To appear in *Coding, Communications, and Broadcasting*,
Lancaster, UK: Research Studies Press, 2000.

Bit Error Probability and Encoder Structure

Seán Coffey and Houshou Chen

Abstract. The problem of decoding a binary linear code used over a binary symmetric channel so as to minimize bit error probability, rather than codeword error probability, is studied. In general, there are substantial differences with the maximum likelihood case. The optimal decoding rule depends on the crossover probability of the channel, the bit error probability depends on the generator matrix used, and the optimal generator matrix itself depends on the crossover probability of the channel.

This paper demonstrates each of the above properties. In addition, an infinite class of codes whose optimal generator matrix is non-systematic when the channel crossover probability is small is demonstrated. A new demonstration of the following facts is presented: if the code has $d^{\perp} > 2$, then for a sufficiently noisy channel, the optimal generator matrix is (uniquely) a (wide-sense) systematic one, and the optimal decoding rule is to ignore the parity checks and to take the information bits as they are received. The infinite class mentioned above has $d^{\perp} > 2$, and thus has different optimal generator matrices at different channel crossover probabilities.

1. Introduction

This paper examines the problem of decoding a binary linear code used over a binary symmetric channel when the goal is to minimize the information bit error probability.

This topic, of obvious fundamental interest, has received new impetus from the success of turbocodes. Although turbocoding systems work very well experimentally, there are many problems in understanding why this is, and how to choose turbocode parameters optimally. One of the essential features of the turbo decoding algorithm is the use of bit error probability decoding rather than maximum likelihood, and the relationship with encoder structure.

Seán (aka John T.) Coffey is with the EECS Department, University of Michigan, Ann Arbor, Michigan 48109. e-mail: scoffey@eecs.umich.edu

Houshou Chen is with the Department of Electrical Engineering, National Chi-Nan University, Nantou, Taiwan 545. e-mail: houshou@ncnu.edu.tw

This work was supported in part by the U.S. Army Research Office under Grant DAAH04-96-1-0377.

We begin with an intuitive sketch of why the information bit decoding problem is relevant to turbocodes. The classical approach to achieving performance close to capacity was to fix the target probability of error, and to form a family of codes of increasing decoding complexity whose performance points on the plot of P_e versus E_b/N_0 moved (horizontally leftwards) towards capacity. Thus we could, for example, take $P_e = 10^{\perp 5}$ and the class of rate $1/2$ convolutional codes. Then by increasing the constraint length, we get a set of performance points moving closer and closer to capacity. Unfortunately, beyond a certain point the decoding complexity becomes prohibitive in this approach.

Turbocodes instead can be thought of as approaching the eventual target point "vertically", i.e., by taking the E_b/N_0 to be fixed and, by repeated decoding, achieving a set of operating points that move downwards to the target error probability. Thus the early iterations for a turbocode operate in a region of very low signal to noise ratio, and very high error probability even for the decoded bits.

One issue arises immediately: the question of when the turbo decoding algorithm "fires up", i.e., the question of when the output decoded bits at the first iteration are even marginally more reliable than the received bits. For clearly, if this is not the case, we expect that the turbo decoding algorithm will yield the originally received bits as decoded word. The critical channel quality at which there is even marginal improvement in reliability, and the relationship with the generator matrix, are of considerable interest.

Here we concentrate on the more tractable case of a binary symmetric channel. In Section 2, we review some of the main properties of the problem. (For an account of the history of contributions to the problem, we refer the reader to [3] and the references contained therein.) In Section 3, we present new demonstrations of results for general classes of codes. We construct an infinite class of codes that have non-systematic optimal generator matrices when the channel is very quiet, and present a new and simplified demonstration of the result that, if the code has dual distance greater than 2, then for sufficiently noisy channels the optimal generator matrix to use is systematic, and the optimal decoding rule is to take the information bits as they are.

2. Basic properties

The information bit decoding problem differs substantially from the maximum likelihood decoding problem. Suppose, for example, that we are presented with a list of all codewords of the linear code C, plus a received word r, and are told that the channel is binary symmetric, with crossover probability

$0 < p < 1/2$. Do we have enough information to make an optimal estimate of the codeword?

If our goal is to minimize the codeword error probability, then the answer is Yes: we find a coset leader for the coset containing r and declare that to be the error pattern. In particular, we do not need to know either the generator matrix that has been used, or the crossover probability of the channel.

If, however, our goal is to minimize the information bit error probability, then the answer is No: the decoded codeword depends on, i.e., varies with, both the generator matrix used and the crossover probability of the channel.

