# Capacity and codes for embedding information in gray-scale signals

[13] P. Kazakov, "Application of polynomials to CRC and spherical codes," Ph.D. dissertation, Tech. Univ. Delft, Delft, The Netherlands, 2000.
[14] T. Kløve, "Reed–Muller codes for error detection: The good, the bad, and the ugly," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1615–1622, Sep. 1996.
[15] T. Kløve and V. Korzhik, *Error Detecting Codes, General Theory and Their Application in Feedback Communication Systems*. Boston, MA: Kluwer, 1995.
[16] S. K. Leung-Yan-Cheong, E. R. Barnes, and D. U. Friedman, "On some properties of the undetected error probability of linear codes," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 1, pp. 110–112, Jan. 1979.
[17] S. K. Leung-Yan-Cheong and M. E. Hellman, "Concerning a bound on undetected error probability," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 235–237, Mar. 1976.
[18] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam, The Netherlands: North-Holland, 1977.
[19] J. Massey, "Coding techniques for digital data networks," in *Proc. Int. Conf. Information Theory and Systems*, vol. 65, Berlin, Germany, Sep. 18–20, 1978.

# Capacity and Codes for Embedding Information in Gray-Scale Signals

Frans M. J. Willems, *Fellow, IEEE,* and Marten van Dijk

***Abstract*—Gray-scale signals can be represented as sequences of integer-valued symbols. If such a symbol has alphabet $\{0, 1, \ldots, 2^B - 1\}$ it can be represented by $B$ binary digits. To embed information in these sequences, we are allowed to distort the symbols. The distortion measure that we consider here is squared error, however, errors larger than $m$ are not allowed. The embedded message must be recoverable with error probability zero. In this setup, there is a so-called "rate–distortion function" that tells us what the largest embedding rate is, given a certain distortion level and parameter $m$. First, we determine this rate–distortion function for $m = 1$ and for $m \to \infty$. Next we compare the performance of "low-bits modulation" to the rate–distortion function for $m \to \infty$. Then embedding codes are proposed based on i) ternary Hamming codes and on the ii) ternary Golay code. We show that all these codes are optimal in the sense that they achieve the smallest possible distortion at a given rate for fixed block length for any $m$.**

***Index Terms*—Data embedding, embedding distortion, gray-scale symbols, low-bits modulation, rate–distortion function, side information, squared-error distortion.**

## I. Introduction

In 1999, it was observed that data embedding is closely related to the information-theoretical concept of "channels with side information. " For example, Chen [2], Chen and Wornell [3], and Moulin and O'Sullivan [9] realized that (in the Gaussian case) there is a connection between data embedding and Costa's "writing on dirty paper" [4]. Costa's achievability proof can be seen as a special case of the proof of Gelfand and Pinsker [7]. Heegard and El Gamal [8] studied codes based on Gelfand–Pinsker theory for computer memories with defects.

Coding theorems for data-embedding situations appeared in Chen [2] (specialized to the Gaussian case), Moulin and O'Sullivan [9], Barron [1], and Willems [10].

In the present correspondence, we will focus on the coding theorem for the case of embedding in gray-scale symbols with squared-error distortion. We focus on the information embedding in the absence of communication noise. This makes it possible to achieve zero probability of error. Although the "noise-free" setup was also investigated by Chen [2] and Barron [1], the result that we present here is somewhat stronger than theirs as we will soon see. After proving our coding theorem we will propose some coding techniques.

Manuscript received August 26, 2002; revised May 27, 2004. The material in this correspondence was presented in part at the 39th Allerton Conference on Communication, Control, and Computing, Monticello, IL, October 2001.

F. M. J. Willems is with the Eindhoven University of Technology, and Philips Research Laboratories, Eindhoven, 5600 MB, The Netherlands (e-mail: f.m.j.willems@tue.nl).

M. van Dijk is with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA, and Philips Research Laboratories, Eindhoven, The Netherlands (e-mail: marten@csail.mit.edu).

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.842707

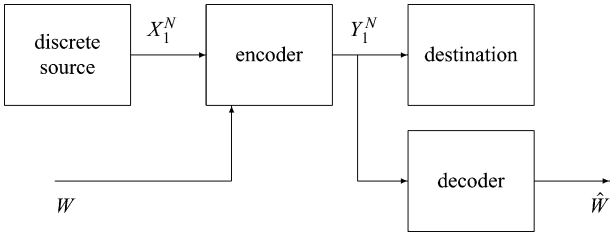


Fig. 1. A model of an information-embedding system.

## II. System Description

The embedding system that we study is shown in Fig. 1. A discrete source emits the *host sequence* $x_1^N = (x_1, x_2, \ldots, x_N)$ where $N$ is called the block length. The symbols $x_n$ for $n = 1, N$ assume values from the finite alphabet $\mathcal{X}$ which is a subset of the set $\mathbb{Z}$ of all integer numbers. We make no assumptions about the probability distribution $\{P(x_1^N), x_1^N \in \mathcal{X}^N\}$. A message source produces the *message index* $w \in \{1, 2, \ldots, M\}$ with probability $1/M$, independent of $x_1^N$. Based on the message index $w$ and the host sequence $x_1^N$ the encoder (embedder) produces the *composite sequence*

