Coding and Compression: a Happy Union of Theory and Practice

Jorma Risssanen*   Bin Yu†

1 Introduction

The mathematical theory behind coding and compression began in 1948, a little more than fifty years ago, with the publication of Claude Shannon's (1948) paper "A Mathematical Theory of Communication" in the *Bell Systems Technical Journal*. This paper laid down the foundations for what is now known as information theory in a mathematical framework that is *probabilistic* (see e.g. Cover and Thomas 1991, Verdú 1998). That is, Shannon modeled the signal or message process by a random process and a communication channel by a random transition matrix that may distort the message. In the five decades that followed, information theory provided fundamental limits for communication in general and coding and compression in particular. These limits, predicted by information theory under probabilistic models, are now being approached in real products such as computer modems. Since these limits or fundamental communication quantities such as entropy and channel capacity vary from signal process to signal process or from channel to channel, they have to be estimated for each communication set-up. In this sense, information theory is intrinsically *statistical*. Moreover, the algorithmic theory of information has inspired an extension of Shannon's ideas that provides a formal measure of information of the kind long sought for in statistical inference and modeling. This measure has led to the Minimum Description Length (MDL) principle for modeling in general and model selection in particular (see Rissanen 1978, Rissanen 1989, Barron, Rissanen and Yu 1998, Hansen and Yu 1998).

A coding or compression algorithm is used when one surfs the web, listens to a CD, uses a cellular phone or works on a computer. In particular, when a music file is downloaded through

*Jorma Rissanen is Fellow, IBM Reserch at San Jose, CA. (Email: *RISSANEN@ALMADEN.IBM.COM*).

†Bin Yu is Member of Technical Staff, Bell Laboratories, Lucent Technologies, and Associate Professor of Statistics, University of California at Berkeley. (Email: *binyu@stat.berkeley.edu*). Her research is supported in part by the NSF grant DMS-9803063. The authors are grateful to Diane Lambert for her very helpful comments.

the internet, a losslessly compressed file (often having a much smaller size) is transmitted instead of the original file. Lossless compression works because the music signal is statistically redundant, and this redundancy can be removed through statistical prediction. For digital signals, integer prediction can be easily carried out based on the past signals that are available to both the sender and receiver, so we need to transmit only the residuals from the prediction. These residuals can be coded at a much lower rate than the original signal (see e.g. Edler, Huang, Schuller and Yu 2000).

2 Entropy and lossless coding

Shannon considers messages or signals to be concatenations of symbols from a set $A = \{a_1, \ldots, a_m\}$, called an *alphabet*. For example, the alphabet A for an English message contains the Roman letters and grammatic separation symbols. For an 8-bit digital music signal, A contains the integers from 0 to 255. A lossless *code* is an invertible function $C : A \to B^*$ the set of binary strings or *codewords* and can be represented by nodes in a binary tree as in Fig. 1 (a). It can be extended to sequences $x^n = x_1, \ldots, x_n$, also written as x,

$$C : A^* \to B^*$$

by *concatenation*: $C(xx_{n+1}) = C(x)C(x_{n+1})$. To make the extended code uniquely decodable without the use of separating commas, we impose the restriction that *no codeword is a prefix of another*. Each codeword node in the tree is then a leaf or end-node. See Fig. 1 (b) for an example.

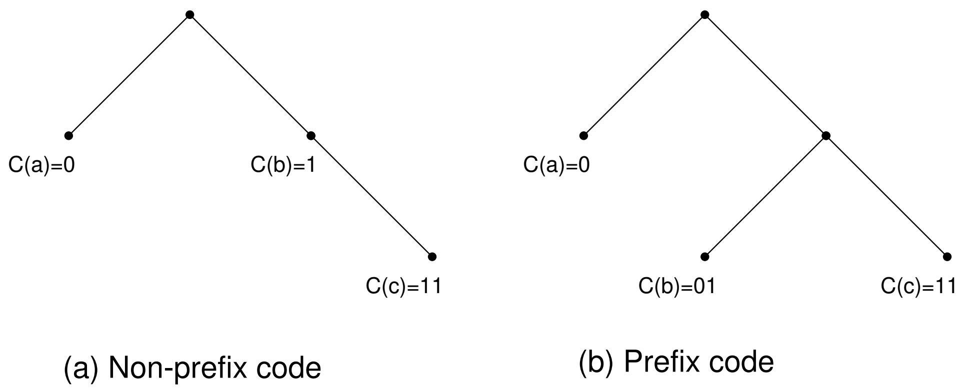

Figure 1. Examples of non-prefix and prefix codes for alphabet $A = \{a, b, c\}$.

