(Reprint [with correction of some typographical errors] of pp. 1-17 in *Advanced Methods for Satellite and Deep Space Communications* (Ed. J. Hagenauer) , Lecture Notes in Control and Information Sciences 182. Heidelberg and New York: Springer, 1992)

Deep-Space Communications and Coding: A Marriage Made in Heaven

James L. Massey
Signal and Information Processing Laboratory
Swiss Federal Institute of Technology
CH-8092 Zürich, Switzerland

1 Introduction

It is almost a quarter of a century since the launch in 1968 of NASA's Pioneer 9 spacecraft on the first mission into deep-space that relied on coding to enhance communications on the critical downlink channel. [The channel code used was a binary convolutional code that was decoded with sequential decoding--we will have much to say about this code in the sequel.] The success of this channel coding system had repercussions that extended far beyond NASA's space program. It is no exaggeration to say that the Pioneer 9 mission provided communications engineers with the first incontrovertible demonstration of the practical utility of channel coding techniques and thereby paved the way for the successful application of coding to many other channels.

Shannon, in his 1948 paper [1] that established the new field of information theory, gave a mathematical proof that every communications channel could be characterized by a single parameter C, the channel capacity, in the manner that information could be sent over this channel to a destination as reliably as desired at any rate R (measured, say, in information bits per second) provided that $R < C$, but that for any rate R greater than C there was an irreducible unreliability for information transmission. Shannon's work showed that, to achieve efficient (i. e., R close to C) and reliable use of a channel, it was necessary in general to "code" the information for transmission over the channel in the sense that each information bit must influence many transmitted digits. Almost immediately, there began an intensive search (that still continues unabated) by many investigators to find good channel codes. Unfortunately, most of these researchers concentrated (and still do concentrate) on the "wrong" channel, viz. on the binary symmetric channel (BSC) [or its non-binary equivalents]. The BSC has both a binary input alphabet and a binary output alphabet and is characterized by a single parameter p $(0 \leq p \leq 1/2)$ in the manner that each transmitted binary digit has probability p of being changed in transmission, independently of what has happened to the previous transmitted digits (i. e., the channel is memoryless). It was natural then to think of such a channel as introducing "errors" into the transmitted data stream and to suppose that it was the purpose of the coding system to "correct" these errors. The term "error-correcting code" came into (and remains) in widespread use to describe such channel codes--although with scarcely any reflection one must conclude that one cannot even talk about "errors" in transmission unless the channel input and output alphabets coincide (as unhappily they do for the BSC). More circumspect writers have adopted the term "error-control code" in place of "error-correcting code," but this is only a trifle less misleading. Where are the errors to be controlled? It seems to us much wiser to use the less suggestive, but more precise, term "channel code" to describe the code used to map the information bits into the sequence of digits to be transmitted over the channel, as we have done already in our opening paragraph.

There had been attempts prior to 1968 to make practical use of channel coding. Codex Corporation, founded in 1962 in Cambridge, Mass., became the first company dedicated to this goal and also the first to encounter the widespread skepticism among communications engineers about the practicality of channel coding. The considerable commercial success that has been enjoyed by this company (which is now a division of Motorola, Inc.) stems less from its pioneering activity in channel coding than from its judicious decision in the late 1960's to expand its technical activity into the development of high-speed modems for telephone channels.

Why did deep-space communications provide the setting for demonstrating the practical benefits of channel coding? Why were the "heavens" of deep-space virtually predestined to be the proving grounds for channel coding techniques? Why was this wedding of channel coding to the deep-space channel, in the words of our title, a "marriage made in heaven"? There are many reasons, the most important of which are as follows:

- The deep-space channel is accurately described by a mathematical channel model, the additive white Gaussian noise (AWGN) channel, that was introduced by Shannon in 1948.

- It was well understood theoretically by the early 1960's what one must do to use this channel efficiently and reliably and what gains could be achieved by channel coding.

- The available bandwidth on the deep-space channel was so great that binary transmission could be efficiently used.

- The NASA communications engineers in the mid-1960's understood that, for efficient use of the deep-space channel, it was necessary to design the modulation system and the channel coding system in a coordinated way and they were willing to make the resulting necessary changes in the demodulators that they had previously been using.

- Good binary convolutional codes were available by the mid-1960's and, more importantly, an effective and practical algorithm was known for decoding these codes on the AWGN channel.

- The only complex part of the efficient channel coding systems for the AWGN channel that were known by the mid-1960's is the decoder, which for the downlink deep-space channel is located at the earth station where complexity is of much less importance than it is in the spacecraft.

- Every dB in "coding gain" on the downlink deep-space channel is so valuable (in the mid-1960's it was reckoned at about $1,000,000 per dB) that even a small gain, such as the 2.2 dB that was provided by the Mariner '69 channel coding system, was a strong economic incentive for developing and implementing such a channel coding system.

