Capacity-Approaching Non-Binary Balanced Codes using Auxiliary Data

Filip Palunčić and B. T. Maharaj

***Abstract*—It is known that, for large user word lengths, auxiliary data can be used to recover most of the redundancy loss of Knuth's simple balancing method as compared to the optimal redundancy of balanced codes for the binary case. Here, this important result is extended in a number of ways. Firstly, an upper bound for the amount of auxiliary data is derived that is valid for all codeword lengths. This result is primarily of theoretical interest, as it defines the probability distribution of the number of balancing indices that results in optimal redundancy. This result is equally valid for particular non-binary generalizations of Knuth's balancing method. Secondly, an asymptotically exact expression for the amount of auxiliary data for the ternary case of a variable length realization of the modified balanced code construction by Pelusi *et al.* is derived, that, in all respects, is the analogue of the result obtained for the binary case. The derivation is based on a generalization of the binary random walk to the ternary case and a simple modification of an existing generalization of Knuth's method for non-binary balanced codes. Finally, a conjecture is proposed regarding the probability distribution of the number of balancing indices for any alphabet size.**

***Index Terms*—Generalized Knuth balancing schemes, non-binary balanced codes, auxiliary data, random walk**

I. Introduction

Binary balanced codes consist of balanced sequences which contain an equal number of the two binary symbols. Such codes have found applications in magnetic and optical storage media [1], [2]. A naive approach to the encoding and decoding of balanced codes is simply to use look-up tables. While this approach can result in optimal codes in terms of redundancy, since the size of such tables grows exponentially with codeword length, this method is only appropriate for short codeword lengths. An alternative approach was proposed by Knuth [3], which is conducive to the encoding and decoding of long codeword lengths. The simplicity and elegance of Knuth's balancing method does come at the cost of greater redundancy. For a user word length m, the number of required redundant symbols is approximately $\log_2 m$, whereas the redundancy of the full binary balanced set of words of length m is approximately $\frac{1}{2}\log_2 m + \frac{1}{2}\log_2 \frac{\pi}{2}$ (cf. [4, p. 1673]). Therefore, the redundancy of Knuth's balancing method is approximately a factor of two greater than that of the full balanced set.

Knuth's first binary balancing method consists of converting a user sequence of length m to a balanced sequence of equal length by inverting (complementing) each bit from the beginning of the user sequence, one at a time, until a balanced sequence is obtained. Knuth showed that at least one index exists such that the initial complemented segment and the subsequent non-complemented segment result in a balanced sequence. This index is encoded as a balanced sequence which is concatenated, as a prefix or suffix, to the altered user word to form a balanced codeword. An alternative approach, also proposed by Knuth, does not require the altered user word and the prefix/suffix to be balanced, but does require the "unbalance" of the two parts to cancel, so that the concatenation thereof is balanced. Knuth's binary balancing methods can then be succinctly classified into two types:

1) The two constituent components (prefix/suffix and altered user word) are both balanced,
2) The two constituent components (prefix/suffix and altered user word) need not be balanced.

The latter type has the potential of having lower redundancy, as the former type is more restrictive than the latter. Results and improvements relating to the former type are contained in [4], [5], while those relating to the latter are contained in [6]–[9]. The result of [10] is pertinent to both types.

Weber and Immink [4] introduced the concept of auxiliary data for Knuth's simple balancing scheme, whose attractiveness stems from its retention of the simplicity of Knuth's balancing method whilst recovering most of the redundancy loss as compared to the redundancy of the full balanced set for large user word lengths. This technique utilizes the degrees of freedom offered in selecting from potentially multiple balancing indices and exploiting this to transmit additional user data per codeword, thereby reducing the overall redundancy. Generalizations of Knuth's binary balancing methods for non-binary alphabets have been proposed in [11, §5.2], [12]–[16]. As stated by Weber *et al.*, within the context of defining and presenting results related to various classes of non-binary balanced codes, "It is an interesting research challenge to investigate whether such techniques (*viz. auxiliary data [note by authors of this article]*) are also applicable in non-binary cases" [17, §IV-F]. In this respect, our main contribution is the derivation of the asymptotic amount of auxiliary data for a variable length realization of a modification of the ternary code construction of Pelusi *et al.* [16]. The importance of this result is that, as in the binary case, the use of auxiliary data allows for the definition of ternary balanced codes that are capacity approaching for large user word lengths. Furthermore, this result allows us to formulate a conjecture regarding the asymptotic amount of auxiliary data, and the corresponding

F. Palunčić and B. T. Maharaj are with the Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria 0002, South Africa (e-mail: fpaluncic@gmail.com, sunil.maharaj@up.ac.za).

asymptotic probability distribution of the number of balancing indices, for any non-binary alphabet size. Although we were not able to prove this conjecture formally, it is supported by simulation results. An attractive feature of this conjecture is that it reduces to our result for the ternary case and to the result of Weber and Immink [4] for the binary case.

In this article, we aim to contribute further aspects regarding auxiliary data and to extend this approach to non-binary generalizations of Knuth's method. In particular, after a short overview of Knuth's method, auxiliary data and existing pertinent non-binary generalizations of Knuth's method (Section II), we derive an upper bound for the amount of auxiliary data for the binary and non-binary cases (Section III), present a generalized random walk as applicable to ternary sequences (Section IV), and derive an asymptotic expression for the amount of auxiliary data for the ternary case based on a variable length realization of a modification of the construction by Pelusi *et al.* [16], that is parallel to that for the binary case derived by Weber and Immink [4] (Section V). The results and open problems are summarized in Section VI. Finally, we note that, like Weber and Immink [4] for the binary case, we limit ourselves to non-binary generalizations of Knuth's method of the first type, i.e. where both the prefix/suffix and altered user word are balanced.

II. Overview of Knuth's Method, Auxiliary Data and Non-Binary Balanced Codes

This section serves as an introduction to the pertinent concepts and known results and to establish some of the notation used in this article. We denote by $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \{0,1\}^m$ a binary sequence of length m, where m is even. Then, the sequence $\mathbf{x}$ is said to be balanced if

$$\sigma(\mathbf{x}) \triangleq \sum_{i=1}^{m} x_i = m/2.$$

A. Knuth's Binary Balancing Method

Let $\mathbf{x}^{(k)}$ denote $\mathbf{x}$ with the first k bits complemented, i.e. $\mathbf{x}^{(k)} \triangleq (\bar{x}_1, \ldots, \bar{x}_k, x_{k+1}, \ldots, x_m)$, where $\bar{x}_i \triangleq 1 - x_i$ is the complement of x_i. Then, Knuth's method iterates through k, $1 \le k \le m$, starting at $k = 1$, until the first $\sigma(\mathbf{x}^{(k)}) = m/2$ is found. Any k such that $\mathbf{x}^{(k)}$ is balanced is referred to as a *balancing index*. It is guaranteed that for any sequence $\mathbf{x}$ there exists at least one balancing index k [3, pp. 51–52] (cf. also [4, §II]). In fact, there may be multiple balancing indices k, but Knuth's method does not exploit this possibility. The auxiliary data approach [4] uses these degrees of freedom in selecting the balancing index to transmit additional user data per codeword.