$$y_1^N = f(x_1^N, w). \tag{1}$$

The symbols from $y_1^N = (y_1, y_2, \ldots, y_N)$ assume values in the finite alphabet $\mathcal{Y}$ which is a subset of the set of all integers $\mathbb{Z}$. We require that from the composite sequence $y_1^N$ the embedded message can always be reconstructed without error, i.e., there is a decoder producing the estimate $\hat{w} = g(y_1^N)$ such that

$$\Pr\{\hat{W} \neq W\} = 0. \tag{2}$$

Moreover, the composite sequence $y_1^N$ must always be "close" to the host sequence $x_1^N$, i.e., the maximum average distortion

$$D_* = \max_{x_1^N} \sum_w \Pr\{W = w\} d(x_1^N, f(x_1^N, w)) \tag{3}$$

should be small. Here

$$d(x_1^N, y_1^N) \triangleq \frac{1}{N} \sum_{n=1,N} D(y_n - x_n) \tag{4}$$

is the distortion between the sequences $x_1^N$ and $y_1^N$, for some specified *distortion mapping* $D(\cdot)$ defined over the integers. Since both $y_n$ and $x_n$ are integers the difference $y_n - x_n$ is also an integer. Note that distortion measure is a *difference* measure. The error (difference) should not exceed the *maximum error* $m$, i.e., the distortion mapping satisfies $D(z) = \infty$ for integers $z \notin \mathcal{Z}$ where $\mathcal{Z} \subset \mathbb{Z}$ and $\mathcal{Z} \triangleq \{-m, 1 - m, \ldots, m\}$. Here $m$ is a positive integer. We can now be more specific about the reproduction alphabet which is defined as

$$\mathcal{Y} \triangleq \{x + z | x \in \mathcal{X}, z \in \mathcal{Z}\}. \tag{5}$$

Finally, we define the *embedding rate* $R$ as

$$R = \frac{1}{N} \log_2 M. \tag{6}$$

## III. The "Rate–Distortion" Function

We are obviously interested in finding codes that combine a large embedding rate $R$ with a small maximum average distortion $D_*$. We can now define a rate–distortion function in the following way. First, we define the distortion–rate pair $(\Delta, \rho)$ to be *admissible* if, for all $\epsilon > 0$ there exist, for all large enough $N$, encoders and decoders such that their embedding rate and maximum average distortion satisfy

$$\begin{aligned} R &\geq \rho - \epsilon \\ D_* &\leq \Delta + \epsilon. \end{aligned} \tag{7}$$

Only finite $\Delta$ are of interest. The rate–distortion function $\rho_m(\Delta)$ is now defined as the largest $\rho$ such that the pair $(\Delta, \rho)$ is admissible. The subscript $m$ specifies the maximum error that is allowed.

*Theorem 1:* For our rate–distortion function $\rho_m(\Delta)$ we can show that $\rho_m(\Delta) = r_m(\Delta)$ where

$$r_m(\Delta) \triangleq \max_{\{P(z):\sum_{z\in\mathcal{Z}} P(z)D(z)\leq\Delta\}} H(Z). \tag{8}$$

Here $\{P(z), z \in \mathcal{Z}\}$ is a probability distribution over the set $\mathcal{Z} = \{-m, 1-m, \ldots, m\}$.

From this theorem it follows[1] that the rate–distortion function $\rho_m(\Delta)$ is nonnegative, nondecreasing in $\Delta$, and convex-$\cap$ in $\Delta$.

The situation that we study here also was investigated by Barron [1] and Chen [2]. They refer to this case as the *noise-free case*. However, it should be noted that here we show that error probability 0 is achievable. Moreover, we measure our distortion averaged over all messages and maximized over the host sequences. Even when averaging over all messages is replaced by maximizing, our result holds. Finally, it should be noted that the distribution of $x_1^N$ is not relevant for our result.

## IV. Proof of the Theorem

### A. Admissibility Proof

*A:* We start by considering the side-information model depicted in Fig. 2. The sets $\mathcal{U}$ and $\mathcal{S}$ are assumed to be finite. Now words $u \in \mathcal{U}$ can be written on a medium. However, this medium generates the state $s \in \mathcal{S}$ and accepts only words $u \in \mathcal{U}_s \subseteq \mathcal{U}$ in that case. It is assumed that

$$|\mathcal{U}_s| = A \tag{9}$$

so whatever the actual state is, there are always $A$ words that can be written onto the medium. To transmit the message $w \in \{1, 2, \ldots, M\}$, assuming that $M \leq A$, the writer produces the word $u = f(w, s) \in \mathcal{U}_s$. The reader inspects $u$ and determines $\hat{w} = g(u)$. No errors are allowed thus, it is required that always $\hat{w} = w$.

How large can the number of messages $M$ be now? We can find a lower bound to this number by applying *random labeling*. We can give

[1]The convexity follows from assuming that if $P'(z)$ achieves maximum entropy $H'(Z)$ for distortions $\leq \Delta'$ and $P''(z)$ achieves maximum entropy $H''(Z)$ for distortions $\leq \Delta''$ then $\alpha P'(z) + (1-\alpha)P''(z)$ for $0 < \alpha < 1$ achieves distortion $\leq \alpha\Delta' + (1-\alpha)\Delta''$ and entropy not smaller than $\alpha H'(Z) + (1-\alpha)H''(Z)$ by the convexity of entropy in the distribution.