We will consider most of the above reasons in some detail in the sequel--there are lessons therein that are still of great relevance today and are all-too-often forgotten. We begin in the next section by taking a careful look at the deep-space channel itself, then considering the interplay between coding and modulation systems used on this channel. We also make a fairly intensive study of bandwidth issues for the deep-space channel that leads us to the important distinction between Fourier bandwidth and Shannon bandwidth. The final section is devoted to a detailed look at the two codes in question, the Pioneer 9 code and the Mariner '69 code, followed by a comparison of their merits. It is something of a quirk in technical history that the Pioneer 9 convolutional coding system became the first channel coding system to be used in deep-space, as this had not been intended by NASA. That honor had been planned for the binary block code used in NASA's Mariner spacecraft that was launched in 1969. Why the convolutional code nevertheless won the race into space is explained in the final paragraph of this "tale of two codes."

2 The Deep-Space Channel

2.1 Channel Capacity Considerations

We have already mentioned that the deep-space channel is accurately described by Shannon's additive white Gaussian noise (AWGN) channel model. The correspondence is so good in fact that no one has ever observed any deviation of the deep-space channel from this mathematical model. In this model, the received signal is the sum of the transmitted signal and a white Gaussian noise process of one-sided power spectral density N_o (watts/Hz).

The transmitted signal is constrained to lie in a Fourier bandwidth of W (Hz) or less and to have an average power of S (watts) or less. Shannon [1] computed the capacity of this channel to be

$$C_W = W \log_2 (1 + \frac{S}{W\ N_o}) \quad \text{(bits / sec)}, \tag{1}$$

which is one of the most famous formulas in communication theory--and also one of the most abused. It is obvious intuitively (increasing the available Fourier bandwidth can only help the sender) and easy to check mathematically that C_W increases monotonically with W, taking as its maximum the value

$$C_\infty = \frac{1}{\mathit{ln}\ 2} \frac{S}{N_o} \approx 1.44 \frac{S}{N_o} \quad \text{(bits / sec)}. \tag{2}$$

Suppose now that one is transmitting information bits at a rate R (bits / sec) very close to this maximum capacity. Then, because the power is S (joules / sec), the energy per information bit, E_b, is just $E_b = S / R \approx S / C_\infty = \mathit{ln}\ 2\ N_o \approx 0.69\ N_o$ (joules). Equivalently, the *signal-to-noise ratio* is

$$\frac{E_b}{N_o} \approx 0.69 \ \text{(or -1.6 dB)}, \tag{3}$$

which is the minimum signal-to-noise ratio required for arbitrarily reliable communication and is often referred to as the *Shannon limit* for the AWGN channel. All of this was well known in the early 1960's, cf. [4, p. 162].

2.2 The Interplay between Coding and Modulation Systems

Suppose now that digital transmission is used on the AWGN channel and that the modulator emits transmitted symbols [or, more precisely, the waveforms that represent these symbols] at the rate of r (symbols / sec). It follows that $S = r \times E = R \times E_b$, where E (joules / symbol) is the average energy of the waveforms used for a transmitted symbol, so that E_b and E are related as $E_b = E\,(r / R)$. Incidentally, one of the benefits derived from channel coding for the deep-space channel is that it accustomed perspicacious communications engineers to evaluate the performance of their communications systems in terms of the "true" signal-to-noise ratio defined as E_b / N_o, which is a fundamental performance parameter, rather than in terms of the transmitted signal-to-noise ratio, E / N_o, which is not fundamental at all for comparison of system performances--although it had been customarily so used (and is still so used unfortunately often).

Suppose next that there are q signals in the modulation *signal set*, i.e. in the set of waveforms used to represent the q different values of a modulation symbol. It is convenient to represent these signals or waveforms as vectors in n-dimensional Euclidean space in the manner introduced by Shannon [2] and exploited to great effect by Wozencraft and Jacobs [3], i. e., by their coefficient vectors with respect to their representation as a linear combination of the signals in some orthonormal set of n signals. Let $\mathbf{s}_0, \mathbf{s}_1, \ldots , \mathbf{s}_{q-1}$ so represent the modulation signal set. The binary one-dimensional (or scalar) signal set $\mathbf{s}_0 = +\sqrt{E}$ and $\mathbf{s}_1 = -\sqrt{E}$ is called the *binary antipodal signal set* and represents, for instance, the waveforms used in binary phase-shift-keying (BPSK) modulation. BPSK modulation is very attractive for use in space communications because its constant-envelope character greatly simplifies the required transmitter amplifying hardware in the spacecraft.