To retrieve the original user word $\mathbf{x}$ from $\mathbf{x}^{(k)}$, it is sufficient to know the value of the balancing index k. This information is encoded by means of a balanced prefix/suffix $\mathbf{p} \in \{0,1\}^p$ (i.e. $\sigma(\mathbf{p}) = p/2$) of even length p. With each balancing index k, $1 \le k \le m$, is associated a unique balanced prefix/suffix $\mathbf{p}_k$. Therefore, it follows that the number of possible balancing indices m must be less or equal to the number of balanced sequences of length p, i.e.

$$m \le \binom{p}{p/2}. \tag{1}$$

In the rest of this article, we assume that the balancing index is encoded using a suffix. Therefore, subsequently, depending on implementation, each reference to suffix can be replaced by prefix.

B. Auxiliary Data

As mentioned previously, auxiliary data can be used to recover most of the rate loss of Knuth's balancing method by exploiting the degrees of freedom in selecting the balancing index.

Denote by the random variable V_m the number of balancing indices of binary sequences of length m, m even. Using combinatorial arguments and assuming that all user words are equiprobable and independent, Weber and Immink [4, Thm. 3] have derived the explicit expression

$$P(V_m = v) = 2^{v+1-m}\binom{m-1-v}{m/2-v}, \tag{2}$$

where $1 \le V_m \le m/2$ [4, Thm. 2]. If a user word provides v balancing indices, the option of selecting amongst these v indices allows for an additional $\log_2 v$ bits to be transmitted. Since the number of balancing indices is user word dependent, the average amount of auxiliary data is given by

$$H_{\text{aux}}(m) \triangleq \sum_{v=1}^{m/2} P(V_m = v) \log_2 v. \tag{3}$$

Using (2), Weber and Immink [4, p. 1676] obtained an approximation of $P(V_m = v)$ through a series of intermediate approximations. Then, by replacing the summation in (3) with the corresponding integral, Weber and Immink [4, Eq. (12)] derived an approximation $H_{\text{aux}}(m) \approx \frac{1}{2}\log_2 m - 0.916$.

As shown in the Appendix, using an alternative derivation, the approximations of $P(V_m = v)$ used by Weber and Immink [4] are in fact asymptotically exact. In particular, it is shown that

$$\frac{V_m}{\sqrt{m}} \xrightarrow{d} Y,$$

where $\xrightarrow{d}$ denotes convergence in distribution, $Y \triangleq |Z|$, where $|Z|$ is the absolute value of Z, and Z is a standard normally distributed random variable. This result may also be stated equivalently as

$$P(V_m \ge y\sqrt{m}) \to 2[1 - \Phi(y)],$$

where $\Phi(y)$ is the standard normal cumulative distribution.

A consequence of this result is that the approximation of $H_{\text{aux}}(m)$ as used by Weber and Immink [4, Eq. (10)]

$$H_{\text{aux}}(m) \approx \sqrt{\frac{2}{\pi(m-2)}} \int_{v=0}^{\infty} e^{\frac{-v^2}{2m}} \log_2 v \, \mathrm{d}v$$

is also asymptotically exact. In particular, since $P(V_m \geq y\sqrt{m}) \to 2[1-\Phi(y)]$, the approximation

$$H_{\text{aux}}(m) \approx \sqrt{\frac{2}{\pi}} \int_{v=0}^{\infty} e^{\frac{-v^2}{2}} \log_2(v\sqrt{m})\,\mathrm{d}v \tag{4}$$

is exact as $m \to \infty$ in the sense that the ratio of the two sides of (4) tend to unity as $m \to \infty$. By setting $u \triangleq v\sqrt{m}$, it follows that the right side of (4) is equivalent to

$$\sqrt{\frac{2}{\pi m}} \int_{u=0}^{\infty} e^{\frac{-u^2}{2m}} \log_2 u\,\mathrm{d}u = \frac{1}{2}\log_2 m - \frac{1}{2}\left(1+\frac{\gamma}{\ln 2}\right), \tag{5}$$

which is the same as [4, Eq. (11)], where the evaluation of the improper integral follows from [18, Eq. (4.333), p. 575] and $\gamma \doteq 0.57722$ is Euler's constant. Therefore

$$H_{\text{aux}}(m) \sim \frac{1}{2}\log_2 m - \frac{1}{2}\left(1+\frac{\gamma}{\ln 2}\right), \tag{6}$$

where $a(m) \sim b(m)$ means that $\lim_{m\to\infty} a(m)/b(m) = 1$. Furthermore, the redundancy of Knuth's method, p, is minimized when (1) is satisfied as an equality. Then, by using Stirling's formula

$$n! = \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \left(1+O\Big(\frac{1}{n}\Big)\right),$$

we obtain that

$$p = \log_2 m + \frac{1}{2}\log_2 p + \frac{1}{2}\log_2\frac{\pi}{2} + \log_2\left(1+O\Big(\frac{1}{p}\Big)\right). \tag{7}$$

The above equation implies that $p = \log_2 m + \Theta(\log\log m)$ (cf. [16, §IV]). And, so

$$\lim_{m\to\infty} p - \log_2 m - \frac{1}{2}\log_2\log_2 m - \frac{1}{2}\log_2\frac{\pi}{2} = 0. \tag{8}$$

On the other hand, let r denote the redundancy of the full balanced set consisting of words of length n. Then, using Stirling's formula, it follows that

$$\begin{aligned} r &= n - m \\ &= n - \log_2\binom{n}{n/2} \\ &= \frac{1}{2}\log_2 n + \frac{1}{2}\log_2\frac{\pi}{2} + \log_2\left(1+O\Big(\frac{1}{n}\Big)\right) \qquad (9)\\ &= \frac{1}{2}\log_2(m+r) + \frac{1}{2}\log_2\frac{\pi}{2} + \log_2\left(1+O\Big(\frac{1}{m+r}\Big)\right). \qquad (10)\end{aligned}$$

From (9), it is evident that $r = O(\log_2 n)$, and since $r < m$ for sufficiently large m, it follows from (10) that $r = O(\log_2 m)$. Then, it is easily shown that

$$\lim_{m\to\infty} r - \frac{1}{2}\log_2 m - \frac{1}{2}\log_2\frac{\pi}{2} = 0. \tag{11}$$

Using the right-hand side of (6) as an approximation of the amount of auxiliary data for large user word lengths and comparing (8) and (11), we conclude that, for large user word lengths, auxiliary data is able to recover approximately $\frac{1}{2}\log_2 m$ information bits, which is most of the redundancy loss of Knuth's method as compared to the optimal redundancy.

C. Non-Binary Balanced Codes

We use a q-ary alphabet $\mathbb{Z}_q = \{0,1,\ldots,q-1\}$, where $q \geq 2$. Then, a q-ary sequence $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_q^m$ is balanced if

$$\sigma(\mathbf{x}) = \sum_{i=1}^{m} x_i = \frac{(q-1)m}{2}. \tag{12}$$

As in [15], we disregard the case where q is even and m is odd, as in this case (12) cannot be satisfied.