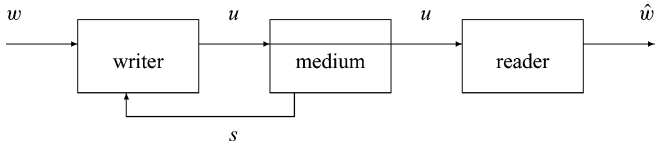


Fig. 2. A model for writing with side-information.

each word $u \in \mathcal{U}$ a message label $w \in \{1, \ldots, M\}$. Consider all such labelings. The *total* number of labelings $T$ is

$$T = M^{|\mathcal{U}|}. \tag{10}$$

*Bad* labelings do not have a word $u \in \mathcal{U}_s$ with label $w$ for at least one $(w, s)$-pair. Hence for the number of bad labelings $B$ we get

$$\begin{aligned} B &\leq \sum_w \sum_s (M-1)^A M^{|\mathcal{U}|-A} \\ &= M|\mathcal{S}|(M-1)^A M^{|\mathcal{U}|-A} \leq A|\mathcal{S}|(M-1)^A M^{|\mathcal{U}|-A}. \end{aligned} \tag{11}$$

Now at least one good labeling exists if $B < T$ hence if there are fewer bad labelings than the total number of labelings. Thus, there is at least one good labeling if

$$\ln(A|\mathcal{S}|) + A\ln\left(\frac{M-1}{M}\right) < 0. \tag{12}$$

Using $\ln(\frac{M-1}{M}) \leq -\frac{1}{M}$, it follows that there exists a good labeling if

$$\ln(A|\mathcal{S}|) - \frac{A}{M} < 0 \tag{13}$$

or, equivalently, if

$$M < \frac{A}{\ln(A|\mathcal{S}|)}. \tag{14}$$

Note that (14) is independent of $\mathcal{U}$. The number of messages $M$ that can be conveyed is essentially determined by $A$, the number of sequences that can be written for any $s$.

*B:* Fix the distribution $\{P(z), z \in \mathcal{Z}\}$. This distribution determines the entropy and the expected distortion

$$H(Z) = \sum_{z\in\mathcal{Z}} P(z)\log_2\frac{1}{P(z)}$$

and

$$E[D(Z)] = \sum_{z\in\mathcal{Z}} P(z)D(z). \tag{15}$$

Next fix the constant $0 < \delta < 1/2$ and consider the set of $\delta$-typical $Z$-sequences which is defined as (16) at the bottom of the page. Then we know (see, e.g., Cover and Thomas [5, Ch. 3]) that $\Pr\{Z_1^N \in \mathcal{T}_\delta\} \geq 1 - \delta$ for all sufficiently large $N$. This leads to the conclusion that

$$|\mathcal{T}_\delta| \geq \frac{1/2}{2^{-N(H(Z)-\delta)}} \geq 2^{N(H(Z)-\delta)-1} \geq 2^{N(H(Z)-2\delta)} \tag{17}$$

for all sufficiently large $N$.

$$\mathcal{T}_\delta \triangleq \left\{ z_1^N : \left|\frac{1}{N}\log_2\frac{1}{P^N(z_1^N)} - H(Z)\right| \leq \delta \wedge \left|\frac{1}{N}\sum_{n=1,N} D(z_n) - E[D(Z)]\right| \leq \delta \right\}. \tag{16}$$

*C:* Now we combine the findings of A and B. For any sequence $x_1^N$ there are at least $A = 2^{N(H(Z)-2\delta)}$ sequences $y_1^N = x_1^N + z_1^N$ with distortion

$$d(x_1^N, y_1^N) = \frac{1}{N}\sum_{n=1,N} D(y_n - x_n) = \frac{1}{N}\sum_{n=1,N} D(z_n) \le E[D(Z)] + \delta. \tag{18}$$

Consequently, $D_* \le E[D(Z)] + \delta$. By the random labeling argument we know that

$$\begin{aligned}&\frac{1}{N}\log_2 M\\ &< \frac{1}{N}\log_2 A - \frac{1}{N}\log_2 \ln(A|\mathcal{S}|)\\ &= H(Z) - 2\delta + \frac{1}{N}\log_2(N(H(Z)-2\delta)\ln 2 + N\ln|\mathcal{X}|) \qquad (19)\\ &= H(Z) - 2\delta - \frac{1}{N}\log_2 N - \frac{1}{N}\log_2[(H(Z)-2\delta)\ln 2 + \ln|\mathcal{X}|] \qquad (20)\end{aligned}$$

should hold (here $\mathcal{X}^N$ plays the role of $\mathcal{S}$ and $\mathcal{Y}^N$ that of $\mathcal{U}$). This condition is satisfied for all $N$ large enough as long as

$$\frac{1}{N}\log_2 M \le H(Z) - 3\delta. \tag{21}$$

Note that this proves the admissibility of pairs $(E[D(Z)], H(Z))$.