This *prefix* requirement implies the important Kraft inequality, which appeared in Kraft's 1949 Masters thesis at MIT:

$$\sum_i 2^{-n_i} \leq 1, \tag{1}$$

where $n_i = |C(a_i)|$ denotes the length of the codeword $C(a_i)$ and in units of *bits* for binary digits, a term suggested by John W. Tukey (see Cover and Thomas, 1991). The Kraft inequality holds even for countable alphabets. Because of this inequality the codeword lengths of a prefix code define a (sub) probability distribution with $Q(a_i) \sim 2^{-n_i}$. Even the converse is true, in the sense that for any set of integers $n_1, \ldots, n_m$ satisfying the Kraft inequality, and in particular for $n_i = -\lceil \log Q(a_i) \rceil$ obtained from any distribution, there exists a prefix code with the codeword lengths defined by the integers. This (sub) probability Q should be viewed as a means to design a prefix code. It is not necessarily the message generating distribution. When the data or message is assumed to be an independently and identically distributed (iid) sequence with distribution P, an important practical coding problem is to design a prefix code C with a minimum expected code length:

$$L^*(P) = \min_C \sum_i P(a_i)|C(a_i)|.$$

The optimal lengths can be found by Huffman's algorithm (see Cover and Thomas, 1991), but far more important is the following remarkable property, proved readily with Jensen's inequality. For any prefix code the expected code length $L(Q)$ satisfies

$$L(Q) \geq -\sum_i P(a_i) \log_2 P(a_i) \equiv H(P), \tag{2}$$

with equality holding if and only if $Q = P$ or $|C(a_i)| = -\log_2 P(a_i)$ for all a_i, taking $0 \log_2 0 = 0$. The lower bound $H(P)$ is the *entropy.* Since the integers $\lceil -\log P(a_i) \rceil$ defining a prefix code exceed the *ideal* code length $-\log P(a_i)$ by at most 1 bit, the optimal code satisfies the inequalities $H(P) \leq L^*(P) \leq H(P)+1$. In terms of the n-tuples or the product alphabet, the joint distribution has entropy $nH(P)$, within 1 bit of $L^*(P^n)$. Thus $H(P)$ gives the lossless compression limit (per symbol).

Concatenation codes are not well suited for small alphabets nor data modeled by a nonindependent random processes even when the data generating distribution P is known, But, there is a different kind of code, *Arithmetic Code*, that is well suited for coding binary alphabets and all types of random processes (see Cover and Thomas 1991). The original version was introduced in Rissanen (1976); for a practical version we refer to Rissanen and Mohiuddin (1989).

For example, consider a binary alphabet A and let $\{P_n(x^n)\}$ denote a collection of nonzero probability functions that define a random process on A, written simply as $P(x^n)$. When the strings x^n are sorted alphabetically, the cumulative probability $C(x^n) = \sum_{y^n < x^n} P(y^n)$ is a strictly increasing function and can be taken as the code of x^n. A binary tree with $s0$ as the left descendant node of s gives the recursion

$$\begin{aligned} C(x^n 0) &= C(x^n) && (3) \\ C(x^n 1) &= C(x^n) + P(x^n 0) && (4) \\ P(x^n i) &= P(x^n)P(i|x^n),\ i = 0, 1. && (5) \end{aligned}$$