The use of digital modulation on the AWGN channel effectively converts the channel to a discrete-time additive Gaussian noise channel in which the received signal $\mathbf{r}$ in each modulation interval is the sum of the transmitted signal $\mathbf{s}_i$ and a noise vector $\mathbf{n}$ whose components are independent Gaussian random variables with 0 mean and variance $N_o/2$. The capacity C (bits/use) of this channel was also well-known in the early 1960's. In particular, it was known that if one uses a one-dimensional signal set according to a probability distribution on the signals such that the received signal $\mathbf{r}$ well-approximates a zero-mean Gaussian random variable, then (cf. [4, p. 147]) the capacity C is given by

$$C \approx \frac{1}{2}\log_2\left(1 + \frac{2E}{N_o}\right) \quad \text{(bits/use).} \tag{4}$$

One sees from (4) that C/E, the capacity per joule, decreases as E/No increases. In the region of energy-efficient operation, which is roughly the region $0 < E/N_o \leq 1/2$ (-3 dB), the capacity (4) becomes

$$C = 1.44\frac{E}{N_o} \quad \text{(bits/use).} \tag{5}$$

The corresponding capacity per unit of time is thus

$$rC = 1.44\frac{rE}{N_o} = 1.44\frac{S}{N_o} = C_\infty \quad \text{(bits/sec)} \tag{6}$$

where we have made use of (2). Thus, one sees that in the region of energy-efficient operation, which is roughly the region $0 < E/N_o \leq 1/2$, one pays no penalty in capacity for using one-dimensional digital modulation. If the one-dimensional signal set is the binary antipodal signal set and the two signals therein are equally likely, then the received signal $\mathbf{r} = \mathbf{s} + \mathbf{n}$ always has zero mean but will have the required approximately Gaussian distribution only if the standard deviation $\sqrt{N_o/2}$ of the Gaussian noise $\mathbf{n}$ is somewhat greater than the magnitude $\sqrt{E}$ of **s**, i. e., roughly again when $0 < E/N_o \leq 1/2$. The conclusion that *binary antipodal modulation is energy-efficient on the deep-space channel just when one-dimensional modulation is energy-efficient* i. e., when the transmitted signal-to-noise ratio E/N_o is about -3 dB or less, was well-known to information theorists in the early 1960's.

Suppose that information bits are sent *uncoded* over the deep-space channel with binary antipodal modulation, which implies that the number of modulation symbols per second, r, is equal to the number of information bits per second, R and hence that $E_b = E$. The probability of detecting the information bits at the receiver reduces to that of detecting equally likely binary antipodal signals, $+\sqrt{E_b}$ and $-\sqrt{E_b}$, in the presence of Gaussian noise with variance $N_0/2$, the error probability for which (cf. [3, p. 82]) is given by

$$P_b = Q(\sqrt{2E_b/N_o}\,) \tag{7}$$

where

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2}\, d\alpha < \frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2} \tag{8}$$

and where the inequality in (8) is a virtual equality for $x \geq 2$ (cf. [3, p. 83]). If one operates in the region of potentially efficient channel use, $E_b/N_o = E/N_o \leq 1/2$ (-3 dB), one sees from (7) and (8) that the information bits can be recovered at the receiver with an error probability of at best $P_b = Q(1) \approx 2.4 \times 10^{-1}$, which is orders-of-magnitude too large to be acceptable on the downlink deep-space channel or in almost any communications system for that matter. The inescapable conclusion is that *if one wants to signal both energy-efficiently and reliably with binary modulation on the deep-space channel, then one must use channel coding.* Indeed, it was the realization of this fact in the early 1960's that impelled NASA to begin to plan for the use of channel coding in the spacecraft in its Mariner series that would be launched in 1969. With the long lead time required to obtain space approval for design changes, it was already too late to influence spacecraft in the Mariner series with earlier launch dates--these spacecraft continued to use uncoded binary antipodal signalling on the downlink.

The relationship of demodulation to the channel coding system is much more subtle than that of modulation. The reason for this is that the demodulator can spoil the channel for decoding if one is not extremely careful about its design, a fact that is still not sufficiently appreciated. Before channel coding theory "reared its ugly head", communications engineers designed their digital demodulators to make optimum decisions about the transmitted modulation symbols, i. e. they did what is now generally called *hard-decision demodulation*. The combination of modulator, waveform channel and hard-decision demodulator creates a discrete channel whose input and output alphabets coincide. In particular, when binary antipodal modulation is used on the deep-space channel with hard-decision demodulation, the resulting discrete channel is the binary symmetric channel (BSC) about which we have already made disparaging remarks; the "error probability" p on this channel is given by p = $Q(2E/N_o)$. [One sees here that the BSC does not occur naturally in nature; it is created by the designers of energy-inefficient modulation systems!] Nonetheless, most communications engineers continued long after 1948 to assume (and many today still do assume) that the proper goal of of a digital demodulator is to make optimum decisions about the transmitted modulation symbols, i. e., to do hard-decision demodulation. This slothful thinking meshed very nicely with the "error-correcting codes" school of channel coding theorists since hard-decisions give the "errors" that they were eager to correct. It was, however, well-known to many information theorists in the early 1960's that hard-decision demodulation entails a substantial loss in capacity compared to what can be achieved by a more thoughtful form of demodulation. For instance, it was well-known that binary antipodal signalling on the deep-space channel used with hard-decision demodulation achieves a capacity smaller than C_∞ by a factor of $2/\pi$ (2.0 dB) in the energy-efficient range of operation, cf. [4, p. 211]. Because the coding system for Mariner '69 was calculated to provide only 2.2 dB of gain over uncoded binary transmission even when used with optimum demodulation, the use of hard-decision demodulation in Mariner '69 was out of the question!