*D:* Actually, we have even proved more than that. Observe that not only the maximal average distortion $D_*$ is bounded by $E[D(Z)] + \delta$, but that this bound also holds for the maximal distortion

$$\max_{x_1^N, w} d(x_1^N, f(x_1^N, w)).$$

### B. Converse

Suppose that $x_1^N$ gives rise to the maximum average distortion $D_*$ which is finite. Fix this $x_1^N$. Now we introduce the random variable $Z$ that assumes integer values and has distribution

$$\Pr\{Z = z\} = \frac{1}{N}\sum_{n=1,N} \Pr\{Y_n - x_n = z | X_1^N = x_1^N\} \tag{22}$$

for integer $z$. Note that the probability is with respect to the message $W$. For this random variable, we can write

$$\begin{aligned}D_* &= \sum_w \Pr\{W = w\} d(x_1^N, f(x_1^N, w)) = E[d(x_1^N, Y_1^N)]\\ &= \frac{1}{N}\sum_{n=1,N}\sum_z \Pr\{Y_n - x_n = z | X_1^N = x_1^N\} D(z)\\ &= \sum_z \Pr\{Z = z\} D(z). \qquad (23)\end{aligned}$$

Since $D_* < \infty$, this implies that $\Pr\{Z = z\} = 0$ for $z \notin \mathcal{Z}$. Moreover

$$\begin{aligned}\log_2 M &= H(W) = H(W|X_1^N = x_1^N)\\ &= H(W|X_1^N = x_1^N) - H(W|X_1^N = x_1^N, Y_1^N)\\ &= I(W; Y_1^N | X_1^N = x_1^N) \le H(Y_1^N | X_1^N = x_1^N)\\ &= H(Y_1^N - x_1^N | X_1^N = x_1^N)\\ &\le \sum_{n=1,N} H(Y_n - x_n | X_1^N = x_1^N) \le NH(Z) \qquad (24)\end{aligned}$$

where the last inequality follows from the convexity of the entropy in the probability distribution. Note that $H(W|X_1^N = x_1^N) = H(W)$ by the independence of $W$ and $X_1^N$. Moreover, $H(W|X_1^N = x_1^N, Y_1^N) = 0$ since $g(Y_1^N) = W$. This concludes the converse.

Although we have focussed on error probability equal to zero (see (2)) it can be shown by adapting the converse that, in the average-error case, we will not find larger values of $\rho_m(\Delta)$ for any $\Delta$.

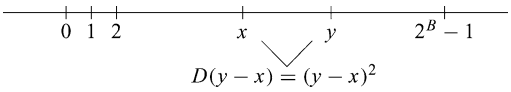


Fig. 3. The gray-scale alphabet $\mathcal{G}$ and squared-error distortion.

## V. Gray-Scale Symbols, Squared-Error Distortion

Gray-scale symbols assume values from an integer alphabet $\mathcal{G} = \{0, 1, \ldots, 2^B - 1\}$ for some positive integer $B$. Each symbol can be represented by a vector of $B$ binary digits. Now let

$$\mathcal{X} \triangleq \{m, m+1, \ldots, 2^B - 1 - m\}$$

then $\mathcal{Y} = \mathcal{G}$. This implies that $B \ge \log_2(2m+1)$. Note that in practise also host symbols smaller than $m$ and larger than $2^B - 1 - m$ can occur with positive probability. Here we assume that this is not the case! In Section IX, we will discuss this point in more detail. If a gray-scale symbol $x \in \mathcal{X}$ is changed into a symbol $y \in \mathcal{Y}$ the resulting *squared-error distortion* is

$$D(y-x) = \begin{cases}(y-x)^2, & |y-x| \le m\\ \infty, & \text{otherwise}\end{cases} \tag{25}$$

see Fig. 3. Again $m$ is the maximum allowable error.

We want to find out next how $r_m(\Delta)$ behaves. Note that by definition $r_{m'}(\Delta) \le r_{m''}(\Delta)$ if $m' < m''$ for all finite $\Delta$. Therefore, first we consider the case where all errors are allowed, thus, $m = \infty$. Therefore, let $p_z \triangleq P(z)$ for all integer $z$, then

$$H(Z) = \sum_z p_z \log_2 \frac{1}{p_z} \quad \text{and} \quad E[D(Z)] = \sum_z p_z z^2. \tag{26}$$

We now must maximize $H(Z)$ under the constraint $E[D(Z)] \le \Delta$. Therefore, consider for some $\alpha > 0$ the "Gaussian" distribution $\{p_z^*, z \in \mathbb{Z}\}$ where

$$p_z^* \triangleq \exp(-\alpha z^2)/\beta \tag{27}$$

with $\beta = \sum_z \exp(-\alpha z^2)$, having variance $\sum_z p_z^* z^2 = \Delta$. Then, for any distribution $\{p_z, z \in \mathbb{Z}\}$ with variance $\sum_z p_z z^2 \le \Delta$, we can write

$$\begin{aligned}\sum_z p_z \ln\frac{1}{p_z^*} &= \sum_z p_z \ln(\beta \exp(\alpha z^2))\\ &= \ln\beta + \alpha\sum_z p_z z^2 \le \ln\beta + \alpha\Delta \qquad (28)\end{aligned}$$

and, therefore, its entropy (in nats)

$$\sum_z p_z \ln\frac{1}{p_z} = \sum_z p_z \ln\frac{p_z^*}{p_z} + \sum_z p_z \ln\frac{1}{p_z^*} \le \ln\beta + \alpha\Delta. \tag{29}$$

Note that the term $\sum_z p_z \ln p_z^*/p_z$ is minus a divergence and therefore nonpositive. Equality is achieved only for $\{p_z, z \in \mathbb{Z}\}$ equal to the Gaussian distribution $\{p_z^*, z \in \mathbb{Z}\}$. Therefore, to determine the rate–distortion function $r_\infty(\Delta)$ we only need to vary $\alpha$ and compute the entropy and variance of $\{p_z^*, z \in \mathbb{Z}\}$. We can now make a plot of the squared-error rate–distortion function $r_\infty(\Delta)$, see Fig. 4. This plot shows that to achieve an embedding rate of 1 bit/symbol we need a maximum average distortion of at least $\approx 0.22$. Note that the plot shows that this statement holds for all $m$.

