# Admissibility Properties of Gilbert's Encoding for Unknown Source Probabilities

THOMAS M. COVER

*Abstract*—**Huffman coding provides optimal source encoding for events of specified source probabilities. Gilbert has proposed a scheme for source encoding when these source probabilities are imperfectly known. This correspondence shows that Gilbert's scheme is a Bayes scheme against a certain natural prior distribution on the source probabilities. Consequently, since this prior distribution has nonzero mass everywhere, Gilbert's scheme is admissible in the sense that no source encoding has lower average code length for every choice of source probabilities.**

## I. Introduction

### Coding for Known $p$

Consider the random variable $X$, where $\Pr\{X = x_i\} = p_i$, $i = 1,2,\cdots,N$, $p_i \geq 0$, $\Sigma p_i = 1$. Let $\boldsymbol{p} = (p_1,p_2,\cdots,p_N)$, and let $\mathscr{C} = \mathscr{C}(\boldsymbol{p})$ denote an instantaneous code (see Abramson [1]) for $X$, where $\mathscr{C}$ consists of $N$ codewords over a given alphabet, instantaneously decodable, with word lengths $l_1(\mathscr{C}),l_2(\mathscr{C}),\cdots,l_N(\mathscr{C})$. Define the *average code length*

$$L(\mathscr{C},\boldsymbol{p}) = \sum_{i=1}^{N} p_i l_i(\mathscr{C}). \tag{1}$$

Let the minimum average code length with respect to $\boldsymbol{p}$ be denoted by

$$L^*(\boldsymbol{p}) = \inf_{\mathscr{C}} L(\mathscr{C},\boldsymbol{p}), \tag{2}$$

and let $\mathscr{C}^*(\boldsymbol{p})$ denote an encoding that achieves $L^*(\boldsymbol{p})$. It is known that Huffman encoding achieves $L^*$ and that

$$H(\boldsymbol{p}) \leq L^*(\boldsymbol{p}) \leq H(\boldsymbol{p}) + 1, \tag{3}$$

where $H$ is the entropy function.

## II. Random $p$

Now let $\boldsymbol{p}$ be a random variable drawn according to density $g(\boldsymbol{p})$. Then the *expected* code length for the encoding $\mathscr{C}$ is simply

$$E_{\boldsymbol{p}}\{L\} = E_{\boldsymbol{p}}\left\{\sum_{i=1}^{N} p_i l_i(\mathscr{C})\right\} = \sum_{i=1}^{N} E\{p_i\} l_i(\mathscr{C}). \tag{4}$$

Thus, by the definition given in (2) it follows that the minimum expected average code length

$$\min_{\mathscr{C}} E_{\boldsymbol{p}}\{L\}, \tag{5}$$

is achieved by $\mathscr{C}^*(E\{\boldsymbol{p}\})$, i.e., the Huffman code with respect to the expected event probabilities $E\{p_i\}$.

## III. Random $p$ and Sample $\{X_i\}$

Suppose now that $\boldsymbol{p}$ and $g(\boldsymbol{p})$ are unknown, but that $n$ independent identically distributed observations $X_1,X_2,\cdots,X_n$ are drawn according to the given fixed but unknown distribution $\boldsymbol{p}$. Let $n_i$ equal the number of occurrences of $X_j = x_i$ in the $n = \Sigma n_i$ trials. Although $\boldsymbol{n} = (n_1/n,\cdots,n_N/n)$ is an unbiased estimate of $\boldsymbol{p}$, the encoding $\mathscr{C}^*(\boldsymbol{n})$ will be very poor in those cases in which $\boldsymbol{n}$ badly underestimates some of the components of $\boldsymbol{p}$, thus assigning these events very long codewords.

In order to avoid this danger, Gilbert [2, p. 32] suggests that $\boldsymbol{n}$ should be modified according to the formula

$$\hat{p}_i = (n_i + \lambda_i)/(n + \Sigma\lambda_i), \qquad i = 1,2,\cdots,N, \tag{6}$$

and that a $\mathscr{C}^*(\hat{\boldsymbol{p}})$ code be used.

Manuscript received June 24, 1971. This work was performed at the Bell Telephone Laboratories, Murray Hill, N.J.

The author is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, Calif. 94305. He is currently on sabbatical leave at the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02114, and at Harvard University, Boston, Mass.

