Binary Balanced Codes Approaching Capacity

Ebenezer Esenogho, Elie N. Mambou Theo G. Swart, and Hendrick C. Ferreira
Center for Telecommunication, Dept. of Electrical and Electronic Engineering Science, University of Johannesburg
P. O. Box 524, Auckland Park, 2006, South Africa
Email: {ebenezere, emambou, tgswart, hcferreira}@uj.ac.za

Abstract — In this paper, the construction of binary balanced codes is revisited. Binary balanced codes refer to sets of bipolar codewords where the number of “1”s in each codeword equals that of “0”s. The first algorithm for balancing codes was proposed by Knuth in 1986; however, its redundancy is almost two times larger than that of the full set of balanced codewords. We will present an efficient and simple construction with a redundancy approaching the minimal achievable one

Index Terms — Balanced codes, redundancy, binary alphabet, parallel decoding scheme, Knuth’s scheme

I. INTRODUCTION

Balanced codes have been widely studied due to its applicability in the field of communication and storage structures such as optical and magnetic recording devices like Blu-Ray, DVD and CD [1]; error correction and detection [2], [3]; cable transmission [4] and noise attenuation in VLSI systems. The decoding of balanced codes is fast, and it is done in parallel which avoids latency in communication.

All works on balancing codes have evolved from the celebrated Knuth’s algorithm [5], which stipulates that any codeword can be balanced by inverting the bits up to a certain point referred to as the balancing point. The value of that point is encoded as the prefix and appended as parity bits to the encoded information word. Two schemes have been suggested by Knuth to find these parity bits. These parity bits representing the prefix, should be determined in such a way that the overall codeword is balanced. For the parallel scheme, the index of the balancing point is encoded as the prefix whereas in the serial or sequential scheme, parity bits indicate the information word’s weight.

In [6], an improved design of Knuth’s parallel scheme was proposed; this consists of splitting parity bits into sets, where each of them represents a portion of inverted bits. In [7], the information word is split into two categories that are almost balanced codewords and those that are not. The inverting process is applied on the almost balanced codewords while for the others, a variable length code called tail is appended to achieve overall balancing.

In [8], a scheme was described to improve the spectral performance for balanced codes; this consists of generating some alternative balanced codewords representing each information word, this technique is called multimode code, and the selected word is that with the minimum squared weight, this improved the spectral performance and reduced low components by up to 3dB.

In [9], two schemes were described to improve the redundancy of Knuth’s algorithm. The first one used the distribution of the prefix index; knowing that the balancing point may be not unique given an information word, it has been proven that this distribution for equiprobable words is not uniform and presents a redundancy slightly less than that of Knuth’s scheme. In the second proposition, the multiplicity of balancing points is used to transmit auxiliary data. This last scheme was re-explored in [10] and renamed as bit recycling for Knuth’s algorithm (BRKA); the multiplicity of balancing points was exploited efficiency to lower the redundancy of Knuth’s scheme.

Immink and Weber presented a very efficient encoding for both variable and fixed length prefixes in [11]. This scheme consisted of associating every word to a balanced codeword according to the Knuth’s algorithm, then ranked all words associated to a balanced codeword. The rank of each word was encoded as prefix. A variation of this method was proposed in [12], following the same rule with the slight difference that only non-balanced words were associated to balanced codewords; this resulted in a small gain in redundancy.

In this paper, we present an efficient algorithm for associating words to a balanced codeword which results in a prefix length achieving the minimal redundancy for large information word length.

The rest of this paper is organized as follows: some work background is discussed in Section II. The new prefix coding construction for balancing words is presented in Section III. Section IV outlined some performance analysis and discussions of the proposed scheme. Finally, the paper is concluded in Section V.