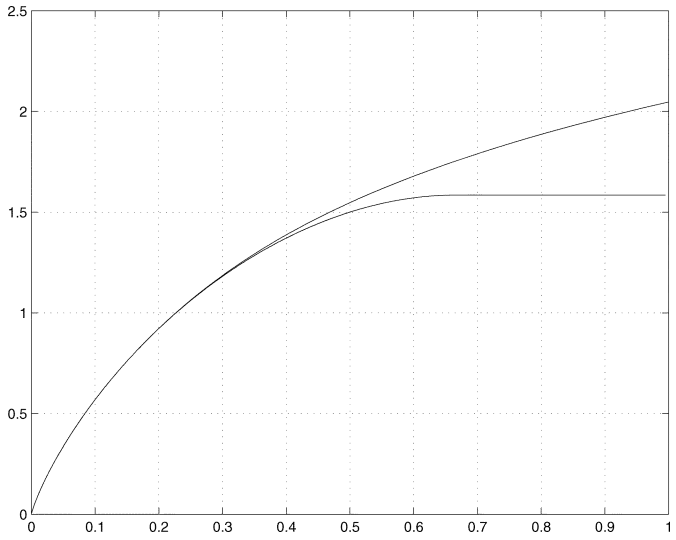

Fig. 4. Rate-distortion functions $r_\infty(\Delta)$ (upper curve) and $r_1(\Delta)$ (lower curve) for squared-error and difference-one distortion. On the horizontally axis is the distortion, vertically the rate in bits/symbol.

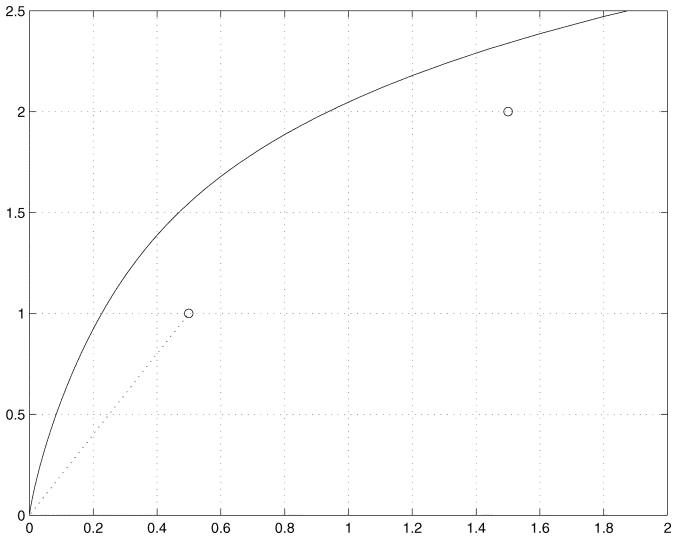

Fig. 5. Rate-distortion function $r_\infty(\Delta)$ together with two LBM distortion–rate pairs (o), and time-sharing the $R = 1$ LBM method (…).

Next we consider the case where $m = 1$. Now the error cannot be larger than one, hence,

$$D(y-x) = \begin{cases} 0, & \text{for } y = x \\ 1, & \text{for } |y-x| = 1 \\ \infty, & \text{else.} \end{cases} \tag{30}$$

This measure could be called "difference-one" distortion. It can be shown that the rate–distortion function for this measure is

$$r_1(\Delta) = \begin{cases} h(\Delta) + \Delta, & \text{for } 0 \le \Delta \le 2/3 \\ \log_2 3, & \text{for } \Delta > 2/3 \end{cases} \tag{31}$$

where $h(\alpha) = -\alpha \log_2 \alpha - (1-\alpha)\log_2(1-\alpha)$ is the binary-entropy function. This follows from

$$\begin{aligned} H(Z) &= h(1-p_0) + (1-p_0)h\left(\frac{p_1}{1-p_0}\right) \\ &\le h(1-p_0) + (1-p_0) \\ &\le h(\Delta) + \Delta, \qquad \text{for } \Delta \le 2/3 \end{aligned} \tag{32}$$

if we note that $1-p_0 = p_{-1}+p_1 \le \Delta$. Equality appears for $p_{-1} = p_1$ and $1 - p_0 = \Delta$. Note that $h(\Delta) + \Delta$ achieves its maximum for $\Delta = 2/3$.

In Fig. 4, this rate–distortion function is plotted together with the squared-error rate–distortion function $\rho_\infty(\Delta)$. Although for given distortion $\Delta$ the squared-error rate–distortion function $r_\infty(\Delta)$ is larger than the error-one rate–distortion function $r_1(\Delta)$, it can be seen that the difference is very small for small values of the distortion level. This demonstrates that for small distortion levels in the squared-error case, no matter how large the maximum allowable error $m$ is, it is not very useful to consider codes that have $|y_n - x_n| > 1$ for some components $n$. This fact will be used when we construct codes in the next sections.