What should a good demodulator do? Fano answered this question very well in the early 1960's: "the capacity of the discrete channel [that results from the combination of the digital modulator, waveform channel and demodulator] should not be unduly smaller than the capacity of the original [waveform] channel" [4, p. 211]. The real goal of the designer of the demodulation system should be to create a good channel for the channel coding system--a point that we have written about at greater length elsewhere [5]. If the capacity of the resulting discrete channel is taken as the design criterion, then one must conclude that the optimum demodulator is a straight wire because any quantization of the received signal can only reduce capacity. In fact, to ease the decoding problem for the channel code, one should in fact design the demodulator to make as coarse a quantization of the received signal as possible consistent with Fano's adage that "the capacity of the [resulting] discrete channel should not be unduly smaller than the capacity of the original [waveform] channel." Already in the early sixties it was well-known among some information theorists, cf. [6], that 8 demodulator quantization levels were enough, when used for binary antipodal signalling on the deep-space channel, to reduce the capacity loss to a negligible 0.1 dB in the energy efficient range $E/N_o \leq 1/2$ (-3 dB). However, using only 4 quantization levels would yield an additional loss of about 0.3 dB. It was natural then to choose 8 level demodulation, or "3-bit soft-decision demodulation" as it is generally called because the soft-decision demodulator's output can be thought to consist of the hard-decision binary digit together with 2 binary digits of information about the quality of this hard decision. Almost all coding systems that have been made for the deep-space channel and similar channels have operated with such 3-bit soft decisions.

2.3 Bandwidth Considerations

We have already observed that the user of the deep-space channel has virtually unlimited bandwidth at his disposal. This does not mean, of course, that he should use as much bandwidth as possible. Rather, analogous to Fano's dictum for demodulator quantization, he should in fact use as little bandwidth as possible consistent with not unduly decreasing the capacity, C_∞, of the original waveform channel. One reason for this is that the receiver's radio-frequency "front-end" must have as wide a bandwidth as the transmitted signal and, when this bandwidth becomes too great, the large Gaussian noise present at the receiver's front-end makes it virtually impossible to realize the coherent demodulation that is required to reap the theoretically available benefits of greater bandwith--a kind of "Catch-22" of large bandwidth. We will presently see another cogent and related reason for using as little bandwidth as possible consistent with not unduly decreasing capacity.

The *code rate* R (bits/digit) of a binary channel code is the average number of information bits per binary digit produced by the code (on a long-time basis). The minimal requirement that the encoding be invertible specifies that $R \leq 1$, where $R = 1$ corresponds to uncoded transmission. Suppose the code is used with binary modulation so that each encoded binary digit selects one modulation symbol. Then, because r is the number of modulation symbols per second, $\mathsf{R} = \mathsf{r} \times R$ is the information transmission rate in bits per second and is the number that is specified in advance to the designers of the communication system. But the Fourier bandwidth of the resulting sequence of waveforms is directly proportional to $\mathsf{r} = \mathsf{R} / R$ rather than to R itself, which leads to the inescapable conclusion that when designing a coding system for the deep-space channel, one should choose the maximum code rate consistent with not unduly decreasing the capacity per second of the corresponding discrete channel from its value C_∞ for the infinite-bandwidth waveform channel.

2.4 Fourier Bandwidth and Shannon Bandwidth

To proceed further with our consideration of bandwidth for the deep-space channel, it is helpful to make the important distinction between Fourier bandwidth and Shannon bandwidth. Shannon [1], [2] in fact identified bandwidth with the number of signal-set dimensions that are transmitted per second; we shall use the symbol B to denote this quantity that we will call the *Shannon bandwith* of the transmitted signal. If the modulation signal set is n-dimensional, then $B = \mathsf{r} \times n$, as follows from the fact that r is the number of modulation signals transmitted per second. Shannon's equation (1) can be written more fundamentally in terms of Shannon bandwidth as

$$C_B = \frac{1}{2} B \log_2 (1 + \frac{S}{B N_o/2}) \quad \text{(bits/sec).} \tag{9}$$

The consistency between (1) and (9) can be seen as follows. According to the Shannon-Nyquist sampling theorem (i.e., Shannon's completion [1] of the theorem begun by Nyquist [7]), at most 2W orthogonal signals can be sent per second in a frequency band of Fourier bandwidth W. Thus, $B \leq 2W$ with equality if the orthogonal signals are translates by $1/(2W)$ seconds of the familiar sinc signal that has a flat spectrum over the given frequency band, i. e., if the transmitted signals fill the available Fourier bandwidth as completely as possible. Using this maximum $B = 2W$ in (9) yields (1), as indeed it must because C_B as given by (9) increases monotonically with B and hence, if a constraint W on the Fourier bandwidth is given, one must choose the maximum Shannon bandwidth B consistent with that constraint. Examination of Shannon's arguments in [1] show in fact that he first proved the capacity formula (9), then made use of the sampling theorem to deduce (1). Equation (9) is indeed the more fundamental of these two capacity formulas since it holds regardless of whether or not one chooses modulation signals that completely fill out the Fourier bandwidth. Communications engineers must live with the constraints on Fourier bandwidth specified by regulatory agencies, but the performance of their systems depends much more on their Shannon bandwidth rather than on their Fourier bandwidth.

