbol) pair to a new state; the code is the binary representation of the final state. Table-based ANS (tANS) precomputes these transitions, enabling single-instruction throughput on modern CPUs and GPUs — a crucial practical advantage over arithmetic coding's sequential interval narrowing (Strutz & Schreiber, 2024; Auli-Llinaš, 2023). ANS is now used in Zstandard, HEVC range coding variants, and hardware image sensors. Mathematically, it is equivalent to arithmetic coding in terms of compression ratio but radically different in implementation.

Shannon's bound $H(X) \leq L$ is achieved under assumptions of known distribution and asymptotically large block length. Real compressors violate both. Probabilities must be estimated from data, introducing statistical model error; blocks are finite, incurring boundary overhead; and arithmetic must be integer, introducing rounding. The gap between theoretical and observed compression performance is analysed precisely for arithmetic coding and universal codes, with achievable lengths of the form $nH(\theta n)+(d/2)\log n+O$ for parametric models, the second term being the cost of estimating d parameters from n symbols (Ao et al., 2025; Li, 2023).

A more subtle issue concerns what entropy is being measured. The naive Shannon entropy of pixel grey-values in a satellite image ignores spatial correlations; a sophisticated predictor

exploiting those correlations can compress beyond what first-order H(X) predicts without violating Shannon's theorem. It is simply operating on a higher-order, lower-entropy source (Cheng & Li, 2021). This motivates configurational entropy and higher-order entropy rates as tighter compression bounds for structured data.

The relationship between Shannon entropy and Kolmogorov complexity sharpens this picture. Shannon's H(X) is an average over a probabilistic model: computable, but dependent on models. Kolmogorov complexity is the length of the shortest program that outputs a specific sequence: free from models, but uncomputable. In practice, universal compressors such as Lempel–Ziv converge to the entropy rate of the source without knowing its distribution, bridging the two frameworks asymptotically (Cover & Thomas, 2005; Rissanen & Yu, 2000). Claims that novel 'complexity' measures escape information-theoretic bounds have been shown to reduce, on formal analysis, to Shannon entropy or Lempel–Ziv compression (Abrahão et al., 2024). This distinction reflects a deeper divide between probabilistic and algorithmic notions of information.

Entropy coding is not merely inspired by information theory; it is its mathematical realisation. The chain from Shannon's entropy formula (1st equation) through the Kraft inequality (3rd equation), the AEP, and the source coding theorem (2nd equation) is a sequence of theorems, each transforming an abstract probabilistic quantity into a concrete coding constraint. Huffman coding achieves the integer-length optimum; arithmetic coding closes the gap to H(X) at the cost of sequential processing; ANS recovers that compression at modern hardware speeds. The limit H(X) organises theory and practice alike: every real compressor is measured against it, and every gap between the two is a precisely characterised information-theoretic quantity. Shannon's 1948 paper did not merely describe how to compress data. It proved that data compression has a floor, identified what that floor is, and showed exactly how to approach it. The significance of Shannon's result is not merely that compression has a limit, but that this limit is both mathematically inevitable and algorithmically attainable.

References

Abrahão, F., Hernández-Orozco, S., Kiani, N., Tegnér, J., & Zenil, H. (2024). Assembly Theory is an approximation to algorithmic complexity based on LZ compression. *ArXiv*. https://doi.org/10.48550/arxiv.2403.06629

Affeldt, R., Hagiwara, M., & Sénizergues, J. (2014). Formalization of Shannon's theorems. Journal of Automated Reasoning, 53, 63–103. https://doi.org/10.1007/s10817-013-9298-1

Affeldt, R., & Hagiwara, M. (2012). Formalization of Shannon's theorems in SSReflect-Coq. Lecture Notes in Computer Science, 233–249. https://doi.org/10.1007/978-3-642-32347-8_16

Ao, Z., Sun, Z., & Sun, J. (2025). A predictive method for estimating the limits of lossless data compression. 2025 Data Compression Conference (DCC), 359. https://doi.org/10.1109/dcc62719.2025.00047

Auli-Llànès, F. (2023). Fast and efficient entropy coding architectures for massive data compression. Technologies. https://doi.org/10.3390/technologies11050132

Cheng, X., & Li, Z. (2021). Predicting the lossless compression ratio of remote sensing images with configurational entropy. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 11936–11953. https://doi.org/10.1109/jstars.2021.3123650

Cover, T., & Thomas, J. (2005). Elements of Information Theory. Wiley. https://doi.org/10.1002/047174882x

Lesne, A. (2014). Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Mathematical Structures in Computer Science, 24. https://doi.org/10.1017/s0960129512000783

Li, L. (2023). Empirical lossless compression bound of a data sequence. Entropy, 27. https://doi.org/10.3390/e27080864

Rissanen, J., & Yu, B. (2000). Coding and compression: A happy union of theory and practice. Journal of the American Statistical Association, 95, 986–989. https://doi.org/10.1080/01621459.2000.10474290

Said, A. (2023). Introduction to arithmetic coding — Theory and practice. ArXiv. https://doi.org/10.48550/arxiv.2302.00819

Schwarz, H., Coban, M., Karczewicz, M., Chuang, T., Bossen, F., Alshin, A., Lainema, J., Helmrich, C., & Wiegand, T. (2021). Quantization and entropy coding in the Versatile Video Coding (VVC) standard. IEEE Transactions on Circuits and Systems for Video Technology, 31, 3891–3906. https://doi.org/10.1109/tcsvt.2021.3072202

Strutz, T., & Schreiber, N. (2024). Investigations on algorithm selection for interval-based coding methods. Multimedia Tools and Applications, 84, 46553–46580. https://doi.org/10.1007/s11042-025-20971-3

Related reading:

Information Theory - Robert B. Ash
Point Cloud Compression: Technologies and Standardization - Ge Li, Wei Gao, Wen Gao