Based on the cardinality of the full non-binary balanced set (derived by Star [19, Thm. 2] within the context of compositions and first related to non-binary balanced sets by Capocelli *et al.* [20, Lemma 2.3]), i.e. the set consisting of all non-binary balanced sequences of a given length, the redundancy of the full q-ary balanced set consisting of sequences of length m can be approximated by (cf. the alternative derivations in [17, Thm. 3 & Cor. 4] and [21, Eq. (46)])

$$\frac{1}{2}\log_q m + \frac{1}{2}\log_q\frac{\pi(q^2-1)}{6}$$

for large m. This approximation is asymptotically exact in the sense that the ratio of the actual redundancy to the approximation tends to unity as word length m tends to infinity.

Amongst the proposed non-binary balancing codes which are generalizations of Knuth's binary balancing schemes (e.g. [11, §5.2], [12]–[16]), we limit ourselves to codes of type 1 as defined in Section I, i.e. where the altered user word and suffix are both balanced, namely the codes by Swart and Weber [15] and Pelusi *et al.* [16, §II].

Swart and Weber [15] base their generalization on the observation that Knuth's binary complementation scheme can be equivalently described as the binary addition of the user word with a binary sequence of equal length consisting of an all '1's segment followed by an all '0's segment. If

$$[s]^k \triangleq (\overbrace{s,\ldots,s}^{k}),$$

then $\mathbf{x}^{(k)} = \mathbf{x} \oplus_2 ([1]^k, [0]^{m-k})$, $1 \leq k \leq m$, where $\oplus_2$ denotes modulo 2 integer addition and $(.,.)$ denotes the concatenation of two sequences. The generalization to the non-binary case follows as

$$\mathbf{x}^{(k)} \triangleq \mathbf{x} \oplus_q [i]^m \oplus_q ([1]^j, [0]^{m-j}), \tag{13}$$

where $\mathbf{x} \in \mathbb{Z}_q^m$, $i \in \{0,1,\ldots,q-1\}$, $j \in \{0,1\ldots,m-1\}$, $k(i,j) = im+j$ and $\oplus_q$ denotes modulo q integer addition [15, p. 1565]. It is guaranteed that for any sequence $\mathbf{x} \in \mathbb{Z}_q^m$ there exists at least one k such that $\mathbf{x}^{(k)}$ is balanced [15, Thm. 1].

Pelusi *et al.* [16] take a different approach. Here, we only summarize their simple scheme [16, §II], which is of type 1 as defined in Section I. They complement each symbol, starting from the first symbol, in stages such that the symbol at the final stage is the complement of the original symbol. In the intermediate stages between the original and the complemented symbol, all other symbols of $\mathbb{Z}_q$ are enumerated. At the heart of their scheme is a function

$$\phi : \mathbb{Z}_q \times \{0,1,\ldots,l\} \to \mathbb{Z}_q, \tag{14}$$

which complements each symbol $x \in \mathbb{Z}_q$ in l stages, where $l = q - 1 + (q \bmod 2)$. Then ϕ must satisfy the following properties [16, Eqs. (4)–(6)]:

1) $\forall\, x \in \mathbb{Z}_q$, $\phi(x, 0) = x$ and $\phi(x, l) = \bar{x}$, where $\bar{x} \triangleq (q-1) - x$,
2) $\forall\, x \in \mathbb{Z}_q$ and $\forall\, y \in \{\min[x, \bar{x}], \min[x, \bar{x}] + 1, \ldots, \max[x, \bar{x}]\}$, there exists $j \in \{0, 1, \ldots, l\}$ such that $\phi(x, j) = y$,
3) $\forall\, j \in \{0, 1, \ldots, l\}$ and $\forall\, x, x' \in \mathbb{Z}_q$, if $x \neq x'$ then $\phi(x, j) \neq \phi(x', j)$.

Then

$$\mathbf{x}^{(k)} \triangleq (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{i-1}, \phi(x_i, j), x_{i+1}, \ldots, x_{m-1}, x_m), \tag{15}$$

where $i \in \{1, 2, \ldots, m + (q \bmod 2)\}$, $j \in \{1, 2, \ldots, q-1\}$, $k(i, j) \triangleq (i-1)(q-1) + j$ and $\mathbf{x}^{(0)} \triangleq \mathbf{x}$. For further details regarding this construction, including the bidirectional mappings between k and (i, j) and the optimal properties of ϕ, the reader is referred to [16, §II]. The following example illustrates the complementation process.

Example 1: Consider the ternary word $\mathbf{x} = (2102) \in \mathbb{Z}_3^4$. Using ϕ from [16, Eq. (8)],

$$\begin{pmatrix} \phi(0,0) & \phi(0,1) & \phi(0,2) & \phi(0,3) \\ \phi(1,0) & \phi(1,1) & \phi(1,2) & \phi(1,3) \\ \phi(2,0) & \phi(2,1) & \phi(2,2) & \phi(2,3) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{pmatrix},$$

we have

$$\begin{aligned}
k = 0\,(i = 0, j = 2): &\quad \mathbf{x}^{(0)} = (2102),\ \sigma(\mathbf{x}^{(0)}) = 5, \\
k = 1\,(i = 1, j = 1): &\quad \mathbf{x}^{(1)} = (0102),\ \sigma(\mathbf{x}^{(1)}) = 3, \\
k = 2\,(i = 1, j = 2): &\quad \mathbf{x}^{(2)} = (\mathbf{1102}),\ \sigma(\mathbf{x}^{(2)}) = 4, \\
k = 3\,(i = 2, j = 1): &\quad \mathbf{x}^{(3)} = (\mathbf{0202}),\ \sigma(\mathbf{x}^{(3)}) = 4, \\
k = 4\,(i = 2, j = 2): &\quad \mathbf{x}^{(4)} = (0002),\ \sigma(\mathbf{x}^{(4)}) = 2, \\
k = 5\,(i = 3, j = 1): &\quad \mathbf{x}^{(5)} = (\mathbf{0112}),\ \sigma(\mathbf{x}^{(5)}) = 4, \\
k = 6\,(i = 3, j = 2): &\quad \mathbf{x}^{(6)} = (0122),\ \sigma(\mathbf{x}^{(6)}) = 5, \\
k = 7\,(i = 4, j = 1): &\quad \mathbf{x}^{(7)} = (0120),\ \sigma(\mathbf{x}^{(7)}) = 3, \\
k = 8\,(i = 4, j = 2): &\quad \mathbf{x}^{(8)} = (\mathbf{0121}),\ \sigma(\mathbf{x}^{(8)}) = 4, \\
k = 9\,(i = 5, j = 1): &\quad \mathbf{x}^{(9)} = (0120),\ \sigma(\mathbf{x}^{(9)}) = 3.
\end{aligned}$$

Balanced words ($\sigma(\mathbf{x}^{(k)}) = 4$), which are shown in bold, are obtained for $k = 2, 3, 5, 8$. □

These two non-binary balancing schemes can be seen, in a certain sense, as orthogonal to each other. Whereas the scheme by Pelusi *et al.* [16] processes completely a symbol at a given index before moving to the next index, the scheme by Swart and Weber [15] processes a given "level" (corresponding to some symbol in $\mathbb{Z}_q$) for all indices before moving on to the next level, effectively returning to the same symbol index multiple times. In their own way, both these schemes are generalizations of Knuth's binary balancing method.

While the results of the next section are applicable to both these schemes, further results expounded thereafter relate only to the construction by Pelusi *et al.* [16] and seem to have no apparent analogue for the construction by Swart and Weber [15].