II. BACKGROUND

Let \( x = (x_1, ..., x_k) \) be a bipolar sequence of length \( k \) and \( p = (p_1, p_2, ..., p_r) \), the prefix of length \( r \) to be appended to \( x. c = (c_1, c_2, ..., c_r) \), of length \( n = k + r \) is the transmitted codeword comprised of the encoding of \( x \) denoted as \( x' \) appended with \( pc \). All these words are defined within the alphabet \( A^2 \) where \( A^2 = \{-1, 1\} \). Let \( d(x) \) refer to the sum of all digits in \( x \), also called the disparity of \( x \). The word \( x \) is said to be balanced if \( d(x) = \sum_{i=1}^{k} x_i = 0 \).
Similarly, the disparity of the first $j$ bits of $x$, also called running digital sum (RDS), is denoted as $d_j(x)$; and $d_j(x) = \sum_{i=1}^{k} x_i$ where $1 \leq j \leq k$.

For the scope of this paper, the information word length is considered as even.

A. Knuth’s Balancing Scheme

The celebrated Knuth’s scheme consists of complementing a word bit up to certain point. This is equivalent of splitting a word into two segments, the first one has its bits flipped and the second is unchanged. It was shown in [5] that this simple and efficient procedure will always generate at least one balanced codeword. If $e$ is the index of the first balancing point then, the disparity of $x$ is given by:

$$d(x) = -\sum_{i=1}^{e} x_i + \sum_{i=e+1}^{k} x_i$$

(1)

Those summations reflect the two segments that build a balanced codeword. Because $d_j(x) = d_j(x) \pm 2$, it is always achievable to find an index $e$ corresponding to a balancing point such that $d(x) = 0$. Because this index might be unique given a word, by convention, the Knuth’s algorithm only considers the first one while inverting from least index bits.

In parallel scheme, the index $e$ is encoded as the prefix and appended to $x'$; and the length of that prefix, $r$, is given by:

$$r = \log_2 k, \text{ for } m \ll 1$$

(2)

The redundancy of a full set of balanced codewords of length $k$, denoted as $H_o(k)$, equals

$$H_o(k) = k - \log_2 \left(\frac{k}{2}\right)$$

(3)

An approximation of $H_0(k)$ was given in [5] as

$$H_o(k) \approx \frac{1}{2}\log_2 k + 0.326, \text{ for } m \ll 1$$

(4)

For large $k$, the Knuth’s scheme redundancy is almost twice larger than $H_o(k)$.

B. Efficient Binary Balanced Codewords

Let $x^e$ be the word $xx$ where first/bits are inverted. If $e$ represents the index of the first balancing point, then $x' = x^e$ is the codeword balanced through Knuth’s scheme. There are $k$ different ways of inverting the word $x$. In [11], it was established that some words from the set $x^1, x^2, ... x^k$ can be associated to the balanced word $x'$ following the inverting up to the first balancing point’s index as per Knuth’s scheme.

Let $s(x')$ be the set of all words associated to a balanced codeword, $x'$, $s(x') = x^1, x^2, ... x^k$ with $1 \leq j \leq k$; and $|s(x')|$, its cardinality.

Example 1 For $k=6$, $s(001110) = \{101110, 110000, 110001, 111110\}$, and $|s(001110)| = 4$; Similarly, $s(101010) = \{001010, 010101\}$, and $|s(001110)| = 2$.

The prefix of the encoded word corresponds to the information word rank within the subset $s(x')$. It was shown in [11] that the size of $s(x')$ is such that: $2 \leq |s(x')| \leq \frac{k}{2} + 1$ where $|s(x')| = \max\{d_j(x)\} - \min\{d_j(x)\}$ with $\max\{d_j(x)\}$ and $\min\{d_j(x)\}$ being the maximum and minimum RDS values of $x$ respectively.

For the fixed length scheme, the prefix has exactly $\log_2(k/2 + 1)$ digits while in variable scheme, the prefix length varies between 1 and $\log_2 k$.

In [12], based on the observation that every balanced code word is always associated to a balanced word, a modification of [11] was proposed for packet transmission systems where $1 \leq |s(x')| \leq k/2$.

Example 2 For $k = 6$, considering the same subsets as in Example 1, $s(001110) = \{101110, 110000, 111110\}$, and $|s(001110)| = 3$. Similarly, $s(101010) = \{001010\}$, and $|s(101010)| = 1$.