## VI. Low-Bits Modulation

Traditional embedding methods assume that a gray-scale symbol $x \in \mathcal{G}$ is represented as a binary vector $b_{B-1}, \ldots, b_1, b_0$ such that $x = \sum_{i=0,B-1} b_i 2^i$. Messages are now embedded only in the least significant bits (LSBs) of this binary representation, thus, $y$ is chosen as the symbol closest to $x$ such that its $R$ LSBs contain the message $w$ ($R$ is a positive integer). Therefore, this form of embedding is called low-bits modulation (LBM), see, e.g., Chen [2]. The following tables now show what happens for $R = 1$. Observe that for $R = 1$ the distortion $D_* = 1/2$.

| $x$ | $w =$ | 0 | 1 |
|---|---|---|---|
| … | | | |
| $8 = 1000$ | $y =$ | 8 | 9 |
| $9 = 1001$ | | 8 | 9 |
| … | | | |

| $x$ | $w =$ | 0 | 1 |
|---|---|---|---|
| … | | | |
| $8 = 1000$ | $D =$ | 0 | 1 |
| $9 = 1001$ | | 1 | 0 |
| … | | | |

Similarly, for $R = 2$ the distortion is $D_* = (0+1+1+4)/4 = 3/2$ since there is a message that achieves distortion 0, two messages achieve distortion 1, and a fourth message achieves distortion $2^2 = 4$. The distortion–rate pairs $(D_*, R) = (1/2, 1)$ and $(3/2, 2)$ are plotted in Fig. 5 denoted by o's. Using the LBM scheme $(D_*, R) = (1/2, 1)$ only for a fraction of the symbols, we achieve

$$R(D_*) = 2D_*, \text{ for } 0 \le D_* \le 1/2. \tag{33}$$

The resulting distortion–rate pairs are plotted in the figure with a dotted line. The problem we want to address next is: "How can we do better than $R(D_*)/D_* = 2$?" Note that the $R = 1$ LBM scheme can be operated with maximum error $m = 1$.

## VII. Ternary Embedding Methods

We start this section with a definition. A gray-scale symbol $x$ (or $y$) is said to be *in class* $c$ iff $x \bmod 3 = c$ where $c = 0, 1, 2$. Now we are ready to discuss uncoded ternary embedding and after that two coded ternary embedding methods.

### A. Uncoded Ternary Modulation

Suppose that message $w \in \{0, 1, 2\}$ is to be embedded in the gray-scale symbol $x$. Then the decoder determines the message $w$ simply by looking at the class of $y$. If $x$ is in class $w$ then $y = x$ is chosen by

the embedder, otherwise, the embedder changes $x$ into the symbol $y$ in class $w$, such that $|y-x|$ is minimal, see the following tables.

| $x$ | $w=$ | 0 | 1 | 2 |
|---|---|---|---|---|
| ... | | | | |
| 9 | $y=$ | 9 | 10 | 8 |
| 10 | | 9 | 10 | 11 |
| 11 | | 12 | 10 | 11 |
| ... | | | | |

| $x$ | $w=$ | 0 | 1 | 2 |
|---|---|---|---|---|
| ... | | | | |
| 9 | $D=$ | 0 | 1 | 1 |
| 10 | | 1 | 0 | 1 |
| 11 | | 1 | 1 | 0 |
| ... | | | | |

The obtained embedding rate $R=\log_2 3\approx 1.5850$. The corresponding distortion $D_*=2/3$. This results in the ratio $R/D_*=3/2\log_2 3\approx 2.3774$ which is already quite good! Note that the maximum error of uncoded ternary embedding $m=1$ again.

### B. Embedding Based on Ternary Hamming Codes

To describe the embedding code consider the $(13,10,3)$ ternary Hamming code. This code has 27 cosets. Associate a message index to each coset. Focus now on a certain message index $w\in\{0,1,\ldots,26\}$ and the corresponding coset $\mathcal{C}_w$. Consider the vector containing the classes $c(x_n)$ of the 13 host symbols $x_1,x_2,\ldots,x_{13}$. Denote this ternary vector by $(c_x)_1^{13}$. Now determine the class vector $(c_y)_1^{13}\in\mathcal{C}_w$ which is closest to $(c_x)_1^{13}$ in Hamming sense. To obtain the composite sequence $y_1,y_2,\ldots,y_{13}$ just replace $x_n$ by the closest symbol $y_n$ in class $(c_y)_n$ for $n=1,13$.

First we determine the maximum average distortion $D_*$ of this embedding method. The Hamming code is perfect and has minimum Hamming distance $d_{H,\min}=3$, thus, we will find a word $(c_y)_1^{13}\in\mathcal{C}_i$ at Hamming distance 1 from $(c_x)_1^{13}$ with probability $26/27$ or a word at distance 0 with probability $1/27$. If $d_H((c_x)_1^{13},(c_y)_1^{13})=1$ then by construction

$$\sum_{n=1,13}(y_n-x_n)^2=1.$$

Hence, $D_*=26/27\cdot 1/13=2/27\approx 0.0741$. The decoder first determines the vector $(c_y)_1^{13}$ by looking at all the components $y_1,y_2,\ldots,y_{13}$ of the composite sequence. Then it determines the coset to which $(c_y)_1^{13}$ belongs, hence, reliable transmission is possible with rate $R=(\log_2 27)/13\approx 0.3658$ bit/ symbol. Thus, we achieve $(D_*,R)=(0.0741,0.3658)$. The $R/D_*-$ ratio $=4.9378$ which is a factor $2.4689$ larger than LBM.