In fact, minimum bit error probability has the following properties, each of which will be demonstrated later:

I. **The optimal decoding rule depends on p.**

II. **The optimal decoding rule depends on the generator matrix.**

III. **The bit error probability depends on the generator matrix.**

IV. **The optimal generator matrix depends on p.**

By "decoding rule" here, we mean the rule for determining which codeword (which in turn determines which set of k information bits) will be chosen by the decoder given a particular received word.

Consider for illustrative purposes the $[5,3,2]$ code with generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

The standard array is

00000	10011	**01010**	**00101**	11001	10110	**01111**	11100
00001	10010	**01011**	**00100**	11000	10111	**01110**	11101
00010	10001	**01000**	**00111**	11011	10100	**01101**	11110
10000	00011	**11010**	**10101**	01001	00110	**11111**	01100

.

Consider the problem of decoding the first information bit. We divide the code into codewords for which this bit is 0, i.e., $\{00000, 01010, 00101, 01111\}$, and codewords for which this bit is 1. We divide each of the cosets correspondingly, as in the standard array above.

Suppose we receive the word 10000; then the error pattern belongs to the last coset. If the error pattern is one of $\{10000, 11010, 10101, 11111\}$, i.e., one of those listed in bold face, the transmitted codeword has first information bit 0; if not, i.e., if the error pattern is one of $\{00011, 01001, 00110, 01100\}$, the first information bit is 1. To decide which is more probable, we compare the total probability of each of these sets of possible error patterns, choosing

the one that is more likely. In this case, the sets have probabilities $p(1-p)^4+2p^3(1-p)^2+p^5$ (for the set of words in bold face) and $4p^2(1-p)^3$ (for the words in regular font). The first set is more probable if $p \leq 0.228$, while the second is more probable if $p > 0.228$. This demonstrates property I above: *the optimal decoding rule depends on the value of p.* Note that for maximum likelihood decoding, one word in the coset rather than one set is chosen, and the word of weight 1 is optimum for all p; for minimum bit error probability decoding, however, the four words of weight 2 collectively outweigh the single word of weight 1 in probability when the channel is noisy enough.

In general, given a received word r, we estimate the jth information bit u_j as the bit that maximizes

$$\max_{u_j=0,1} \sum_{\substack{u\in\mathcal{U} \\ u_j}} p^{|r+uG|}(1-p)^{n\perp|r+uG|}, \tag{1}$$

choosing arbitrarily if the two are equal. The probability of receiving a word in a given coset and also making a decoding error for the jth information bit is the sum of the probabilities of the less probable set in that coset. (Note that the partition, and hence the probability, depends on the information bit being decoded. Also the estimate of the information bit depends on the received word r; however, the probability that this estimate is incorrect depends only on the coset.)

The overall probability of the jth information bit being in error is then obtained by summing over all cosets. Thus

$$p_{\text{inf}}^{(j,G)} = \sum_{q\in Q} \min \left(\sum_{\substack{u\in\mathcal{U} \\ u_j=1}} p^{|q+uG|}(1-p)^{n\perp|q+uG|}, \sum_{\substack{u\in\mathcal{U} \\ u_j=0}} p^{|q+uG|}(1-p)^{n\perp|q+uG|} \right), \tag{2}$$

where Q is a set of coset representatives, and $\mathcal{U}$ is the set of possible information sequences.

For this code, it may be verified that for each of the other cosets, and for each of the other information bits, the optimal information bit decision is to use a coset leader as estimate of the error pattern. The average information bit error probability, $\overline{p}_{\text{inf}}^{(G)} = \left(p_{\text{inf}}^{(1)} + p_{\text{inf}}^{(2)} + p_{\text{inf}}^{(3)}\right)/3$, is then

$$\overline{p}_{\text{inf}}^{(G)} = \begin{cases} \frac{1}{3}\left(2p+8p^2-20p^3+20p^4-8p^5\right) & \text{for } p \leq 0.228 \\ p & \text{for } p > 0.228. \end{cases}$$

2.1 Generator matrices

It will emerge that it is important in this problem to define the term "systematic" carefully. Classically this means that the generator matrix is of the form $G = (I_k|P)$, i.e., that all the columns of the $k \times k$ identity matrix are present in the leftmost k positions of the generator matrix. We will refer to this as "narrow-sense systematic." A generator matrix will be said to be "wide-sense systematic" if the columns of the $k \times k$ identity matrix are each present somewhere in the generator matrix. This is sometimes referred to as a "separable" generator matrix.