We are finally in position to see how the code rate R affects the capacity of the deep-space channel. Because binary antipodal signalling uses a one-dimensional ($n = 1$) signal set, we have $B = r$ and thus Shannon's capacity formula (9) becomes

$$C_r = \frac{1}{2} r \log_2 (1 + \frac{S}{r N_o/2}) \quad \text{(bits/sec).}$$

Recalling that $S = r \times E = r \times R \times E_b$, we see that this can be rewritten as

$$C_r = \frac{1}{2} \frac{S}{R E_b} \log_2 (1 + \frac{R E_b}{N_o/2}) \quad \text{(bits/sec).} \tag{10}$$

Letting the code rate R tend to 0, we see that C_r tends to the limit C_∞ given by (2)--as of course it must since, for fixed signal power S and fixed energy per information bit E_b, $B = r$ tends to infinity as R tends to 0. Moreover, we see that, for any given non-zero rate R, the capacity reduction factor $\gamma = C_r/C_\infty$ due to the resulting finite bandwidth is given by

$$\gamma = \frac{\mathit{ln}\,(1 + \frac{R E_b}{N_o/2})}{\frac{R E_b}{N_o/2}} . \tag{11}$$

Suppose next that we are operating near the Shannon limit, i. e., that $E_b/N_o \approx \mathit{ln}\ 2 \approx 0.69$. The capacity reduction factor γ then becomes

$$\gamma \approx \frac{\mathit{ln}\ (1 + 1.386\ R)}{1.386\ R}\ , \tag{12}$$

which is our desired result for specifying how much loss will be suffered due to finite bandwidth by a coding system that operates near the Shannon limit.

For a code rate R = 1/2, which is the rate that was selected for the Pioneer 9 channel coding system, equation (12) shows a capacity reduction factor $\gamma = 0.76$ (-1.2 dB) as the penalty paid for the resulting finite bandwidth. For a code rate R = 6/32, which is the rate of the Mariner '69 channel code, the reduction factor is only $\gamma = .889$ (-0.51 dB). One might suppose then that the Mariner '69 code rate was a wiser choice than the Pioneer 9 code rate. In fact, however, the Mariner '69 code was chosen for reasons, which will be explained below, that had nothing to do with minimizing the capacity loss due to finite bandwidth. It would, in fact, have been more desirable to use the code rate R = 1/2 in spite of the 0.7 dB increased capacity loss, since the symbol energy $E = R \times E_b$ at rate R = 6/32 is so much smaller than at R = 1/2 (4.2 dB smaller for the same E_b) that the phase-lock loops that are used to perform coherent demodulation of BPSK signals tend to lose lock unacceptably often when the lower code rate is used. One of the lessons of these first applications of channel coding in deep-space was that the use of an energy-efficient channel coding system places unusually severe demands on the phase-tracking loops in the demodulator because of the very low energy E of the transmitted waveforms.

3 A Tale of Two Codes

We will now take a closer look at the specific channel codes that were chosen for the Pioneer 9 and Mariner '69 downlink channels. Our discussion in the previous section has already told us two important considerations for these channel coding systems, viz.,

- The decoder must make use of soft-decisions on the transmitted digits.

- The code rate should be at most 1/2, but preferably as near 1/2 as possible.

3.1 The Mariner '69 Code

The first choice that had to be made by the designers of the Mariner '69 channel coding system was whether to use a block code or a convolutional code. Several factors militated against the choice of a convolutional code. At the time that the decision on the Mariner '69 code had to be made, the only general soft-decision decoding technique for convolutional codes that was known was Wozencraft's original sequential decoding algorithm, cf. [8]. The much more efficient and easily implemented Fano algorithm [9] for sequential decoding was still in the process of discovery and development. Wozencraft's algorithm had already been implemented in special purpose hardware at the M.I.T. Lincoln Laboratory, but for application on telephone channels (for which it turned out to be not well suited) rather than for the deep-space channel (for which it would have been well suited). Memory in a channel tends to cause large increases in the computation required to do sequential decoding; the deep-space channel is memoryless but the telephone channel is certainly not. In any case, there was no convincing evidence at the time when the decision on the Mariner '69 code had to be made that sequential decoding would be a good and practical choice. The Mariner '69 code designers at the CalTech Jet Propulsion Laboratory (JPL) in Pasadena, California, opted for a block code, a decision that we find difficult to fault even in retrospect.