We now wish to show that $\mathscr{C}^*(\hat{\boldsymbol{p}})$ minimizes

$$E\{L(\mathscr{C},\boldsymbol{p}) \mid X_1,X_2,\cdots,X_n\}, \tag{7}$$

the expected average code length conditioned on the data, if in fact $\boldsymbol{p}$ is drawn according to a Dirichlet prior distribution with parameters $(\lambda_1,\lambda_2,\cdots,\lambda_N)$, defined by the density function

$$g(p_1,p_2,\cdots,p_N) = C(\lambda_1,\lambda_2,\cdots,\lambda_N)p_1^{\lambda_1-1}p_2^{\lambda_1-1}\cdots p_N^{\lambda_N-1},$$
$$p_i \geq 0,\ \Sigma p_i = 1,\ \lambda_i \geq 1, \tag{8}$$

where $C(\lambda_1,\cdots,\lambda_N)$ is a suitable normalization constant. This follows from application of Bayes' rule, from which it follows that the posterior distribution of $\boldsymbol{p}$ given $X$ is again a Dirichlet distribution, this time with parameters $\lambda_i + n_i$, given by

$$g(p_1,p_2,\cdots,p_N \mid X_1,X_2,\cdots,X_n)$$
$$= C(\lambda_1 + n_1,\ \lambda_2 + n_2,\cdots,\ \lambda_N + n_N)p_1^{\lambda_1+n_1-1}p_2^{\lambda_2+n_2-1}\cdots,$$
$$p_N^{\lambda_N+n_N-1}, \qquad p_i \geq 0,\ \Sigma p_i = 1. \tag{9}$$

Now if $\boldsymbol{p}$ is drawn according to a Dirichlet distribution with parameters $\lambda$, we easily calculate

$$Ep_i = \lambda_i/\Sigma\lambda_i, \qquad i = 1,2,\cdots,N. \tag{10}$$

Thus, from (9) we have

$$E\{p_i \mid X_1,X_2,\cdots,X_n\} = \frac{\lambda_i + n_i}{\Sigma\lambda_i + n} \triangleq \hat{p}_i, \qquad i = 1,2,\cdots,N. \tag{11}$$

Finally,

$$E\{L(\mathscr{C}^*(\hat{\boldsymbol{p}})) \mid X_1,X_2,\cdots,X_n\} = E\{\Sigma p_i l_i(\mathscr{C}^*(\hat{\boldsymbol{p}})) \mid X_1,X_2,\cdots,X_n\}$$
$$= \Sigma\hat{p}_i l_i(\mathscr{C}^*(\hat{\boldsymbol{p}})) = L^*(\hat{p}). \tag{12}$$

Thus, if $\boldsymbol{p}$ is Dirichlet with parameters $\lambda$, and $\boldsymbol{n}$ is observed, then by (12) and Section II, $E\{L(\mathscr{C}) \mid X_1,X_2,\cdots,X_n\}$ is minimized, over all codes $\mathscr{C}$, by the code $\mathscr{C}^*(\hat{\boldsymbol{p}})$, which is the Huffman encoding with respect to $\hat{\boldsymbol{p}}$. Thus $\mathscr{C}^*(\hat{\boldsymbol{p}})$ is a Bayes rule.

As pointed out in [2], one may also think of $\mathscr{C}^*(\hat{\boldsymbol{p}})$ as being the natural rule with respect to a sample $(n_1,n_2,\cdots,n_N)$ to which $\lambda = (\lambda_1,\lambda_2,\cdots,\lambda_N)$ additional observations have been added. $\lambda = (1,1,\cdots,1)$ corresponds to a uniform prior distribution.

*Example:* Suppose that all values $p_1$, $p_2$, $p_3$, $p_4$ are *a priori* equally likely. Thus $\boldsymbol{p}$ is Dirichlet with parameters (1,1,1,1) and $\hat{\boldsymbol{p}} = (\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$. The optimal binary encoding of events $x_1$, $x_2$, $x_3$, $x_4$ is thus

$$x_1 \rightarrow 00$$
$$x_2 \rightarrow 01$$
$$x_3 \rightarrow 10$$
$$x_4 \rightarrow 11.$$

Now suppose that the string of events $x_1x_1x_2x_1$ is observed. Now $\boldsymbol{p}$ (given the data) is Dirichlet with parameters (4,2,1,1), and the conditional expected probability is given by $\hat{\boldsymbol{p}} = (\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8})$. The optimal code given the new information is

$$x_1 \rightarrow 0$$
$$x_2 \rightarrow 10$$
$$x_3 \rightarrow 110$$
$$x_4 \rightarrow 111.$$

## IV. Fixed Unknown $p$ and Sample $\{X_i\}$

Since $g(\boldsymbol{p})$ puts positive mass everywhere for any $\lambda \geq \mathbf{1}$, it follows that $\mathscr{C}^*(\hat{\boldsymbol{p}})$ is *admissible* in the sense that there exists no other encoding $\hat{\mathscr{C}}$ (based on $X_1,X_2,\cdots,X_n$) such that

$$E\{\hat{\mathscr{C}} \mid \boldsymbol{p}\} \leq E\{\mathscr{C}^*(\hat{\boldsymbol{p}}) \mid \boldsymbol{p}\}, \qquad \forall \boldsymbol{p}, \tag{13}$$

with sharp inequality for a set of nonzero Lebesgue measure. (Otherwise, taking expectations over $\boldsymbol{p} \sim g(\boldsymbol{p})$ yields a contradiction.) Thus

the encoding scheme based on a sample $n$ and *arbitrary* weighting $\lambda \geq 1$ is undominated by any other encoding.