We can design a series of codes for modulating the class, based on ternary Hamming codes. For a given value $\mu=2,3,\ldots$, i.e., the number of parity-check equations, the codeword length is $(3^\mu-1)/2$. Therefore,

$$R=\frac{2\mu\log_2 3}{3^\mu-1}\quad\text{and}\quad D_*=\frac{2}{3^\mu}. \tag{34}$$

Hence,

$$R/D_*=\frac{\mu 3^\mu\log_2 3}{3^\mu-1} \tag{35}$$

see Fig. 6. Note that we can achieve an arbitrary large ratio $R/D_*$ by increasing $\mu$. Note also that the maximum error $m=1$ again.

### C. Embedding Using the Ternary Golay Code

Instead of a ternary Hamming code we can also use the $(11,6,5)$ ternary Golay code in an embedding system. Again, we put the message in the syndrome, which is five trits long, but now the class must be changed in at most two positions. This leads to the following rate and distortion:

$$R=\frac{5\log_2 3}{11}\approx 0.7204$$

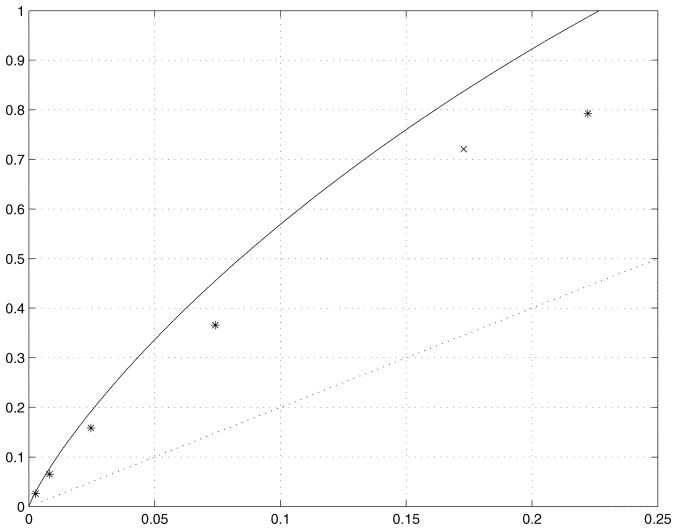


Fig. 6. Rate-distortion curve $r_\infty(\Delta)$, time-sharing $R=1$-LBM (...), ternary Hamming codes (*), and the ternary Golay code (×).

and

$$D_*=\frac{\binom{11}{1}\cdot 2\cdot 1+\binom{11}{2}\cdot 4\cdot 2}{243\cdot 11}=\frac{42}{243}\approx 0.1728. \tag{36}$$

Hence, $R/D_*\approx 4.1682$, see Fig. 6. Again $m=1$.

## VIII. Optimality

We call an embedding code with block length $N$, embedding rate $R$, and maximal average distortion $D_*$ *optimal*, if all other codes with the same block length $N$ have rate $R'\le R$ or distortion $D'_*\ge D_*$. We assume that distortion is of squared-error type as in (25), $m$ is arbitrary.

*A:* Now consider a code with block length $N$. Suppose that the source produces a host sequence $x_1^N$. Suppose that the embedding rate $R=\log_2(2N+1)/N$ so there are $2N+1$ composite sequences $y_1^N$ all representing a different message. Note that there is only one sequence $y_1^N=x_1^N$ such that $d(x_1^N,y_1^N)=0$. The smallest nonzero distortion $d(x_1^N,y_1^N)=1/N$ is achieved for $2N$ sequences $y_1^N$ that differ from $x_1^N$ in exactly one component $n\in\{1,2,\ldots,N\}$ by one, i.e., $|y_n-x_n|=1$. Therefore, the smallest possible maximum average distortion $D_*=2/(2N+1)$. Hence, the proposed embedding method based on ternary Hamming codes for $\mu=2,3,\ldots$ is optimal. Moreover, this holds for uncoded ternary transmission. This holds for any $m$.

*B:* Again consider a code with block length $N$ and assume that the source generates host sequence $x_1^N$. Suppose that the number of messages $M$ is $1+2N+4\binom{N}{2}=1+2N^2$. Again, there is only one sequence $y_1^N$ with distortion $d(x_1^N,y_1^N)=0$, and $2N$ sequences with distortion $d(x_1^N,y_1^N)=1/N$. The next smallest possible squared-error distortion $d(x_1^N,y_1^N)=2/N$ is achieved by $4\binom{N}{2}$ sequences $y_1^N$. Consequently, the rate $R=\log_2(1+2N^2)/N$ should lead to a maximum average distortion of at least $(4N-2)/(1+2N^2)$. The $(11,6,5)$ ternary Golay code achieves this bound and is therefore optimal for any $m$.

## IX. Remark

So far we have assumed that $\mathcal{X}=\{m,m+1,\ldots,2^B-1-m\}$. However, in practise a source will also produce gray-scale symbols smaller than $m$ and larger than $2^B-1-m$. Note, however, that in practise it suffices to take $m=1$, thus only the host symbol 0 and $2^B-1$ cause problems since the composite symbols $-1$ and $2^B$ are not in the gray-scale alphabet $\mathcal{G}$. The strategies based on the (optimal)

ternary codes that we proposed here can however be adapted slightly. Instead of composite value $-1$ we use 2 and instead of $2^B$ we take $2^B - 3$. This will lead to a larger distortion but if the probabilities of host symbols 0 and $2^B - 1$ are not too large the effect can be neglected.