III. Upper Bound for Average Amount of Auxiliary Data

As demonstrated by Weber and Immink [4], the auxiliary data technique is able to recover the major part of the redundancy loss of Knuth's binary balancing method. As already indicated, this is achieved by exploiting the multiplicity of balancing indices of certain user words to transmit auxiliary data. This potentially implies that the set of all balanced words obtained by Knuth's method for all balancing indices is equivalent to the full balanced set consisting of sequences of length $m+p$, where the sub-sequences of length m and p are also balanced. This is precisely what we prove in this section. This result leads naturally to an upper bound for $H_{\text{aux}}(m)$ valid for all $m \geq 2$, which in turn identifies the probability distribution of the number of balancing indices $P(V_m = v)$ which maximizes $H_{\text{aux}}(m)$.

Before presenting the proofs, we set the stage with additional notation. By $\mathcal{B}(m)$ we denote the set of all binary balanced words of length m, m even, i.e.

$$\mathcal{B}(m) \triangleq \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_2^m : \sigma(\mathbf{x}) = m/2\Big\}.$$

Then

$$B(m) \triangleq |\mathcal{B}(m)| = \binom{m}{m/2}.$$

In this article we follow the convention of denoting sets by calligraphic upper case letters and their cardinality by the non-calligraphic counterpart. Similarly, we define $\mathcal{B}(m, p) \triangleq \mathcal{B}(m) \times \mathcal{B}(p)$, m and p even, and hence $B(m, p) \triangleq |\mathcal{B}(m, p)| = B(m) \cdot B(p)$, and note that balanced codewords resulting from Knuth's simple balancing scheme must be from $\mathcal{B}(m, p)$.

The set of m balanced suffixes of length p is denoted by $\mathcal{P}$, i.e.

$$\mathcal{P} \triangleq \Big\{\mathbf{p}_k : \mathbf{p}_k \in \mathcal{B}(p), 1 \leq k \leq m\Big\}.$$

Furthermore, $\boldsymbol{\psi}$ is a bijective mapping between possible balancing indices and corresponding balanced suffixes

$$\boldsymbol{\psi} : \{1, 2, \ldots, m\} \to \mathcal{P}, \boldsymbol{\psi}(k) = \mathbf{p}_k.$$

Finally, since $\mathcal{P} \subseteq \mathcal{B}(p)$ (with equality if $m = \binom{p}{p/2}$), we also define $\mathcal{B}^{(\mathcal{P})}(m) \triangleq \mathcal{B}(m) \times \mathcal{P} \subseteq \mathcal{B}(m, p)$, so that $B^{(\mathcal{P})}(m) \triangleq |\mathcal{B}^{(\mathcal{P})}(m)| = mB(m)$ as $|\mathcal{P}| = m$. It is easy to see that any balanced sequence obtained using Knuth's simple balancing method must be from $\mathcal{B}^{(\mathcal{P})}(m)$.

Denote by $\mathcal{D}_v(m)$ the set of binary sequences of length m, m even, that have v balancing indices, $1 \leq v \leq m/2$, i.e.

$$\mathcal{D}_v(m) \triangleq \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_2^m : \sigma(\mathbf{x}^{(k_i)}) = m/2, \\ 1 \leq k_i \leq m, 1 \leq i \leq v\Big\},$$

and $D_v(m) \triangleq |\mathcal{D}_v(m)|$. We note that

$$P(V_m = v) = D_v(m)/2^m.$$

We aim to show that

$$\sum_{v=1}^{m/2} vD_v(m) = B^{(\mathcal{P})}(m). \tag{16}$$

We first outline the approach for the binary case before formalizing the results for the general case of $\mathbb{Z}_q$. To prove (16), it suffices to show that each unique pair of $\mathbf{x} \in \mathbb{Z}_2^m$ and balancing index k_i for $\mathbf{x}$, $1 \le k_i \le m$ and $1 \le i \le v$, results in a unique balanced sequence using Knuth's balancing method and that all balanced words in $\mathcal{B}^{(\mathcal{P})}(m)$ can be obtained from some $\mathbf{x} \in \mathbb{Z}_2^m$ using Knuth's method. To demonstrate the first part, consider an arbitrary $\mathbf{x} \in \mathcal{D}_v(m)$. Since $\boldsymbol{\psi}$ is a bijection, $\boldsymbol{\psi}(k_i) \neq \boldsymbol{\psi}(k_{i'})$ if $k_i \neq k_{i'}$, and therefore $(\mathbf{x}^{(k_i)}, \mathbf{p}_{k_i}) \neq (\mathbf{x}^{(k_{i'})}, \mathbf{p}_{k_{i'}})$ for all $i \neq i'$, $1 \le i, i' \le v$, where $(\mathbf{a}, \mathbf{b})$ denotes the concatenation of two words $\mathbf{a}$ and $\mathbf{b}$. Furthermore, consider arbitrary words $\mathbf{x}$ and $\mathbf{x}'$, $\mathbf{x} \neq \mathbf{x}'$, where $\mathbf{x} \in \mathcal{D}_v(m)$ and $\mathbf{x}' \in \mathcal{D}_{v'}(m)$ with $v \neq v'$ or $v = v'$. In order for $(\mathbf{x}^{(k_i)}, \mathbf{p}_{k_i}) = (\mathbf{x}'^{(k_{i'})}, \mathbf{p}_{k_{i'}})$, where $1 \le i \le v$ and $1 \le i' \le v'$, the suffixes must be equal $\boldsymbol{\psi}(k_i) = \boldsymbol{\psi}(k_{i'})$, implying that $k_i = k_{i'}$. However, since $\mathbf{x} \neq \mathbf{x}'$, it follows that $\mathbf{x}^{(k_i)} \neq \mathbf{x}'^{(k_i)}$. Therefore, we conclude that a unique pair of user word and balancing index guarantee uniqueness of the balanced word under Knuth's method. On the other hand, the words in $\mathcal{B}^{(\mathcal{P})}(m)$ are obtained by juxtaposing all combinations of words from $\mathcal{B}(m)$ and $\mathcal{P}$. Consider an arbitrary balanced word $\mathbf{y} \in \mathcal{B}(m)$. Then, for each k, $1 \le k \le m$, there exists a user word $\mathbf{x} = \mathbf{y}^{(k)}$, which results in the balanced word $(\mathbf{y}, \mathbf{p}_k)$. Therefore, for a given $\mathbf{y} \in \mathcal{B}(m)$, all suffixes $\boldsymbol{\psi}(k) \in \mathcal{P}$, $1 \le k \le m$, can be obtained. Since this is true for any $\mathbf{y} \in \mathcal{B}(m)$, all balanced words in $\mathcal{B}^{(\mathcal{P})}(m)$ can be obtained using Knuth's method. Combining these two observations, we conclude that (16) is true.