Consider the non-systematic (in both senses) generator matrix for the same code

$$G_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix},$$

obtained by adding row 1 of G to rows 2 and 3. If we are interested in decoding the first information bit, we again compute the subcode for which this first information bit is zero, i.e., $\{00000, 11001, 10110, 01111\}$ and highlight the corresponding columns in the standard array:

00000	10011	01010	00101	**11001**	**10110**	**01111**	11100
00001	10010	01011	00100	**11000**	**10111**	**01110**	11101
00010	10001	01000	00111	**11011**	**10100**	**01101**	11110
10000	00011	11010	10101	**01001**	**00110**	**11111**	01100

Now if the word $r = 10000$ is received, we compare the probability of the highlighted subset within the corresponding coset, i.e., $\{10000, 01001, 00110, 11111\}$ to its complement, i.e., $\{00011, 11010, 10101, 01100\}$. In this case we have $p(1-p)^4 + 2p^2(1-p)^3 + p^5 > 2p^2(1-p)^3 + 2p^3(1-p)^2$ for all $p < 1/2$, so for all p we declare that the error pattern belongs to the highlighted set, and choose the first information bit as 0 accordingly. Thus, in particular, the received word $r = 10000$ is never decoded to the codeword 10011; on the other hand the results of Section 3.2 imply that for p sufficiently close to $1/2$ and a systematic generator matrix, $r = 10000$ must be decoded to 10011. This demonstrates property II: *the optimal decoding rule depends on the generator matrix.*

Repeating for the other cosets, and for the second and third information bits, we find that the average probability of information bit error with this

generator matrix is

$$\bar{p}_{\text{inf}}^{(G_1)} = \frac{1}{3}\left(4p - 2p^2\right).$$

Note that this is greater than p for $0 < p < 1/2$. Thus this generator matrix is substantially worse than the systematic one used earlier. More generally, we deduce property III above: *the bit error probability depends on the generator matrix used.*

3. General results

3.1 The quiet region

It is not always the case that a systematic generator matrix is best. In this section, we construct an infinite class of codes that have an optimal generator matrix that is non-narrow-sense-systematic when p is sufficiently small. From this we indicate how to construct an infinite class of codes that have optimal generator matrices that are not wide-sense-systematic when p is sufficiently small.

We will use a very useful result of Dunning [2]: given a linear code and a channel crossover probability, consider all possible generator matrices and the resulting bit error probabilities under optimal decoding. Arrange each set of k bit error probabilities in non-decreasing order $p_1^G \leq p_2^G \leq \cdots \leq p_k^G$. Then there exists a generator matrix with bit error probabilities achieving the minimum in each position of this ordered list simultaneously, i.e., $p_i^{G^*} \leq p_i^G$ for all i and G.

Consider codes with generator matrices of the following form:

$$G = \left[\begin{array}{c} 111111000\cdots00 \\ \hline G_1 \end{array}\right],$$

where G_1 generates a code with sufficiently large parameters, say $n \geq 1000$, $k \geq 100$, $d \geq 100$. The minimum distance of the code generated by G is 6, and thus the average bit error probability, even under optimal decoding, must be of the form $\bar{p}_{inf} = k \cdot p^3 + O(p^4)$ with $k > 0$. For sufficiently small p, we need to minimize the leading constant k. With any given generator matrix, the codeword of weight 6 is the sum of some subset of the rows of the generator matrix. The ith bit error probability under optimal decoding is of the form $p_i = k_i \cdot p^3 + O(p^4)$ with $k_i > 0$ if and only if the ith row of the generator matrix is involved in the sum that gives the codeword of weight 6. Otherwise the smallest weight codeword that gives an error in the ith information bit has weight at least 94, and so $p_i = O(p^{47})$ even with optimal decoding, i.e., far lower when p is very small.

It is possible to choose the codeword of weight 6 as a row of the generator matrix, as in G above. Then all information bits except one have error probability $O(p^{47})$. By Dunning's result, the optimal generator matrix will therefore have these probabilities, and hence the codeword of weight 6 must involve only one row of the generator matrix. Thus we must have a generator matrix of the form of G above when p is very small. This is non-narrow-sense-systematic.

We now sketch a construction of a code with an optimal generator matrix that is not systematic in the wide sense as $p \to 0$. We select G_1 to be some large-parameter $[n_1, k_1, d_1]$ code as above, but replace each 1 in the generator matrix by a randomly chosen binary string of length 6 and weight 3, and replace each 0 by a string of six zeros, to obtain a new generator matrix G_1' that generates a $[6n_1, k_1, \geq 3d_1]$ code. We form G by adjoining n_1 non-overlapping words of weight 6 to G_1': these words collectively cover every position.