A minor drawback of this code is that it is one-to-one only for strings that end in 1. The main drawback is that even if we begin with conditional probabilities $P(i|x^n)$, each written to a finite precision, the multiplication in the third equation keeps on increasing the precision, and soon enough no register can hold the product. To overcome this, replace the exact multiplication in Equation (5) is replaced by following $\bar{P}(x^n i) = \lfloor \bar{P}(x^n)\bar{P}(i|x^n) \rfloor_q,\ i = 0, 1$, where $\lfloor z \rfloor_q$ denotes the truncation of a fractional binary number z to q digits after the first 1, as in $0.0\ldots01x\ldots x$. It is clear then that the addition in the update of the code string can also be done in a register of width $q+1$, except for a possible overflow, which is dealt with a device called 'bit stuffing'. We see in equation (5) that the code length is essentially the number of leading zeros in $\bar{P}(x^n 0)$ plus the size of the register q needed to write down the probability itself. As a result, the mean per symbol code length exceeds the entropy by no more than 2^{-q-1}.

When the data-generating distribution P is unknown, an adaptive arithmetic code can be designed based on an estimator of P provided that the estimator itself is coded and transmitted. This leads naturally to the topic of the next section.

3 Universal coding and Minimum Description Length Principle

As we have seen, Shannon's theory of lossless coding requires knowledge of the probability distribution generating the data. The theory of *universal* coding applies when a family of distributions is selected. Consequently, statistical modeling is indispensable in universal coding. A two-stage universal code based on a family of distributions consists of the coding of an estimated distribution (the overhead) and the coding of data using the estimated distribution. A predictive universal coding scheme requires a predictor based on the past and the coding of the residuals, but no overhead is needed.

The performance of the distribution estimator or the predictor is evaluated from the coding point of view. The first question is whether any universal codes exist such that the per symbol code length, either in the mean or almost surely, approaches the entropy no matter which distribution in the family generates the data. As a functional of the process, entropy is similar to the ergodic mean. Thus the existence of a universal code requires much less than the estimability of the joint distribution of the process. There are two main types of universal codes in use for data compression. The first, due to Ziv and Lempel (LZ) (1977, 1978), is fast and efficient for data arising from written natural languages and is extensively used for file compression (e.g. *compress* in the unix system). LZ is based on an efficient parsing algorithm that builds an increasing tree of phrases. So far no explicit probabilistic model has been found for the LZ algorithm. The second type of universal code is based on an explicit probability model. One good example is Rissanen's context algorithm, which builds on the finite state machine defined by context trees. It is slower than LZ, but provides better compression for image data. Weinberger, Seroussi and Sapiro's (1998) LOCO algorithm is based on the context modeling idea and has been selected as the new lossless (or nearly lossless) image compression standard JPEG-LS-2000. The mean per symbol length of the LZ code approaches the entropy at the rate of $O(1/\log n)$, n being the length of the string, no matter which process in a large class of stationary ergodic processes generates the data. Rissanen's context code for Markov processes converges at rate $0.5k \log n/n$, where k denotes the number of free parameters. The faster rate of Rissanen's algorithm under the context model and the fact that it works for image data suggests that the context model is adequate for images, but not written natural languages. On the other hand, the superior empirical performance of LZ for written natural languages contrasts with its slow rate under stationary and ergodic assumptions. This questions the appropriateness of such assumptions or even the whole analysis framework. Perhaps, instead of a fixed alphabet, an expanding alphabet has to be modeled to reflect what LZ actually does. This calls for a new class of statistical analyses based on array asymptotic results.

The theory of universal codes suggests that amount of information in a finite set of data, given a class of models, can be measured by the shortest universal code length, called the *stochastic complexity*, for that class (see Barron, Rissanen and Yu 1998). Recent results show that the stochastic complexity is the minimum in expectation for the worst case model. The best universal code defines a *universal model*, which decomposes the data into noninformative 'noise' that cannot be compressed with the given models and the signal that is defined by an information bearing optimal model. This idea generalizes the usual decomposition of the sufficient statistics. Given several classes of models for the signal, the *Minimum Description Length* (MDL) principle searches for the class with the smallest stochastic complexity, which implies the most efficient removal of incompressible noise.