In traversing to the non-binary domain, the above definitions remain valid with the appropriate substitution of $\mathbb{Z}_2$ with $\mathbb{Z}_q$. Then, we have

$$\begin{aligned}
\mathcal{B}(q,m) &\triangleq \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_q^m : \\
&\qquad\qquad \sigma(\mathbf{x}) = (q-1)m/2\Big\}, \\
\mathcal{B}(q,m,p) &\triangleq \mathcal{B}(q,m) \times \mathcal{B}(q,p), \\
\mathcal{P}_q &\triangleq \Big\{\mathbf{p}_k : \mathbf{p}_k \in \mathcal{B}(q,p), 0 \le k \le \mu(q,m)\Big\}, \\
\boldsymbol{\psi}_q &: \{0,1,\ldots,\mu(q,m)\} \to \mathcal{P}_q, \boldsymbol{\psi}_q(k) = \mathbf{p}_k, \\
\mathcal{B}^{(\mathcal{P}_q)}(q,m) &\triangleq \mathcal{B}(q,m) \times \mathcal{P}_q,
\end{aligned}$$

where the balancing indices k run from 0 to $\mu(q,m)$, and the maximum balancing index μ is a function of q and m and its value is dependent on the non-binary balancing scheme ($\mu(q,m) = qm-1$ for [15] and $\mu(q,m) = (q-1)m + (q \bmod 2)$ for [16]).

For the alphabet $\mathbb{Z}_q$, we denote the "complementation" of the user word at position $k(i,j)$ by the function

$$\boldsymbol{\beta}_q : \mathbb{Z}_q^m \times \{0,1,\ldots,\mu(q,m)\} \to \mathbb{Z}_q^m, \boldsymbol{\beta}_q(\mathbf{x},k) = \mathbf{y}. \quad (17)$$

For Knuth's binary method $\boldsymbol{\beta}_2(\mathbf{x},k) = \mathbf{x}^{(k)}$, while more generally $\boldsymbol{\beta}_q(\mathbf{x},k) = \mathbf{x}^{(k)}$, where $\mathbf{x}^{(k)}$ is given by (13) for the scheme by Swart and Weber [15] and by (15) for the scheme by Pelusi *et al.* [16]. The inverse of $\boldsymbol{\beta}_q$ is similarly defined, so that $\mathbf{x} = \boldsymbol{\beta}_q^{-1}(\mathbf{y},k)$. We are now in a position to generalize $\mathcal{D}_v(m)$ as

$$\begin{aligned}
\mathcal{D}_v(q,m) &\triangleq \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_q^m : \\
&\sigma(\boldsymbol{\beta}_q(\mathbf{x},k_i)) = (q-1)m/2, 0 \le k_i \le \mu(q,m), 1 \le i \le v\Big\}.
\end{aligned}$$

The following lemma applies to any non-binary generalization of Knuth's method of type 1 (defined in Section I) with a "complementation" function $\boldsymbol{\beta}_q$.

Lemma 1: If $\boldsymbol{\beta}_q(\mathbf{x},k) \neq \boldsymbol{\beta}_q(\mathbf{x}',k)$ for all $\mathbf{x} \neq \mathbf{x}'$, where $\mathbf{x}, \mathbf{x}' \in \mathbb{Z}_q^m$, and for all k, $0 \le k \le \mu(q,m)$, then

$$\sum_{v=1}^{\nu(q,m)} vD_v(q,m) = B^{(\mathcal{P}_q)}(q,m), \quad (18)$$

where $\nu(q,m)$ is the maximum number of balancing indices for a specific user word length m and alphabet size q.

Proof: To prove (18), it suffices to show that a unique pair of user word $\mathbf{x} \in \mathbb{Z}_q^m$ and balancing index $k_l(i,j)$, $0 \le k_l(i,j) \le \mu(q,m)$ and $1 \le l \le v$, results in a unique balanced word under the operation of the generalized version of Knuth's method, i.e. after the application of $\boldsymbol{\beta}_q(\mathbf{x},k_l)$ and $\boldsymbol{\psi}_q(k_l)$, and that every balanced word in $\mathcal{B}^{(\mathcal{P}_q)}(q,m)$ can be obtained from some user word $\mathbf{x} \in \mathbb{Z}_q^m$ under the operation of the generalized version of Knuth's method.

To demonstrate the first part, consider an arbitrary $\mathbf{x} \in \mathcal{D}_v(q,m)$. Since $\boldsymbol{\psi}_q$ is a bijection, $\boldsymbol{\psi}_q(k_l) \neq \boldsymbol{\psi}_q(k_{l'})$ if $k_l \neq k_{l'}$, and therefore $(\boldsymbol{\beta}_q(\mathbf{x},k_l), \mathbf{p}_{k_l}) \neq (\boldsymbol{\beta}_q(\mathbf{x},k_{l'}), \mathbf{p}_{k_{l'}})$ for all $l \neq l'$, $1 \le l, l' \le v$, where $(\mathbf{a}, \mathbf{b})$ denotes the concatenation of two words $\mathbf{a}$ and $\mathbf{b}$. Furthermore, consider arbitrary words $\mathbf{x}$ and $\mathbf{x}'$, $\mathbf{x} \neq \mathbf{x}'$, where $\mathbf{x} \in \mathcal{D}_v(q,m)$ and $\mathbf{x}' \in \mathcal{D}_{v'}(q,m)$ with $v \neq v'$ or $v = v'$. In order for $(\boldsymbol{\beta}_q(\mathbf{x},k_l), \mathbf{p}_{k_l}) = (\boldsymbol{\beta}_q(\mathbf{x}',k_{l'}), \mathbf{p}_{k_{l'}})$, where $1 \le l \le v$ and $1 \le l' \le v'$, the suffixes must be equal $\boldsymbol{\psi}_q(k_l) = \boldsymbol{\psi}_q(k_{l'})$, implying that $k_l = k_{l'}$. However, since $\mathbf{x} \neq \mathbf{x}'$, by assumption it follows that $\boldsymbol{\beta}_q(\mathbf{x},k_l) \neq \boldsymbol{\beta}_q(\mathbf{x}',k_l)$. Therefore, we conclude that a unique pair of user word and balancing index guarantee uniqueness of the balanced word under the operation of $\boldsymbol{\beta}_q$ and $\boldsymbol{\psi}_q$ provided that $\boldsymbol{\beta}_q(\mathbf{x},k) \neq \boldsymbol{\beta}_q(\mathbf{x}',k)$ for all $\mathbf{x} \neq \mathbf{x}'$ and $0 \le k \le \mu(q,m)$.

On the other hand, the words in $\mathcal{B}^{(\mathcal{P}_q)}(q,m)$ are obtained by juxtaposing all combinations of words from $\mathcal{B}(q,m)$ and $\mathcal{P}_q$. Consider an arbitrary balanced word $\mathbf{y} \in \mathcal{B}(q,m)$. Then, for each k, $0 \le k \le \mu(q,m)$, there exists a user word $\mathbf{x} = \boldsymbol{\beta}_q^{-1}(\mathbf{y},k)$, which results in the balanced word $(\mathbf{y}, \mathbf{p}_k)$. Therefore, for a given $\mathbf{y} \in \mathcal{B}(q,m)$, all suffixes $\boldsymbol{\psi}_q(k) \in \mathcal{P}_q$, $0 \le k \le \mu(q,m)$, can be obtained. Since this is true for any $\mathbf{y} \in \mathcal{B}(q,m)$, all balanced words in $\mathcal{B}^{(\mathcal{P}_q)}(q,m)$ can be obtained under the operation of $\boldsymbol{\beta}_q$ and $\boldsymbol{\psi}_q$. These two results combined yield (18). ∎

It is straightforward to demonstrate that the condition of Lemma 1 is satisfied by both the schemes of Swart and Weber [15] and Pelusi *et al.* [16]. For the first scheme,

$$\boldsymbol{\beta}_q(\mathbf{x},k) = \mathbf{x} \oplus_q \mathbf{b}(k),$$

where $\mathbf{b}(k) \triangleq [i]^m \oplus_q ([1]^j, [0]^{m-j})$ and $k(i,j) = im + j$ [cf. (13)]. Then, it immediately follows that $\boldsymbol{\beta}_q(\mathbf{x},k) = \mathbf{x} \oplus_q \mathbf{b}(k) \neq \mathbf{x}' \oplus_q \mathbf{b}(k) = \boldsymbol{\beta}_q(\mathbf{x}',k)$ for all $\mathbf{x} \neq \mathbf{x}'$ and $0 \le k \le qm-1$.