Unfortunately, there was at that time only a few block codes for which a practical soft-decision decoding algorithm was known--a situation that has not changed appreciably in the intervening 30 years! Moreover, these codes generally had very low code rates, which as we have seen is fine for energy-efficiency on the deep-space channel but nonetheless undesirable because of the heavy demand that the expanded bandwidth places on the phase-lock loops required for coherent demodulation of BPSK signals. The most attractive block codes available for practical soft-decision decoding were the first-order Reed-Muller codes [6]. These first-order Reed-Muller codes are binary parity-check codes with blocklength $n = 2^m$, minimum Hamming distance $d_{min} = 2^{m-1}$ and $k = m + 1$ information bits, where $m \geq 2$ is a design parameter. The code rate, $R = k/n = (m + 1)/2^m$, is unpleasantly small except for small m. These codes can be viewed in many different ways but, when used with binary antipodal signalling, they are perhaps best seen as realizing the "biorthogonal signal set" in n dimensions, cf. [3, p. 261].

The n-dimensional "orthogonal signal set of energy E" is a set of n vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in n-dimensional Euclidean space with the property that each pair of vectors is orthogonal (i.e., the inner product between each pair of vectors vanishes) and each vector has squared norm E (i. e., its innner product with itself is E). The simplest construction of this signal set is to choose the n vectors with a single non-zero component equal to $\sqrt{E}$, from which it is easy to see that the squared Euclidean distance $(d_E)^2$ between any two signals is $2 \times E$, but many other constructions are possible. For instance, with n = 4, the rows of the Hadamard matrix

$$H_2 = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}, \tag{13}$$

when scaled by $\sqrt{E/4}$, yield the 4-dimensional orthogonal signal set of energy E. The *n-dimensional biorthogonal signal set of energy* E is the set of $2 \times n$ signals formed by augmenting the orthogonal signal set with the negative of each of its signals. Each signal is at squared Euclidean distance $(d_E)^2 = 2 \times E$ from all the other signals excepts its negative, from which it is at squared Euclidean distance $(d_E)^2 = 4 \times E$.

Consider now sending a signal from any set of equi-energy signals in n-dimensional Euclidean space over a channel such that the received vector **r** is the sum of the transmitted signal and an additive noise vector **n** whose components are independent Gaussian random variables, all with zero mean and the same variance. The maximum-likelihood detection rule (which is the rule that minimizes the error probability in choosing the transmitted signal when all signals are equally likely) is to choose that signal $\mathbf{x}_i$ whose inner product (or "correlation") with **r** is greatest, cf. [3, p. 234]. For the "Hadamard signal set" of the preceding paragraph, we see that $H_2\mathbf{r}$ is (within an unimportant positive scale factor) the vector of correlations between **r** and each of the signals in the signal set. Thus, the maximum-likelihood detection rule reduces to: Choose the signal $\mathbf{x}_i$ corresponding to the location i of the maximum component of $H_2\mathbf{r}$. Decoding the corresponding biorthogonal signal set is almost as simple--the maximum-likelihood detection rule becomes: Choose the signal to be $\mathbf{x}_i$ or $-\mathbf{x}_i$, where i is the location of the maximum-magnitude component of $H_2\mathbf{r}$, according as to whether this component is positive or negative, respectively.

The relationship of the above to the Mariner '69 channel code stems from the fact that the first-order Reed-Muller code of length $n = 2^m$, when the binary codewords are mapped into vectors in n-dimensional Euclidean space in the manner that a binary 0 is mapped into $+\sqrt{E}$ and a binary 1 is mapped into $-\sqrt{E}$, yields precisely the biorthogonal signal set corresponding to the $2^m \times 2^m$ Hadamard matrix H_m, where the Hadamard matrices are defined recursively by

$$H_m = \begin{bmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{bmatrix}, \tag{14}$$

in the same manner as described in our example with H_2. It follows that this code can be decoded with maximum-likelihood soft-decision decoding on the deep-space channel by first mapping the received waveform over the corresponding n modulation symbol intervals to the vector **r** of coefficients in the representation (of the "relevant" portion) of this waveform as a linear combination of the n orthonormal waveforms obtained by translation of the basis waveform for the one-dimensional modulation signal set, then using the optimum detection rule described above, which reduces essentially to the computation of H_m **r**. Since the entries of the matrix H_m are all either +1 or -1, the obvious calculation of H_m **r**. would take a total of n (n - 1) ≈ n^2 additions and subtractions. But this matrix multiplication is precisely the rule of calculation for the so-called *Walsh-Hadamard transform* of **r**, for which there exists a fast-transform method that requires only n log_2 n additions and subtractions. All this was well known in the early 1960's to the engineers at JPL, who built a special purpose digital device that they called the "Green machine" (not after its color but after its designer) to perform the fast Walsh-Hadamard transform on vectors **r** of length $n = 2^5$. This device then served as the decoder for the (n = 32, k = 6) Reed-Muller code used in the Mariner '69 spacecraft, cf. [11].