## V. Conclusion

Gilbert's weighted-sample encoding scheme has been shown to be Bayes. Consequently, it is also admissible.

## References

[1] N. M. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1963.
[2] E. N. Gilbert, "Codes based on inaccurate source probabilities," *IEEE Trans. Inform. Theory*, vol. IT-17, May 1971, pp. 304–314.

## Correction to "Detectors, Bandpass Nonlinearities, and Their Optimization: Inversion of the Chebyshev Transform"

NELSON M. BLACHMAN

In the above paper,[1] on page 402, line 6, $v0(a)$ should read $v_m(a)$. On page 403, the bracketed denominator of (30) should be squared. The upper limit on the integral in ($T20$) of Table I should be $u$ rather than $\mu$. The name in footnote 4 should be Schlömilch.

In connection with the closing sentence of the paper, readers are referred to the results of recent work.[2]

Manuscript received July 9, 1971.
[1] N. M. Blachman, *IEEE Trans. Inform. Theory*, vol. IT-17, July 1971, pp. 398–404.
[2] ——, "The SNR performance of optimum nonlinearities," *Electron. Lett.*, vol. 7, July 15, 1971, pp. 393–395.

## Correction to "Independence of Measurements and the Mean Recognition Accuracy"

B. CHANDRASEKARAN

In the above paper,[1] on page 454, (16) should read as follows.

$$\Pr\left(\prod_{i=1}^{N}((s_i+1)/(m_1+2)) > \prod_{i=1}^{N}((r_i+1)/(m_2+2)) \mid \{p\},\{q\}\right) = \sum_{\text{over } S>R} \Pr(s_1,\cdots,s_N,r_1,\cdots,r_N \mid \{p\},\{q\}), \quad (16)$$

where

$$S = \prod_{i=1}^{N}((s_i+1)/(m_1+2))$$

$$R = \prod_{i=1}^{N}((r_i+1)/(m_2+2)).$$

Equation (17) should then become

$$P_{cr}(x) = \tfrac{1}{2}\{E[p_1\cdots p_N \Pr(S > R) \mid \{p\},\{q\}] + E[q_1\cdots q_N \Pr(S \leq R) \mid \{p\},\{q\}]\}. \quad (17)$$

Equations (18) and (19), as well as the results of the paper, remain unchanged.

Manuscript received July 13, 1971.
The author is with the Department of Computer and Information Science, Ohio State University, Columbus, Ohio 43210.
[1] B. Chandrasekaran, *IEEE Trans. Inform. Theory*, vol. IT-17, July 1971, pp. 452–456.

# Book Review

**Rate Distortion Theory: A Mathematical Basis for Data Compression**—Toby Berger (Englewood Cliffs, N.J.: Prentice-Hall, 1971, 311 pp., $16.95).

ROBERT M. GRAY

The Shannon theory of information is concerned essentially with two fundamental problems: source coding or "what information do I wish to send?," and channel coding or "how do I send it?" The latter problem has received by far the most attention in terms of theorems, algorithms, papers, and books. The former problem—source coding with a fidelity criterion or rate distortion theory—has been relatively neglected until recently. Rate distortion theory can be considered either as an information theoretic approach to data compression or as the study of sending information through noisy channels at rates exceeding capacity. The basic results of rate distortion theory were presented in the two classic Shannon papers,[1,2] and for several years were slightly refined, embellished, and applied in scattered papers of others. These efforts reached a climax with the publication in 1968 of the then most up to date and complete tour of rate distortion theory—Chapter 9 of Gallager's excellent book.[3] Gallager presented Shannon's basis results plus several of his own and of others in a concise unified treatment. However, as Wyner pointed out in his review of Gallager's book, the treatment was a bit too concise to be easily used for teaching purposes.[4]

Now at least an entire book devoted to rate-distortion theory has appeared—*Rate Distortion Theory, a Mathematical Basis for Data Compression*, by Toby Berger. This book is designed to serve two main purposes: 1) to present the "classical" results of the theory for independent letter sources in a tutorial fashion (Chapters 1–3, Sections 4.1–4.4) and 2) to collect and present in a unified manner the many important results derived since the publication of Gallager's book. Berger succeeds admirably in achieving both these goals.

In his book Berger proves himself to be an excellent expository writer. The many discussions providing motivation for the theorems and the techniques used are obviously well thought out and quite helpful. In addition, the deeper results are followed by heuristic explanations, which often succeed in getting across ideas that are fundamental but somewhat buried in the mathematical manipulations

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948, pp. 379–423, 623–656.
[2] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, pt. 4, Mar. 1959, pp. 142–163.
[3] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[4] A. D. Wyner, "Book review of *Information Theory and Reliable Communication* by R. G. Gallager," *IEEE Trans. Inform. Theory*, vol. IT-16, Jan. 1970, pp. 103–104.