Various universal codes have been found to approximate the stochastic complexity to first order ($0.5k \log n/n$) for iid, Markov, time series, and other regular parametric models of dimension k (see Barron, Rissanen and Yu 1998). Three such codes are the two-stage, predictive and mixture codes. In a two-stage code the estimated distribution is the plug-in MLE distribution. In a predictive code the predictive distribution uses the MLE plug-in estimator based on the past. A mixture code is based on the marginal distribution of a model class with respect to a regular prior. There is an important connection between universal coding and model selection. Using MDL with a two-stage code gives the conventional BIC model selection criterion; MDL with a predictive gives the accumulated prediction error criterion; and MDL with a mixture code generalizes the Bayes factor to more than two classes. Thus, MDL provides a way to compare frequentist and Bayesian model selection criteria (see Hansen and Yu 1998).

It remains open how MDL can be used to select among complex models such as Bayesian belief networks. Both information theory results (e.g. universal coding theorems) and statistical evaluation analyses are anticipated.

4 Rate-distortion function and lossy compression

Entropy gives the limiting rate at for lossless compression, but often the practical bandwidth of a communication channel or the available disk space for storage demands a rate lower than the entropy. A compromise is to sacrifice fidelity in the original message or data in exchange for a lower rate. The mathematical study of this trade-off is the Shannon's rate-distortion theory (see Cover and Thomas, 1991), and the practice of coding with loss of information is called lossy compression. For example, an image retrieved from a remote website is compressed via the lossy compression standard JPEG. (A more complete review of the area is given in Berger and Gibson 1998.)

The rate-distortion function is a sort of entropy function when distortion is allowed between the original data and the coded data. Given a distortion distance d and a threshold D, the rate-distortion function is defined in the iid case as

$$R(D) := \min_{\hat{X}} I(X, \hat{X}),$$

where

$$I(X, \hat{X}) = \int \int f(x, \hat{x}) \log f(x, \hat{x}) / [f(x) f(\hat{x})] dx d\hat{x}$$

defines the *mutual information*. Here X is the random variable generating the source messages and $\hat{X}$ is another coded message variable which is within distance D from X on average. It is a remarkable fact that $R(D)$ is the minimum achievable rate at distortion D. Calculating $R(D)$ even in the iid case can be a daunting task, and finding codes that achieve $R(D)$ is beyond reach except for special cases, such as Bernoulli and Gaussian. And, of course, speech and image signals are never well-modeled by such processes. Prediction based on time series models and lossy coding of residuals, which are much closer to iid, are the basic ingredients behind speech coders such as ADPCM, which codes our telephone conversations.

In statistics, data can be transformed to expand the applicability of conventional models tremendously. For decades, transform coding has also been the major approach to lossy image compression. The current JPEG standard uses Discrete Cosine Transform (or DCT) in a blockwise manner, and the next generation JPEG2000 will be wavelet-based. In the transform domain, local iid statistical modeling or context-based prediction and arithmetic coding of the scalar-quantized coefficients has given rise to the state-of-the art wavelet image coders as in LoPresto, see Ramchandran and Orchard (1997) and Yoo, Ortega and Yu (1999).

Statistical modeling will also continue to play an essential role in lossy compression. Better predictions based on more sophisticated models will decrease further the entropy of the residuals to be coded and hence reduce the coding rate. This has spurred attempts to develop an expansion of the information-theoretic framework to encompass both estimation and coding limits.

5 Channel coding and channel capacity

Channel coding is about designing codes and efficient decoding algorithms to combat channel noise. Its goal is to spread the codewords out so that they can be distinguished with high probability when the noisy signal is received. The theory of algebraic coding relies on the same combinatorics as experimental design does. For a channel code, the recovery of codewords from the received signal amounts to hypothesis testing. Given a received signal, maximum likelihood corresponds to hard decoding, which uses the most likely code word, and Bayesian methods correspond to soft decoding, which gives a probability distribution over all possible codewords.

Channel coding was considered in parallel with source coding in Shannon's 1948 work. He modeled a (single-user) communication channel by a conditional probability distribution $f(y|x)$ on the output message y given the input message x. He defined *channel capacity* C to characterize an (iid) probabilistic channel as $C = \max_{w(x)} I(X,Y)$, where the maximization is over input distributions w, which the designer can select, and the channel output Y is determined by $w(x)f(y|x)$.