At this point, we are in possession of all tools necessary to introduce our proposed scheme.

III. PROPOSED PREFIX CODING SCHEME FOR BINARY WORDS

A. Encoding Scheme

The logic behind this scheme consists of associating all words to a balanced codeword, $x'$, in an efficient way such that the size of each subset $s(x')$ will be less or equal to a certain threshold $\lambda$, $|s(x')| \leq \lambda$.

The threshold value, $\lambda$ is set to be the average value of $|s(x')|$. This can be estimated as follow:

Let $k$ be the full balanced set of binary words of length $k$ of cardinality $|B_k|$. The value of $|B_k|$ was derived in [5] as

$$|B_k| = \left(\frac{k}{2}\right)^k = \frac{k!}{2^{k} (k-\frac{1}{2})^k}$$

(5)

$$= \frac{k!}{2^k}$$

Using Stirling approximation,

$$x! \approx \sqrt{2\pi x} - \left(x^e\right)^x$$

(6)

Its follows that,

$$|B_k| \approx \frac{\sqrt{2\pi k} \left(k\right)^k}{\left(\sqrt{2\pi k} - \left(\frac{k}{2}\right)^2\right)^{\frac{k}{2}}}$$

(7)

$$\approx 2^k \sqrt{\frac{2}{\pi k}}$$

©2019 Journal of Communications
Let $C_k$ be the full set of binary words of length $k$ and $|C_k|$ its cardinality. $|C_k| = 2^k$.

Therefore, the average value of $|s(x')|$ equals:

$$\lambda = \frac{|C_k|}{p_k} = \frac{2^k}{2k \sqrt{\frac{k}{\pi}}}$$

$$\approx \frac{\sqrt{nk}}{2} \quad (8)$$

The ceiling of the value will be considered for the rest of the paper.

We propose an algorithm called **association based on word subset relocation (AWSR)** for associating words to balanced codeword subset such that $|s(x')| \leq \lambda$.

The AWSR is a simple and efficient algorithm consisting of listing words of length $k$ and associated to a balanced codeword $x'$ following the lexicographic order. Algorithm 1 presents the AWSR procedure.

**Algorithm 1: Algorithm for AWSR**

<table>
<thead>
<tr>
<th>Input</th>
<th>: $k$, word $x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>: all words associated to $x'$</td>
</tr>
<tr>
<td></td>
<td>initialisation:</td>
</tr>
<tr>
<td>$c = 2^k$</td>
<td>/* find the cardinality of $C_k$ */</td>
</tr>
<tr>
<td>$b = (y_j)$</td>
<td>/* find the cardinality of $B_k$ */</td>
</tr>
<tr>
<td>$\lambda = \frac{c}{b}$</td>
<td>/* calculate $\lambda$ */</td>
</tr>
<tr>
<td>$t = \text{rank}(x)$</td>
<td>/* find the rank of $x$ in the set $C_k$, $1 \leq t \leq c$ */</td>
</tr>
<tr>
<td>$v = \frac{1}{2^{t-1}}$</td>
<td>/* find the rank of the corresponding balanced codeword $x'$ in the set $B_k$, $1 \leq v \leq b$ */</td>
</tr>
<tr>
<td>$p = \log_2 \lambda$</td>
<td>/* loop through all words $x_j$ associated to $x'$ */</td>
</tr>
<tr>
<td>$s(\lambda v - 1) + 1 \leq j &lt; \lambda \times v$</td>
<td>/* print $x_j$ */</td>
</tr>
<tr>
<td>$s(\lambda v)$</td>
<td>end</td>
</tr>
</tbody>
</table>

**Theorem 1** It is always possible to associate all binary words of length $k$ to subsets $(x')$ and $|s(x')| \leq \lambda$.

**Proof:** The following assumption is always verified $\lambda \times (k \frac{k}{2}) \leq 2^k$. This means that all $2^k$ binary words can be associated to at most $(k \frac{k}{2})$ balanced codeword subsets; and each subset has a size of at most $\lambda$.