It remains to see why this specific Reed-Muller code was the one selected. The probability of error in deciding between two equi-energy vectors in additive Gaussian noise (whose components are independent Gaussian random variables with zero means and variances $N_o/2$) depends only on the squared Euclidean distance between these signals and is given by $Q(\sqrt{(d_E)^2/2N_o})$, [as can be inferred from our discussion preceding equation (7) above where $(d_E)^2 = 4E_b$]. If a binary code has minimum Hamming distance d_{min} and its binary codewords are mapped into vectors in n-dimensional Euclidean space in the manner that a binary 0 is mapped into $+\sqrt{E}$ and a binary 1 is mapped into $-\sqrt{E}$, then the minimum squared Euclidean distance between the resulting vectors is

$$(d_{Emin})^2 = 4\,E\,d_{min.} \tag{15}$$

Thus, the probability of a decoding error will be given by $P_e \approx Q(\sqrt{2\,E\,d_{min}/N_o})$, where we have neglected to account for the multiplicity of vectors that may be at this same minimum squared Euclidean distance from the transmitted codeword. But $E = RE_b$, where R is the code rate, so

$$P_e \approx Q(\sqrt{2\,R\,E_b\,d_{min}/N_o}\ ). \tag{16}$$

Comparing (7) and (16) shows that, at least to the first order, the coding gain G_c is given by

$$G_c \approx R\,d_{min}, \tag{17}$$

which is a very useful formula for evaluating binary codes used with soft-decision decoding and binary antipodal signalling on the deep-space channel. Using the parameters of the first-order Reed-Muller codes in (17) gives

$$G_c \approx \frac{m+1}{2}. \tag{18}$$

This shows that it is desirable to choose m as large as possible, i. e., up to the point where the phase-lock loop tracking problem induced by the bandwidth increase can still be overcome. The JPL engineers chose m = 5, for which (18) gives an estimated gain of $G_c \approx 3$ (4.8 dB). The actual gain is somewhat smaller because of the large multiplicity of nearest neighbors in the first-order Reed-Muller codes--there are 62 nearest neighbors for each codeword in the m = 5 code. The actual gain was 2.2 dB at a bit error probability of 5×10^{-4} [Posner gives the figure as 5×10^{-3} in [11] but this is apparently incorrect; the figures given in [11] for the uncoded Mariner IV system (an E/N_0 of 8.5 dB), which had the same bit error probability as the Mariner '9 system, yield a bit error probability of 5×10^{-4}, as can be checked from equation (7).]

3.2 The Pioneer 9 Code

The designers of the Pioneer 9 channel coding system had the advantage of starting work in the mid-1960's when the power and practicality of the Fano sequential decoding algorithm [9] were becoming well known. With little hesitation, these designers settled on a rate R = 1/2 convolutional coding system to be decoded by the Fano algorithm. Sequential decoding in general is a technique by which the decoder works (at least in principle) with the tree of partial encoded sequences that it has previously examined for their "likelihood" with respect to the received sequence, extending each time the most recent explored sequence until its "likelihood" falls below some threshold. Fano in 1964 [9] had made two important contributions to sequential decoding. First, he introduced a "metric" that on intuitive grounds should correspond to a reasonable notion of "likelihood" when applied to possible partial encoded sequences of different length. [It was not until many years later that we were able to prove [12] that this "Fano metric" was precisely the metric that yields the true maximum-likelihood decision as to which partially explored encoded sequence to extend. This delayed theory is illustrative of the theoretical intricacy of sequential decoding!] Second, he introduced his own extension algorithm, which is the soul of simplicity and, at the same time, the most subtle algorithm that this writer has ever seen. Fano's algorithm uses almost no memory, which was an important consideration in the mid-1960's, and remains today the fastest sequential decoding algorithm known.

The main problem with sequential decoding is the variability of the decoding computation. Generally (but not always) when sequential decoding is used, the output of the convolutional encoder is segmented into independent "frames" by occasionally inserting a pattern of M consecutive 0's into the information bit stream to drive the encoder back to the all-zero initial state. This yeilds a finite (but still very large) potential code tree to be explored for each frame. How long it takes the decoder to work its way to the end of this code tree depends on how "noisy" the actual received frame is.