For the latter scheme

$$\boldsymbol{\beta}_q(\mathbf{x},k)=(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_{i-1},\phi(x_i,j),x_{i+1},\ldots,x_{m-1},x_m), \tag{19}$$

where $k(i,j)=(i-1)(q-1)+j$ [cf. (15)]. If $\mathbf{x}=(x_1,x_2,\ldots,x_m)\neq\mathbf{x}'=(x'_1,x'_2,\ldots,x'_m)$, then there exists at least one index λ, $1\leq\lambda\leq m$, such that $x_\lambda\neq x'_\lambda$. Then, in comparing $\boldsymbol{\beta}_q(\mathbf{x},k)$ and $\boldsymbol{\beta}_q(\mathbf{x}',k)$, there are three cases:

1) $\lambda\leq i-1$: $x_\lambda\neq x'_\lambda\Rightarrow\bar{x}_\lambda=(q-1)-x_\lambda\neq(q-1)-x'_\lambda=\bar{x}'_\lambda$,
2) $\lambda=i$: $x_\lambda\neq x'_\lambda\Rightarrow\phi(x_\lambda,j)\neq\phi(x'_\lambda,j)$ by the invertibility property [16, Eq. (6)],
3) $\lambda\geq i+1$: $x_\lambda\neq x'_\lambda$.

Therefore $\boldsymbol{\beta}_q(\mathbf{x},k)\neq\boldsymbol{\beta}_q(\mathbf{x}',k)$ for all $\mathbf{x}\neq\mathbf{x}'$ and $0\leq k\leq(q-1)m+(q \bmod 2)$.

Here we introduce the random variable $V_m^{(q)}$, the non-binary extension of V_m, so that $V_m^{(q)}$ denotes the number of balancing indices of q-ary sequences of length m as obtained by some non-binary generalization of Knuth's method. Then, the average amount of auxiliary data achievable is

$$H_{\text{aux}}(q,m)\triangleq\sum_{v=1}^{\nu(q,m)}P(V_m^{(q)}=v)\log_q v, \tag{20}$$

where $\nu(q,m)$ is the maximum number of balancing indices for a given q and m and $P(V_m^{(q)}=v)=D_v(q,m)/q^m$ for some non-binary generalization of Knuth's method defined by the functions $\boldsymbol{\beta}_q$ and $\boldsymbol{\psi}_q$. The following theorem is valid for any non-binary generalizations of Knuth's method of type 1 (defined in Section I) whose function $\boldsymbol{\beta}_q$ satisfies the condition of Lemma 1.

Theorem 1: For $q\geq 2$ and $m\geq 2$,

$$H_{\text{aux}}(q,m)\leq\log_q\left(\frac{B^{(\mathcal{P}_q)}(q,m)}{q^m}\right). \tag{21}$$

Proof: Since $P(V_m^{(q)}=v)=D_v(q,m)/q^m$, relation (18) can be stated equivalently as

$$\sum_{v=1}^{\nu(q,m)}vP(V_m^{(q)}=v)=\frac{B^{(\mathcal{P}_q)}(q,m)}{q^m}. \tag{22}$$

Furthermore, $H_{\text{aux}}(q,m)$ can be stated equivalently as

$$H_{\text{aux}}(q,m)=\log_q\left(\prod_{v=1}^{\nu(q,m)}v^{P(V_m^{(q)}=v)}\right).$$

The well known weighted arithmetic-geometric mean inequality [22, §2.5] states that

$$\prod_{v=1}^{n}x_v^{w_v}\leq\sum_{v=1}^{n}w_vx_v$$

for $w_v>0$ and $x_v>0$, where $v=1,2,\ldots,n$ and $\sum_{v=1}^n w_v=1$. Equality is attained if, and only if, all the x_v are equal. Setting $x_v=v$, $w_v=P(V_m^{(q)}=v)$, $n=\nu(q,m)$, and noting that $\sum_{v=1}^{\nu(q,m)}P(V_m^{(q)}=v)=1$ and that $\log_q$ is a monotonic function, yields (21). ∎

This upper bound admits a simple interpretation. Noting that $m+p-\log_q B^{(\mathcal{P}_q)}(q,m)$ represents the redundancy of the set consisting of the juxtaposition of all balanced words of length m with all balanced suffixes of length p necessary to represent all balancing indices and that $m+p-\log_q q^m$ represents the redundancy of any non-binary generalized form of Knuth's method, the upper bound on the average auxiliary data $H_{\text{aux}}(q,m)$ from (21) is the difference between these two redundancies. Stated alternatively, it is the maximal redundancy gain achievable over a generalized version of Knuth's method without auxiliary data by using a balanced code consisting of all balanced words from $\mathcal{B}^{(\mathcal{P}_q)}(q,m)$.

Example 2: Consider the case $q=2$ and $m=4$. For each user word, the corresponding balanced words obtained using Knuth's method are (underlining denotes complemented bits):

$$\left\{\begin{matrix}(0000)\\(0001)\\(0010)\\(0011)\\(0100)\\(0101)\\(0110)\\(0111)\\(1000)\\(1001)\\(1010)\\(1011)\\(1100)\\(1101)\\(1110)\\(1111)\end{matrix}\right\}\rightarrow\left\{\begin{matrix}(\underline{11}00)&\\(\underline{1}001)&\\(\underline{1}010)&(\underline{1100})\\(\underline{1100})&\\(\underline{1}100)&(\underline{1}0\underline{1}0)\\(\underline{1}001)&(\underline{1}0\underline{1}0)\\(\underline{1}010)&(\underline{1001})\\(\underline{1}00\underline{1})&\\(\underline{0110})&\\(\underline{0}1\underline{0}1)&(\underline{0110})\\(\underline{0}110)&(\underline{0}1\underline{0}1)\\(\underline{0}011)&(\underline{0}1\underline{0}1)\\(\underline{0011})&\\(\underline{0}101)&(\underline{0011})\\(\underline{0}110)&\\(\underline{00}11)&\end{matrix}\right\}.$$

It follows that $D_1(4)=D_2(4)=8$ and $P(V_4=1)=P(V_4=2)=1/2$. Then $H_{\text{aux}}(4)=\sum_{v=1}^{2}P(V_4=v)\log_2 v=1/2$. On the other hand, $B(4)=6$ and so $\log_2(4\cdot B(4)/2^4)\doteq 0.585$, as expected by (21). □

Equality in (21) is attained if, and only if, all v are equal, i.e. if all $\mathbf{x}\in\mathbb{Z}_q^m$ have the same number of balancing indices under the operation of $\boldsymbol{\beta}_q$. If we denote such a v by $\hat{v}$, then, by noting that $P(V_m^{(q)}=\hat{v})=1$, we have from (22) that

$$\hat{v}=\frac{B^{(\mathcal{P}_q)}(q,m)}{q^m}.$$

Therefore, the probability distribution $P(V_m^{(q)}=v)$ which maximizes $H_{\text{aux}}(q,m)$ is

$$P(V_m^{(q)}=v)=\begin{cases}1, & \text{if } v=\hat{v},\\ 0, & \text{if } v\neq\hat{v}.\end{cases}$$

For $q=2$ and Knuth's binary balancing method, this distribution is impossible for $m>2$ as can be deduced from (2), and so (21) in this case is a strict inequality.