**Example 3** Considering all binary words of length 4; the cardinality of this set is $2^4 = 64$ and there are $\binom{6}{3} = 20$ balanced words amongst of the 64.

The process below presents the associating of all information words of length 6 to balanced words subset following the AWSR algorithm as described above. $p$ are appended prefixes listed in the lexicographic order. It takes a prefix length of $\left\lceil \log_2 \left( \frac{6}{2} \right) + 1 \right\rceil$ to uniquely rank all elements associated a subset.

<table>
<thead>
<tr>
<th>$x'$</th>
<th>000111</th>
<th>000111</th>
<th>001101</th>
<th>001110</th>
<th>010011</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s(x')$</td>
<td>000000</td>
<td>001000</td>
<td>001000</td>
<td>001000</td>
<td>010000</td>
<td>00</td>
</tr>
<tr>
<td>000001</td>
<td>000101</td>
<td>001001</td>
<td>001101</td>
<td>010001</td>
<td>01</td>
<td></td>
</tr>
<tr>
<td>000110</td>
<td>001010</td>
<td>001110</td>
<td>001110</td>
<td>010010</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>000011</td>
<td>000111</td>
<td>001011</td>
<td>001111</td>
<td>010111</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

The AWSR scheme may not lead to all balanced codewords with associated words. Therefore, the encoding process only considers balanced codewords having associated words.

**Example 4 (Encoding)** Consider balancing the word of length 6, $x = 010011$.

Following the AWSR algorithm, $\lambda = 4$ and the prefix length equals 2. The rank of $x$ is $t = 20$. The rank of the corresponding $x'$ is $v = 20/4 = 5$. That is $x' = 010011$ words with ranks between $y_1 = \lambda(v - 1) + 1 = 17$ and $y_2 = 20$ are associated to $x'$; $s(010011) = \{010000, 010011, 010010, 010100\}$. The information word is the fourth one within $s(010011)$ corresponding to the prefix $p = 10$.

Therefore, $x$ is encoded as $c = (p|x') = 10010011$. The transmitted codeword $c$ of length $n = 6$ is balanced.

**B. Decoding Scheme**

Let the received codeword be $c$ of length $n$. The following steps are followed to retrieve the information word from the $c$.

- Compute $\lambda$ through (9) and set the prefix length to $r = \log_2 \lambda$.
- Extract $x'$ as the $n - r$ last digits of $c$. Then find the rank $v$ of $x'$ within $B_k$ where $1 \leq v \leq |B_k|$.
- Find the subset $s(x')$ with associated words to $x'$ of ranks between $t_1 = \lambda(v - 1) + 1$ and $t_2 = 2\lambda \times v$, where $1 \leq t_1, t_2 \leq |B_k|$.
- Extract the prefix as the first $r$ digits of $c$ then use to recover the information word $x$ from the subset $s(x')$.

**Example 5 (Decoding)** Consider recovering the information word from the received codeword $c = 01101100$ of length $n = 6$. $\lambda = 4$, then $r = 2$. The codeword $x' = 101100$ with rank $v = 16$. Words associated to $x'$ are ranked between $t_1 = 61$ and $t_2 = 64$, then $s(101100) = \{111110, 111101, 111110, 111111\}$. The prefix $p = 01$ corresponds to the second word in $s(x')$. 

©2019 Journal of Communications
Therefore, the information word is recovered as \( x = 111101 \).

IV. ANALYSIS AND DISCUSSION OF THE PROPOSED SCHEME

A. Redundancy

The values of \( |B^k| \) correspond to the central binomial coefficients of \( (1 + \alpha)^k \), where \( \alpha \) is an unknown variable. Some of these values can be obtained online from [13].

The minimum redundancy for the full set of balanced codewords was given in (4).

