

# Asymptotic multi-dimensional quantization, rate distortion theory for singular probability measures, design and analysis of data compression algorithms

## »» Tsutomu KAWABATA Laboratory



Tsutomu KAWABATA

### Summary of Research

This message is intended for students who would consider graduate study in this laboratory.

### Research on Source Coding

The purpose of the communication is to transmit the output from a source to a receiver through a communication channel. Examples of source output include continuous-time signals  $X(t)$  (such as sound) and discrete-time discrete-value data  $X_n$  (such as text files).

The first step in information communication is to encode the output from a source using binary symbols (0 and 1). The source output contains a diversity of data types such as electronic texts, images, and sounds. The purpose is to compress and decrease the apparent volume of information. Due to this function, the process is also called source coding or data compression. Source coding simplifies “the communication” to “the transmission of information through a channel.” The source coding theory serves as the basis for various problems, including problems of information transmission.

### Research on Lossless Source Coding

Source coding enables distortion-free or lossless decoding in certain cases: for example, when the source output is a text file. In some cases, distortion-free decoding is inherently impossible: for example, when the source output is a continuous signal. Let’s examine the former case. If  $X_n$  distributes with a probability law, we can examine the expected length of the binary code from the encoder. If the expected length can be minimized over all coding processes, the smallest value can be considered as the volume of essential information contained in the source.

Assuming the probability law hold, is it possible to estimate the value of the essential information volume from  $X_n$ ? If possible, the use of the arithmetic coding method—a technique that converts information into real numbers—would yield the coding to accompany the estimated value. When  $X_n$  is a text file, a group of hierarchical probability laws, called tree sources, is considered to be standard.

### Tree Source, Context Tree Weighting (CTW) Method, Lempel-Ziv Coding

Our laboratory pursues research on the context-tree weighting (CTW) method, an effective estimation method for the tree sources.

If it is computationally unreasonable to approximate with a tree source, we must use a different method to obtain the essential information volume. In such cases, it is effective to employ

the Lempel-Ziv coding method (dictionary method). We also pursue the method, and take both the design-based and the analytical approaches.

The following briefly introduces our research activities pertaining to the design of algorithms for a method called incremental parsing. The incremental parsing process requires an empty dictionary and a character string selected to be coded. The method is very simple. From the beginning of the input character string, the shortest word that is not in the dictionary is cut. Then, that word is registered in the dictionary. This process is repeated. Since a newly cut word consists of the previously registered word and another character, the resulting data can be reduced to a short expression (data compression).

In comparison to the abovementioned basic method, the technique developed by our laboratory achieves effective counting and expression within the tree structure of a dictionary containing new resolved words. Our method not only enables description

### Keywords

Asymptotic vector quantization, source coding, lattice, differential geometry, rate distortion function, singular probability measure, rate distortion dimension, tree source, Lempel-Ziv coding, context tree weighting (CTW) method, enumerative code

Affiliations	Society of Information Theory and its Applications; Institute of Electronics, Information and Communication Engineers; Institute of Electrical and Electronics Engineers (IEEE); Japan Society for Industrial and Applied Mathematics
Member	Tsutomu Kawabata, Professor

with an extremely short procedure, it eliminates a number of drawbacks of the basic method. This newly developed method can be used in a wider range of applications than the incremental parsing method, and its applications continue to expand. In that sense, it can be regarded as a baseline technology.

This method is a byproduct of the analytical research continuously pursued by our laboratory.

### Research on Lossy Source Coding

Next, let's examine a case in which signal distortion is tolerable. In this case, we must investigate the effectiveness of source coding based on the probabilistic assumption. This involves the examination of the tradeoff relationship between compression rate and allowable distortion. Let us introduce two topics on the relationship.

One topic is the research focusing on the relationship that depends on the signal type. For example, singular signals that are neither continuous nor discrete signals appear in actual signals. The theory lacked the capability to cope with such cases, but our laboratory discovered the rate distortion relationship for anomalous signals, including fractal signals.

This compression theory for singular signals that are neither continuous nor discrete signals is capable to open up a new domain of information theory.

Another topic is the research focusing on the relationship in the two-stage quantization system.

Problems in the high-resolution compression of continuous signals are investigated as asymptotic multi-dimensional quantization problems. Those problems translate to the problem of optimally arranging many points in a space.

Our laboratory re-examines the long-neglected two-stage quantization system in an effort to improve the theory and the system.

Source coding is also effective in improving the confidentiality of communication. Improvements in confidentiality, or information hiding theory, will likely generate expanded applications.

#### Advantages

### Highly Original Researches Based on Mathematical Engineering

Information theory is a branch of electrical and electronics engineering and its purpose is to provide the theoretical foundations for digital communications. It places the highest priority on engineering feasibility. Fundamental technologies are crucial to advanced researches in this field.

Our laboratory not just focuses on information theory, whose foundations have been systematized, but looks at the theory from the viewpoint of the mathematical engineering, which includes applied probability, mathematical statistics, and applied geometry. This allows us to work on wider problems. One example. The human genome project organized in the late '80s has increased the search demands for DNA base sequences. To establish a search method, researchers defined the proximity of orderly arranged base sequences as the sum of scores defined on base pairs and sought the largest sum by varying both ends of the subsequence. They then determined a method for searching for the position at which the value becomes largest when the array slides one direction or another.

This method was established as the BLAST method by Stephen Altschul. The BLAST method continues to be used today. Statistical analysis was also performed during that period by Samuel Karlin, Amir Dembo, and Tsutomu Kawabata at Stanford University. The BLAST method is also called Karlin-Altschul statistics.

### A Base for Information Theory: The University of Electro-Communications

Long-range visions must be formulated in order for engineers involved in the development of digital communications technologies to conduct active and meaningful R&D activities on a long-range basis. Information theory involves fundamental principles that allow the formation of long-range visions. Information theory serves, directly and indirectly, as the foundations of research in wide-ranging scientific and engineering fields. In fact, many researchers ultimately arrive at information theory after extensive research across many other fields. The University of Electro-Communications is a home for information theory in Japan, and our laboratory provides opportunities to learn advanced information communication technologies based on the information theory taught at the University.



Students express opinions freely without being confined by conventional ideas or concepts.