Now any codeword involving a row from G_1' must have weight at least d_1, no matter how many of the rows of weight 6 are involved in the sum: this is the reason for the modification of G_1 to G_1'. Once again we can argue that all the weight 6 codewords must be included in the optimal generator matrix for all sufficiently small p. But then the additional k_1 rows cannot be chosen in a way that ensures that all the columns of the $k \times k$ identity matrix are present, i.e., the generator matrix cannot be wide-sense systematic.

3.2 The noisy region

We present a simplified demonstration of the following results, which were first found in a more general form by Kiely, Coffey, and Bell [3]. Delsarte [1] had previously noted, via direct computation of half-coset weight enumerators, that the 'ignore parity checks' decoding rule is optimum for many codes for sufficiently high p when the generator matrix is systematic.

THEOREM: *For any binary linear code with $d^\perp > 2$ transmitted over the binary symmetric channel, the optimal generator matrix to use for p sufficiently close to $1/2$ is wide-sense systematic, and the optimal decoding rule is to ignore the parity check bits and to accept the information bits as they are received.*

Here the "information" bits are those in the positions in which the columns of the identity matrix appear, and the "parity check" bits are the others.

(There is an unfortunate misstatement of the result in the abstract of Kiely *et al.* [3], in which it is stated that a wide-sense systematic generator matrix

is optimal for binary symmetric channels for codes with $d^{\perp} > 2$, without the qualification that this is necessarily true only for sufficiently noisy channels. The result is stated correctly in the conclusions of the same paper.)

Proof: Writing $p = 1/2 - \varepsilon$ and substituting into (2), we find that

$$\begin{aligned} p_{\text{inf}}^{(j,G)} &= \frac{1}{2^n} \sum_{q \in \mathcal{Q}} \min \left(\sum_{u \in \mathcal{U}: u_j = 1} (1-2\varepsilon)^{|q+uG|} (1+2\varepsilon)^{n-|q+uG|}, \right. \\ &\qquad \left. \sum_{u \in \mathcal{U}: u_j = 0} (1-2\varepsilon)^{|q+uG|} (1+2\varepsilon)^{n-|q+uG|} \right) \\ &= \frac{1}{2^n} \sum_{q \in \mathcal{Q}} \min \left(\sum_{u \in \mathcal{U}: u_j = 1} \left(1 + 2\left(n - 2|q+uG|\right)\varepsilon + O\left(\varepsilon^2\right)\right), \right. \\ &\qquad \left. \sum_{u \in \mathcal{U}: u_j = 0} \left(1 + 2\left(n - 2|q+uG|\right)\varepsilon + O\left(\varepsilon^2\right)\right) \right) \\ &= \frac{1}{2} + n\varepsilon - \frac{4}{2^n} \sum_{q \in \mathcal{Q}} \max \left(\sum_{\substack{u \in \mathcal{U} \\ u_j = 1}} |q+uG|, \sum_{\substack{u \in \mathcal{U} \\ u_j = 0}} |q+uG| \right) \varepsilon + O\left(\varepsilon^2\right) \end{aligned}$$

since for all sufficiently small ε, the minimum is achieved by minimizing the first order term.

The optimal estimate of u_j for small ε is then the complement of the choice that maximizes the sum of the weights in the half cosets given by the value of u_j, if there is not a tie, i.e., the optimal estimate of u_j barring a tie is the complement of

$$\arg \max_{u_j = 0,1} \sum_{\substack{u \in \mathcal{U} \\ u_j}} |r + uG|. \tag{3}$$

The generator matrix G must at least achieve the maximum

$$\text{argmax}_G \sum_{j=1}^{k} \max_{u_j = 0,1} \sum_{q \in \mathcal{Q}} \sum_{\substack{u \in \mathcal{U} \\ u_j}} |q + uG|. \tag{4}$$

We fix q and to each of the terms $\sum_{u \in \mathcal{U}: u_j = 0} |q + uG|$ and $\sum_{u \in \mathcal{U}: u_j = 1} |q + uG|$ we add the sum

$$\sum_{\substack{u \in \mathcal{U} \\ u_j = 0}} |uG|. \tag{5}$$

Now the first sum is

$$\sum_{\substack{u\in\mathcal{U}\\ u_j=1}} |q+uG| + \sum_{\substack{u\in\mathcal{U}\\ u_j=0}} |uG| = \sum_{u\in\mathcal{U}} |uG'| \tag{6}$$

where G' is the generator matrix obtained by changing the jth row of G from g_j to g_j+q. Similarly the second sum is

$$\sum_{\substack{u\in\mathcal{U}\\ u_j=0}} |q+uG| + \sum_{\substack{u\in\mathcal{U}\\ u_j=0}} |uG| = \sum_{u\in\mathcal{U}} |uG''| \tag{7}$$

where G'' is the generator matrix obtained by changing the jth row of G from g_j to q.