Let \( N(\tau, k) \) be the number of balanced codewords of length \( k \) such that \( s(x') = \tau \). The average number of bits for the construction in [11] is as follow:

\[
H_1(k) = 2^{-k} \sum_{\tau=2}^{2^k - 1} \tau N(\tau, k) \log_2 \tau
\]

The fixed prefix length scheme from [11] has a redundancy of

\[
\tau = \log_2 \left( \frac{k}{2} + 1 \right)
\]

The scheme proposed in [12] has the following average prefix length

\[
H_2(k) = \sum_{c=1}^{\frac{k}{2}} \tau N(\tau, k) \log_2 \tau / 2^k - \left( \frac{k}{2} \right)
\]

The average number of bits for the method in [10] is given by

\[
H_3(k) = \sum_{c=1}^{\frac{k}{2}} P(c) A(c)
\]

where \( P(c) = 2^{c+1-k} \left( \frac{k-1-c}{k} \right)^c \), \( 1 \leq c \leq k/2 \), \( d = c - 2 \log_2 c \), and \( AV(c) \)

\[
= (c - 2d) \log_2 c \cdot \frac{1}{2 \log_2 c} + 2d \cdot \frac{1}{2 \log_2 c} \cdot \log_2 c
\]

In variable length scheme, the prefix should be balanced; let \( \Delta \tau \) corresponds to the smallest value of length \( k \) such that proposed method is a fixed length scheme that takes exactly \( k/2 \geq \tau \) to uniquely encode every binary word of length \( k \).

This imposes a slight modification on (12) that

\[
H_1'(k) = 2^{-k} \sum_{\tau=1}^{\frac{k}{2}+1} \tau N(\tau, k) \Delta \tau / 2^k - \left( \frac{k}{2} \right)
\]

Similarly, \( H_2'(k) \) is derived from (10) as fellows

\[
H_2'(k) = 2^{-k} \sum_{\tau=1}^{\frac{k}{2}+1} \tau N(\tau, k) \Delta \tau
\]

The proposed method is a fixed length scheme that takes exactly \( \log_2 \lambda \) to uniquely encode every binary word of length \( k \).

\[
r = \log_2 \lambda = \log_2 \left( \frac{\pi k}{2} \right)
\]

Therefore, the information length is expressed as

\[
k = \frac{2(2^r-1)^2}{\pi}
\]

Fig. 1. Comparison of information word-length vs. redundancy for various schemes

Fig. 1 presents the performance of various schemes on the information length vs. redundancy including the proposed method. The redundancy of the proposed scheme presented in (16) outer-performs almost all art-of-state schemes and it is only comparable to the minimum achievable redundancy as presented in (4). These results are significantly close to capacity. And as \( k \) get larger, it achieves the theoretical minimum value of the redundancy.

This is a significant improvement from state-of-art schemes in terms of redundancy at a relatively low complexity. However, the concatenation of the prefix \( p \) to the encoded codeword \( x' \). \( p = (p|x')0 \) may not always guarantee overall balancing. Therefore, few additional digits may be added in order to force overall balancing. In [14], a construction was presented to perform overall balancing based on Gray code and additional symbols; a similar approach may be applied in this work. In [11], it was suggested that the prefix could be balanced through a look-up table or using Knuth’s balancing scheme. We will keep this aspect for future investigations. Nevertheless, even by taking into consideration the overall balancing, the proposed scheme redundancy always converges to the capacity as \( k \) gets larger.

B. Complexity

The complexity of the proposed scheme is essentially equivalent to that of the AWSR algorithm which generates the subset \( s(x') \) of words associated to \( x' \) for encoding and decoding. The construction of each subset \( s(x') \) grows linearly with the information word length \( k \); since \( s(x') \leq \lambda \), we conclude that it takes \( O(\log_2 \lambda) \) to generate each subset \( s(x') \) following the AWSR procedure presented in algorithm 1. The generated
subsets have their words in the ordered list corresponding to the lexicographic order. On the other hand, the ranking of a word $x$ within the ordered subset $s(x_i)$ is a straightforward procedure consisting of counting all words occurring before $x$. Therefore, the overall complexity of the proposed method for the encoding and decoding is $O(\log_2 \lambda)$ operation digits.