## X. Conclusion

For gray-scale symbols and squared-error we have determined the rate–distortion function for maximum error $m = 1$. We have also determined $r_\infty(\Delta)$ which serves as an upper bound for all $r_m(\Delta)$. Moreover, we have constructed embedding codes based on ternary error-correcting codes. We have only looked at small distortions (and rates). We have concentrated on perfect codes since these codes result in simple schemes that are easy to analyze. We have shown that the proposed codes are optimal. More codes can be found in [6].

## Acknowledgment

We thank the reviewers for their comments.

## References

[1] R. J. Barron, "Systematic hybrid analog/digital signal coding," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[2] B. Chen, "Design and analysis of digital watermarking, information embedding, and data hiding systems," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[3] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[4] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[6] M. van Dijk and F. M. J. Willems, "Embedding information in gray-scale images," in *Proc. 22nd Symp. Information Theory in the Benelux*, Enschede, The Netherlands, May 15–16, 2001, pp. 147–154.

[7] S. Gelfand and M. Pinsker, "Coding for a channel with random parameters," *Probl. Control Inf. Theory*, vol. 9, pp. 19–31, 1980.

[8] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.

[9] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," preprint, 1999.

[10] F. M. J. Willems, "An information-theoretical approach to information embedding," in *Proc. 21st Symp. Information Theory in the Benelux*, Wassenaar, The Netherlands, May 25–26, 2000, pp. 255–260.

# A Comment on "Systematic Single Asymmetric Error-Correcting Codes"

Ching-Nung Yang and Guo-Jau Chen

***Abstract*—Bose and Al-Bassam designed quasi-systematic single asymmetric error-correcting codes that are able to encode $k$ information bits in a systematic way, but cannot encode all $2^k$ information words. However, "*Construction I*" of their paper does not provide an efficient way for computing the number of codewords. In this correspondence, we derive a recursive relation based on generating functions in order to compute the maximum number of codewords in such codes. This recursion allows us to determine those numbers for large code lengths which may not be feasible otherwise.**

***Index Terms*—Abelian group, asymmetric code, generating function, systematic code.**

## I. Introduction

A single asymmetric error-correcting code (SAEC) for the binary asymmetric channel is capable of correcting one asymmetric error (i.e., $1 \rightarrow 0$ error or $0 \rightarrow 1$ error). At present, there are some codes which can be used as SAECs, such as Varshamov codes [3] and Constantin–Rao codes [2]. In general, nonsystematic codes have a better coding efficiency than systematic codes. However, systematic codes often are less complex in encoding/decoding. In [1], Bose and Al-Bassam show how to encode a nonsystematic SAEC in a systematic way to achieve a high code rate, and at the same time have the benefits of systematic codes. Such Bose–Al-Bassam codes are quasi-systematic.

To understand the concept of the Bose and Al-Bassam codes, we first review the nonsystematic Constantin–Rao SAEC [2]. Given the Abelian group $G = \{g_0, \ldots, g_n\}$ of order $(n+1)$ and $g_0 = 0$ (the identity element), let $V$ be the set of all binary $n$-tuples, and define a function $f : V \rightarrow G$ as $f(\underline{x}) = \sum_{i=1}^{n} x_i \cdot g_i$ for any $\underline{x} = (x_1, x_2, \ldots, x_n) \in V$. Then, $f$ partitions $V$ into $n+1$ disjoint subsets, $V_l = \{\underline{x} \in V | f(\underline{x}) = g_l\}$, for $l = 0, \ldots, n$, and each subset $V_l$ is a nonsystematic Constantin–Rao SAEC. The Varshamov code [3] is equivalent to the Constantin–Rao code when $G$ is chosen as the cyclic group $Z_{n+1} = \{0, 1, \ldots, n\}$. The following lemma shows which subset has the largest cardinality.

*Lemma 1:* The nonsystematic Constantin–Rao SAEC using the Abelian Group $G$ of order $n+1$ given by $V_0$ has the maximum number of codewords [2].

Going back to the Bose–Al-Bassam codes now, we briefly describe *Construction I* in [1]. Consider the case where $n = 2^r$ and $k = n - r$, where $n$, $k$, and $r$ are the lengths of the codeword, information word, and checking vector, respectively. Let $E = \{0,1\}^k$ be the set of all $k$-bit information words, and $C = \{0,1\}^r$ be the set of all $r$-bit checking vectors. Let $Z'' = (z_1'' z_2'', \ldots, z_r'') = (1, 2, 4, \ldots, 2^{r-1})$ and $Z' = (z_1', z_2', \ldots, z_{2^r - r}')$ consisting of all the elements in $Z_{n+1}$ except the elements in $Z''$ and zero.

Manuscript received June 11, 2003; revised October 29, 2004.

The authors are with the Department of Computer Science and Information Engineering, National Dong Hwa University, Shoufeng, 97401 Taiwan, R.O.C. (e-mail: cnyang@mail.ndhu.edu.tw).

Communicated by K. A. S. Abdel-Ghaffar, Associate Editor for Coding Theory.

Digital Object Identifier 10.1109/TIT.2004.842760