It is well known [4, p. 156] that every position in a binary linear code either assumes the values 0 and 1 equally often, or is always 0. The sum of the weights of the codewords of an $[n,k]$ binary linear code is then at most $n\cdot 2^{k\perp 1}$, with equality if and only if none of the columns of the generator matrix is 0.

We note that the sum in Eq. (5) is the sum of codeword weights for an $[n,k-1]$ code, with generator matrix obtained by deleting the jth row of G. G has no columns consisting of all zeros (since $d^\perp > 1$) and at most one column consisting of a 1 in the jth row and zeros elsewhere (since $d^\perp > 2$). Thus the generator matrix obtained by deleting the jth row of G has at most one zero column; it has exactly one if and only if G contains the jth column of a $k\times k$ identity matrix. We have

$$\sum_{\substack{u\in\mathcal{U}\\ u_j=0}} |uG| = \begin{cases} n\cdot 2^{k\perp 2} & \text{if } G \text{ does not contain } e_j \\ (n-1)\cdot 2^{k\perp 2} & \text{if } G \text{ does contain } e_j. \end{cases} \tag{8}$$

This then shows that

$$\max\left(\sum_{\substack{u\in\mathcal{U}\\ u_j=1}} |q+uG|, \sum_{\substack{u\in\mathcal{U}\\ u_j=0}} |q+uG|\right) \le n\cdot 2^{k\perp 1} - (n-1)2^{k\perp 2}$$

if the code has $d^\perp > 2$, and substituting into the expansion obtained for $p_{\text{inf}}^{(j,G)}$ earlier we find

$$p_{\text{inf}}^{(j,G)} \ge \frac{1}{2} - \varepsilon + O\left(\varepsilon^2\right)$$

and we achieve this if and only if the column e_j is contained in G.

From this, the error probability in the noisy region differs from p by a factor that is at most $O(\varepsilon^2)$; the remaining question is the crucial one of whether the information bit error probability is greater than or less than p.

We take the received word r as coset representative. The optimal choice of u_j is the complement of the one that maximizes the sum in Eq. (3). If e_j is a column in G, this maximum contains a 1 in jth position in every word of the sum, i.e., is the sum in which $\bar{u}_j + r_j = 1$. Thus the optimal choice of u_j is just r_j, i.e., we take the jth information bit as it is. Then it is trivial that the error probability for this jth bit is exactly p in the very noisy region, i.e.,

$$p_{\text{inf}}^{(j,G)} = \frac{1}{2} - \varepsilon.$$

The same reasoning holds for each of the k information bits, and so overall we have, if the code has dual distance greater than 2,

$$\bar{p}_{\text{inf}}^{(G)} \geq p$$

in some region $p_{crit} \leq p < 1/2$, with equality if and only if the generator matrix is wide-sense systematic, i.e., contains all the columns of the $k \times k$ identity matrix.

3.3 Property IV

The general construction in Section 3.1 produces codes with $d^{\perp} > 2$ with high probability. Combining the results of Sections 3.1 and 3.2, we find that for such codes the optimal generator matrix cannot be wide-sense systematic when p is sufficiently small, and must be wide-sense systematic when p is sufficiently large. Thus we have demonstrated property IV: *the optimal generator matrix depends on p*.

REFERENCES

[1] P. Delsarte, "Partial-optimal piecewise decoding of linear codes," *IEEE Trans. on Inform. Theory*, vol. IT-24, no. 1, pp. 70–75, Jan. 1978.

[2] L. A. Dunning, "Encoding and decoding for the minimization of message symbol error rates in linear block codes," *IEEE Trans. on Inform. Theory*, vol. IT-33, no. 1, pp. 91–104, Jan. 1987.

[3] A. B. Kiely, J. T. Coffey, and M. R. Bell, "Optimal information bit decoding of linear block codes," *IEEE Trans. on Inform. Theory*, vol. 41, no. 1, pp. 130–140, Jan. 1995.

[4] R. J. McEliece, *The Theory of Information and Coding.* Reading, Mass.: Addison-Wesley, 1977.