The code that was used in the Pioneer 9 system was a rate R = 1/2 non-systematic convolutional code that had been constructed by S. Lin and H. Lyne, cf. [13, p. 539]. [The term "non-systematic" refers to the fact that the information bits do not appear unchanged among the encoded digits--it is really the encoder that should be called "non-systematic" rather than the code. In this paper we have followed the usual communications practice of not distinguishing between a "code" (i. e., the set of all codewords) and an "encoder," but this difference can be important in coding theory.] The Lin-Lyne code had a memory of M = 20, i. e., the encoder remembered twenty past information bits as well as the current information bit when forming the two encoded binary digits that are emitted for each input information bit. It was known that the minimum Hamming distance d_{min} of this code was 11. This is the minimum distance between two encoded sequences of length $2 \times (M+1) = 22$ digits (one "constraint length") that correspond to different values of the initial information bit. If one uses these code parameters mindlessly in (17), one computes an estimated coding gain of $G_c \approx 5$ (7.0 dB), but this estimate is so optimistic that it must be taken with a grain of salt! [In fact, one should really use in (17) the *free distance*, d_{free}, of the convolutional code, which is the minimum Hamming distance between two encoded sequences of infinite length that correspond to different values of the initial information bit. This would, of course, give an even more optimistic estimate of $G_{c.}$!] The problem is not one of great multiplicity of near neighbors as it was for the Reed-Muller code. Rather, the problem is that one must take into account the fact that although Fano-algorithm sequential decoding is virtually maximum-likelihood decoding *if the decoder is allowed to compute until it makes a decoding decision*, in practice *one always aborts the decoding after some predetermined amount of computation on a frame and announces erasure of the corresponding frame of data*. The actual error probability in the non-erased frames is virtually zero since the frames that would have resulted in decoding errors with maximum-likelihood decoding are generally frames that would require enormous computation to decode. [This latter feature was attractive to the scientists with experiments aboard Pioneer 9--they could really trust any experimental data radioed back from Pioneer 9 that was not erased by the sequential decoder.] The way that computation increases with decreasing signal-to-noise ratio is the primary determiner of the actual signal-to-noise ratio at which one can operate with sequential decoding, and hence of the actual coding gain. For the Pioneer 9 system, the actual gain was about 3.0 dB, cf. [13, p. 539].

3.3 Comparison of the Two Codes and Why the Pioneer 9 Code Was the First into Space

On all counts, the Pioneer 9 channel coding system was better than the Mariner '69 channel coding system. It offered greater coding gain (3.0 dB vs. 2.2 dB) and at the same time its higher rate (R = 1/2 vs. R = 6/32) meant considerably less bandwidth expansion and hence much better tracking by the phase-lock loops in the coherent demodulator. The Fano-algorithm sequential decoder for the Pioneer 9 system was essentially cost-free--it was realized in software during the spare computation time of an already-on-site computer, whereas the "Green machine" for decoding the Mariner '69 code was a non-negligible piece of electronic hardware.

The fact that the superior Pioneer 9 channel coding system got into space sooner than the inferior Mariner '69 channel coding system, even though the former was designed several years later than the latter (which is why it was a better system), is due primarily to the efforts of a single person, D. R. Lumb of the NASA Ames Research Center. Lumb had been quick to appreciate the importance of Fano-algorithm sequential decoding for the deep-space channel. In this, he was influenced by G. D. Forney, Jr., of Codex Corporation and by two Codex consultants, R. G. Gallager and (to a much lesser extent) this writer. After rapid development of the convolutional coding system, Lumb succeeded in getting it aboard Pioneer 9 as *an experiment*. This neatly side-stepped the long approval time that would have been necessary if this coding system had been specified as part of an operational communications system for a spacecraft. The operational communications system for Pioneer 9 was, of course, an uncoded BPSK system. The experimental coding system was activated as soon as Pioneer 9 was launched--it was never turned off!

References

[1] C. E. Shannon, "A Mathematical Theory of Communication", *Bell Sys. Tech. J.*, vol. 27, pp. 379-423 and 623-656, July and Oct. 1948.

[2] C. E. Shannon, "Communication in the Presence of Noise", *Proc. IRE,* vol. 37, pp. 10-21, Jan. 1949.

[3] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[4] R. M. Fano, *Transmission of Information*. Cambridge, Mass.: MIT Press and Wiley, 1961.

[5] J. L. Massey, "Coding and Modulation in Digital Communication", pp. E2(1)-E2(4) in *Proc. Int. Zurich Seminar*, Zürich, Switzerland, March 1974.

[6] I. M. Jacobs, "Sequential Decoding for Efficient Communication from Deep Space", *IEEE Trans. Commun. Tech.*, vol. COM-15, pp. 492-501, August 1967.

[7] H. Nyquist, "Certain Factors Affecting Telegraph Speed", *Bell Sys. Tech. J.*, vol. 3, pp. 324-

346, April 1924.

[8] J. M. Wozencraft and B. Reiffen, *Sequential Decoding.* Cambridge, Mass.: MIT Press and Wiley, 1961.

[9] R. M. Fano, "A Heuristic Discussion of Probabilistic Decoding", *IEEE Trans. Info. Th.,* vol. IT-9, pp. 64-73, April 1963.

[10] I. M. Reed, "A Class of Multiple-Error-Correcting Codes and the Decoding Scheme", *IRE Trans. Info. Th.*, pp. 38-49, Sept. 1954.

[11] E. C. Posner, "Combinatorial Structures in Planetary Reconnaissance", pp. 15-46 in *Error Correcting Codes* (Ed. H. B. Mann). New York: Wiley, 1968.

[12] J. L. Massey, "Variable-Length Codes and the Fano Metric", *IEEE Trans. Info. Th.,* vol. IT-18, pp. 196-198, Jan. 1972.

[13] S. Lin and D. Costello, Jr., *Error Control Coding: Fundamentals and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1983.