Furthermore, for $q=2$, the upper bound from (21) becomes $\log_2(mB(m)/2^m)$, where $|\mathcal{P}|=m$. Then, using Stirling's formula

$$m!\sim\sqrt{2\pi m}\left(\frac{m}{e}\right)^m$$

to evaluate $B(m)=\binom{m}{m/2}$, it follows that

$$\log_2\left(\frac{B^{(\mathcal{P})}(m)}{2^m}\right)\sim\frac{1}{2}\log_2 m-\frac{1}{2}\log_2\frac{\pi}{2}. \tag{23}$$

Comparing the asymptotically exact expressions for $H_{\text{aux}}(m)$ from (6) and the upper bound of $H_{\text{aux}}(m)$ from (23), and noting that $\frac{1}{2}\log_2\frac{\pi}{2} \doteq 0.326 < 0.916 \doteq \frac{1}{2}(1+\gamma/\ln 2)$, we see that the asymptotic expression for the upper bound for $H_{\text{aux}}(m)$ is greater than the asymptotic expression for $H_{\text{aux}}(m)$ from (6).

More generally, by using (see [20, Lemma 2.3], [17, Thm. 3], [21, Eq. (46)])

$$B(q,m) \sim q^m \sqrt{\frac{6}{\pi m(q^2-1)}},$$

it follows that

$$\log_q\left(\frac{B^{(\mathcal{P}_q)}(q,m)}{q^m}\right) \sim \frac{1}{2}\log_q m - \frac{1}{2}\log_q\frac{\pi(q^2-1)}{6q^2}$$

for the scheme from [15] and

$$\log_q\left(\frac{B^{(\mathcal{P}_q)}(q,m)}{q^m}\right) \sim \frac{1}{2}\log_q m - \frac{1}{2}\log_q\frac{\pi(q+1)}{6(q-1)}$$

for the scheme from [16].

IV. Generalized Random Walk for Ternary Balanced Codes

Before considering a generalized random walk for the ternary case, we define a one-dimensional random walk for binary sequences and demonstrate its link with balanced sequences and the number of balancing indices. For the binary case, a random walk can be defined as a discrete stochastic process [23, p. 74]

$$S_m \triangleq \sum_{i=1}^{m} X_i, \qquad S_0 \triangleq 0, \tag{24}$$

where $P(X_i = 0) = P(X_i = 1) = 1/2$ and $m \geq 0$. We will call the index i of a random walk epoch i. Clearly, a binary sequence $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_2^m$ of length m corresponding to a random walk S_j, $1 \leq j \leq m$, where X_i corresponds to x_i for $1 \leq i \leq m$, is balanced if, and only if, $S_m = m/2$. As a support for the visual representation of random walks, we also introduce

$$S'_m \triangleq \sum_{i=1}^{m} X'_i, \qquad S'_0 \triangleq 0, \tag{25}$$

where $X'_i = -1$ if $X_i = 0$ and $X'_i = +1$ if $X_i = 1$. The sequence corresponding to the random walk S'_m is balanced if, and only if, $S'_m = 0$. Note that we use the same graphical representation for both random walks S_m and S'_m.

A plot of a random walk corresponding to some binary word allows for the application of a simple technique to determine the number of balancing indices for that word. We introduce

$$S'^{(k)}_m \triangleq \sum_{i=1}^{k} \bar{X}'_i + \sum_{i=k+1}^{m} X'_i, \qquad S'^{(k)}_0 \triangleq 0, \tag{26}$$

where $1 \leq k \leq m$ and $\bar{X}'_i \triangleq -X'_i$, to denote the random walk corresponding to the complementation process. As with Knuth's balancing method, complement the bits one at a time until an epoch k_1 is reached such that the altered word is balanced, i.e. $S'^{(k_1)}_m = 0$ (this means that the random walk consisting of the complement of the original random walk prior to epoch k_1 and the original random walk subsequent to epoch k_1 starts and ends at the same level, i.e. $S'^{(k_1)}_m = S'^{(k_1)}_0 = 0$). Suppose that at this epoch $S'_{k_1} = l$. If we continue with the complementation process after epoch k_1, suppose that the next balanced word is obtained at epoch $k_2 > k_1$, i.e. $S'^{(k_2)}_m = 0$. It is clear that this can only occur if the original random walk between epochs k_1 and k_2 consists of an equal number of the two binary symbols, so that $S'_{k_2} = l$. This is equally true for all subsequent balancing indices. Then, the number of balancing indices for that word is the number of times the random walk S'_j, $1 \leq j \leq m$, equals l, i.e. $|\{j : S'_j = l, 1 \leq j \leq m\}|$. The following example illustrates this.

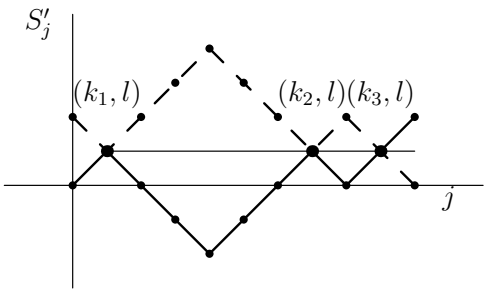

Fig. 1. Random walk of (1000111011) with three balancing indices at epochs $k_1 = 1, k_2 = 7, k_3 = 9$ (indicated by larger dots) and $l = 1$. The solid line represents the random walk of the original word, while the dashed line that of its complement. The random walk of the complement has been shifted to start at the level where the random walk of the original word ends.