V. CONCLUSION

We have presented an efficient and simple scheme for constructing binary balanced codes. This is a fixed length scheme that has a prefix length of $\log_2 \lambda$. This method is attractive as it does not make use of look-up tables or enumerative encoding. This is less redundant scheme compared to various existing ones at a relatively low complexity. We have shown the association based on word subset relocation AWSR algorithm which frames the foundation of this work. Also, the decoding can be performed in parallel which reduces latency and saves up memory.

Future works include investigating how to achieve the minimum redundancy by combining the multiplicity of balancing points to the proposed AWSR algorithm.

ACKNOWLEDGEMENT

This work is supported partially by the Centre for Telecommunication under the Global Excellence Stature program under the University of Johannesburg.

REFERENCES

Mr. Elie Ngomseu Mambou graduated from the prestigious University of Johannesburg, South Africa. He holds in his academic portfolio, two bachelor’s degree in Information and Technology and Electrical and Engineering with distinctions in 2014. In 2016, he bagged two master’s degree in Telecommunication Engineering from Beijing Institute of Technology, China and University of Johannesburg in with distinction. He is a recipient of the Chancellor's award for being the best graduating student in the entire University of Johannesburg in 2017 academic session. He has a solid academic background and has been into active teaching, research and development in institutions of higher learning to date with various international conferences and journals publications couple with past industrial experience. Currently, he is rounding up his PhD degree at the University of Johannesburg, South Africa. His research interests are in coding, information theory, Fifth Generation (5G) Wireless Networks, Cognitive Radio Networks, Smart Grid Networks, IoT/IoE, SDN/SDR, Wireless Sensor Networks, Artificial Intelligence, Mobile Computing, Visible light communication. He is a registered as a graduate Engineer with ECSA and a member of SAIEE/IEEE region 8.

Theo G. Swart was born and educated in South Africa. He received the B. Ing., M. Ing.,and PhD degrees in Electrical and Electronic Engineering from the Rand Afrikaans University now University of Johannesburg. He is currently employed as an Associate Professor in the Department of Electrical and Electronic Engineering Science at the University of Johannesburg. He currently has over 100 research papers already published in international Journals and conference proceedings. Prof Theo G. Swart is currently the Director for Centre for Telecommunication (CTT) Research group. He worked closely with his mentor and supervisor, H C Ferreira on the introduction and development a new theme in Information Theory, namely coding techniques for constructing combined channel codes, where error correction and channel properties are considered jointly. He also, work as a member of the pioneering initiator and stimulator of the research fields of Information Theory and Power Line Communications in South Africa. He has been an organizer for the IEEE Information Theory Society and power line communications within South Africa and Africa. His research interest is but not limited to Digital communications power-line communications; cognitive radio networks; smart grid; visible light communications; information theory and emphasis on coding techniques.

Hendrik Christoffel Ferreira is a Professor in Digital Communications and Information Theory at the University of Johannesburg, Johannesburg, South Africa. He studied electrical engineering at the University of Pretoria, South Africa, where he obtained his Ph.D. in 1980. He worked as a visiting researcher at Likebait in San Diego. He joined the Rand Afrikaans University in 1983, where, in 1989, he was appointed full professor. In recognition of his excellence in research and educating post-graduate students, he has been appointed as a research professor at the University of Johannesburg in 2007. He is a Fellow of the SAIEE, the South African Institute of Electrical Engineers. He has published over to 300 research papers on topics such as digital communications, power line communications, vehicular communication systems. With his work he introduced and developed a new theme in Information Theory, namely coding techniques for constructing combined channel codes, where error correction and channel properties are considered jointly. Ferreira is a pioneering initiator and stimulator of the research fields of Information Theory and Power Line Communications in South Africa. He has also been an organizer for the IEEE Information Theory Society and Power Line Communications within South Africa and Africa. He is a member of the Technical Committee for Power Line Communications of the IEEE Communications Society, and he has served on the Technical Program Committee of several IEEE conferences, including the IEEE (ISIT) International Symposium on Information Theory, the IEEE (ISPLC) International Symposium on Power Line Communications, and the IEEE Africon and Chinacom conferences. He is a Fellow of SAIEE and Senior Member of IEEE region 8.