Shannon's channel coding theorem showed, contrary to the popular belief at the time, that if and only if the average transmission rate is below C, iid messages can be transmitted with error probability decreasing to zero by increasing the size of the message block. This theorem gives the operational meaning of C as the upper bound at which a channel can carry reliable communication. Shannon (1948) calculated the capacity of the additive Gaussian white noise channel with a power constraint. This capacity is being approached by sophisticated coding methods in applications such as space and satellite communications, audio/video transmission, CDs and CD-ROM's .

Models for wireless communication channels, such as those in cellular phones, have to be time-varying and accommodate multi-users (see Biglieri, Proakis and Shamai 1998, Lapidoth and Narayan 1998). Their decoding is essentially statistical inference with time-varying time series models. Joint source-channel coding gives better data throughput, but complicates decoding even further. With limited bandwidth to share, any decoding mileage is worthwhile. This demands efficient statistical inference procedures for special time-varying time series models, and fast algorithms in applications such as cellular phones. It is safe to say that wireless communication promises to be another good example of the happy union of theory and practice.

REFERENCES

Barron, A.R., Rissanen, J., and Yu, B. (1998). "The MDL Principle in Modeling and Coding," (Special Commemorative Issue: Information Theory: 1948-1998) *IEEE. Trans. Inform. Th.*, Vol.44, 2743-2760.

Berger, T. and Gibson, J. D. (1998). "Lossy Source Coding," (Special Commemorative Issue: Information Theory: 1948-1998) *IEEE. Trans. Inform. Th.*, Vol.44, 2693-2723.

Biglieri, E., Proakis, T., and Shamai, S. (1998). "Fading Channels: Information Theoretic and Communication Aspects," (Special Commemorative Issue: Information Theory: 1948-1998) *IEEE. Trans. Inform. Th.*, Vol.44, 2619-2692.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory,* Wiley.

Edler, B., Huang, D., Schuller, G., and Yu, B. (2000). "Perceptually Lossless Sound Coding: Separating the Removals of Irrelevance and Redundancy," in preparation.

Hansen, M. and Yu, B. (1998). "Model selection and Minimum Description Length principle," *J. Amer. Statist. Assoc*, (submitted).

LoPresto, S, Ramchandran, K, and Orchard, M. (1997). "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," *Proc. Data Compression Conference*, Snowbird, Utah.

Lapidoth, A. and Narayan, P. "Reliable Communication under Channel Uncertainty," (Special Commemorative Issue: Information Theory: 1948-1998) *IEEE. Trans. Inform. Th.*, Vol.44, 2148-2177.

Rissanen, J. (1976). " Generalized Kraft-inequality and Arithmetic Coding," *IBM J. Res. Devel.,* Vol.20, 198-203.

Rissanen, J. (1978). "Modeling by Shortest Data Description," *Automatica*, Vol.14, 465–471.

Rissanen, J. (1989). *Stochastic Complexity and Statistical Inquiry.* Singapore: World Scientific.

Rissanen, J. and Mohiuddin, K. (1989). "A Multiplication-Free Multialphabet Arithmetic Code," *IEEE Trans. Communications,* Vol.37, 93-98.

Shannon, C. E. (1948). "A Mathematical Theory of Communication", *Bell Syst. Tech. J.* , Vol.27, 379-423 and 623-656.

Verdú, S. (1998). "Fifty Years of Information Theory," (Special Commemorative Issue: Information Theory: 1948-1998) *IEEE. Trans. Inform. Th.*, Vol.44, 2057-2078.

Weinberger, M. J., Seroussi, G., and Sapiro, G. (1998). "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Trans. Image Proc.,* (submitted). Available as Hewlett-Packard Laboratories Technical Report No. HPL-98-193.

Yoo, Y., Ortega, A. and Yu, B. (1999). " Image subband coding using context-based classification and adaptive quantization," *IEEE Trans. Image Processing*, Vol.8, 1702-1215.

Ziv, J. and Lempel, A. (1977). "A universal algorithm for sequential data compression," Vol.33, 337 - 343.

Ziv, J. and Lempel, A. (1978). "Compression of individual sequences via variable-rate coding," Vol. 33, 530 - 536.