Example 3: Consider the binary word (1000111011). The random walk of this word and its complement is depicted in Fig. 1. The segment ↗ denotes the symbol '1' (equivalently '+1' when determining S'_j), while ↘ denotes the symbol '0' (equivalently '−1' when determining S'_j). By complementing the bits as in Knuth's balancing procedure, we find that the smallest epoch at which a balanced word is attained is $k_1 = 1$ (as can be verified from Fig. 1 by noting that the word with the first bit complemented starts and ends at the same level, i.e. $S'^{(1)}_{10} = 0$). Note that $S'_1 = 1$. By continuing the complementation process, we find that the next epoch that results in a balanced word is $k_2 = 7$, i.e. $S'^{(7)}_{10} = 0$. This also occurs at level $l = 1$, i.e. $S'_7 = 1$. The last balanced word is at $k_3 = 9$ (i.e. $S'^{(9)}_{10} = 0$), $l = 1$ (i.e. $S'_9 = 1$). It is easy to see, that once the first balancing epoch k_1 is reached, the next balancing epoch k_2 is attained only if the complemented random walk (dashed line) returns to level $l = 1$. This is true for all subsequent balancing epochs. □

This observation is important, as the formula in (2) derived by Weber and Immink is based on the decomposition of a binary word of given length having v balancing indices into sub-words which are balanced, but whose random walks have no internal balancing indices/epochs and the remainder of the word having a single balancing index/epoch [4, p. 1678]. This is equivalent to the situation depicted in Fig. 1, where the balanced sub-words with no internal balancing indices/epochs are the sub-words from epoch k_1 to k_2 and from epoch k_2 to k_3.

In attempting to generalize the binary random walk to the ternary case, the aim is to retain the properties and characteristics described above. As we demonstrate hereafter, this can be achieved using a one-dimensional random walk of length $2m$. This generalization is based on the non-binary balancing scheme by Pelusi *et al.* [16], as this construction proved conducive to such a generalization. The idea is to decompose a ternary word in $\mathbb{Z}_3^m$ into two binary words of equal length so that the ternary word is the sum of the constituent binary words. The random walks of the binary constituent words are combined in a particular manner to yield the generalized random walk corresponding to the ternary word. Before formalizing these ideas, we give the following example as an introduction to the idea.

Example 4: Consider again the ternary word $(2102) \in \mathbb{Z}_3^4$ as in Example 1. This word is uniquely decomposed as $(2102) = (1101) + (1001)$. Then, we apply Knuth's complementation to the constituent binary words concurrently in the following manner (the values of k in brackets after k' denote the complementation stage of the balancing scheme by Pelusi *et al.* [16] from Example 1 which produce the same ternary word):

$$
\begin{aligned}
k' = 0\,(k = 0) &: \left\{ \begin{array}{c} (1101) \\ +(1001) \end{array} \right\} = (2102), \\
k' = 1\,(k = 2) &: \left\{ \begin{array}{c} (\underline{0}101) \\ +(1001) \\ \hline (1101) \\ +(\underline{0}001) \end{array} \right\} = (\mathbf{1102}), \\
k' = 2\,(k = 1) &: \left\{ \begin{array}{c} (\underline{0}101) \\ +(\underline{0}001) \end{array} \right\} = (0102), \\
k' = 3\,(k = 4) &: \left\{ \begin{array}{c} (\underline{00}01) \\ +(\underline{0}001) \end{array} \right\} = (0002), \\
k' = 4\,(k = 3) &: \left\{ \begin{array}{c} (\underline{0}101) \\ +(\underline{01}01) \end{array} \right\} = (\mathbf{0202}), \\
k' = 5\,(k = 5) &: \left\{ \begin{array}{c} (\underline{001}1) \\ +(\underline{01}01) \\ \hline (\underline{00}01) \\ +(\underline{011}1) \end{array} \right\} = (\mathbf{0112}), \\
k' = 6\,(k = 6) &: \left\{ \begin{array}{c} (\underline{001}1) \\ +(\underline{011}1) \end{array} \right\} = (0122), \\
k' = 7\,(k = 8) &: \left\{ \begin{array}{c} (\underline{00}10) \\ +(\underline{011}1) \\ \hline (\underline{001}1) \\ +(\underline{011}0) \end{array} \right\} = (\mathbf{0121}), \\
k' = 8\,(k = 7, 9) &: \left\{ \begin{array}{c} (\underline{00}10) \\ +(\underline{011}0) \end{array} \right\} = (0120).
\end{aligned}
$$

As usual, underlining denotes complemented symbols, while non-binary balanced words are represented in bold. Therefore, the word (2102) has four balancing indices ($k' = 1, 4, 5, 7$) under this complementation scheme. Notice that at stages $k' = 1, 5, 7$ (corresponding to the first complementation stage of symbols from $\{0, 2\}$), it is irrelevant in which of the two binary constituent words the next symbol is complemented first, as both options produce the same ternary symbol. For the symbol '1', as represented by stages $k' = 3, 4$, we complement the corresponding symbol in one constituent binary word and then, at the next stage, uncomplement the symbol from the previous stage and complement the corresponding symbol of the other binary constituent word. The order in which this is done is irrelevant, so stages $k' = 3, 4$ could be reversed. Finally, at stage $k' = 8$, where both constituent binary words have been fully complemented, we obtain (0120), which is the ternary complement of the original word. Here, $k' = 9$ is not needed (cf. $k = 9$ from Example 1) as the complement of the ternary symbol is obtained after two, and not three, stages. It is easily seen that both complementation schemes produce the same sequences, albeit in a different order. □

Based on the ideas espoused in the above example, we formalize the complementation process based on binary decomposition for each symbol in $\mathbb{Z}_3 = \{0, 1, 2\}$ as follows:

- 2:

$$
\begin{aligned}
&2 \to 1 \to 0 \\
&\equiv (1,1) \to (0,1) \to (0,0) \\
&\equiv \nearrow \to \curlyvee \to \searrow,
\end{aligned}
\tag{27}
$$

- 0:

$$
\begin{aligned}
&0 \to 1 \to 2 \\
&\equiv (0,0) \to (1,0) \to (1,1) \\
&\equiv \searrow \to \curlywedge \to \nearrow,
\end{aligned}
\tag{28}
$$

- 1:

$$
\begin{aligned}
&1 \to 2 \to 0 \to 1 \\
&\equiv (0,1) \to (1,1) \to (0,0) \to (1,0) \\
&\equiv \nwarrow \to \nearrow \to \searrow \to \nwarrow \\
&\equiv 1 \to 0 \to 2 \to 1 \\
&\equiv (1,0) \to (0,0) \to (1,1) \to (0,1) \\
&\equiv \nwarrow \to \searrow \to \nearrow \to \nwarrow.
\end{aligned}
\tag{29}
$$

Above, we have also depicted the random walk segments corresponding to the symbols. Each ternary symbol is composed of two bars each corresponding to a binary symbol. The definitions for $\{0, 2\}$ are self-explanatory, while that of $\{1\}$ shows a significant deviation in that it introduces discontinuity into the random walk. We deliberately introduce this discontinuity to signify the different behavior of $\{1\}$, which deviates from the other symbols $\{0, 2\}$ in requiring, after the complementation of the first binary symbol, the reversion to the original state, before the complementation of the second binary symbol.

Then, the random walk of a ternary word is defined as the concatenation of the random walk segments corresponding to symbols in $\mathbb{Z}_3$ from (27)–(29). Formally, we define the random walk over the ternary alphabet $\mathbb{Z}_3$ as a discrete stochastic process

$$
\begin{aligned}
S_{2m}^{(\mathbb{Z}_3)} &\triangleq \sum_{i=1}^{2m} X_{\lceil i/2 \rceil}^{(\mathbb{Z}_2, [(i+1) \bmod 2]+1)} \\
&= \sum_{i=1}^{m} X_i^{(\mathbb{Z}_3)}, \qquad S_0^{(\mathbb{Z}_3)} \triangleq 0,
\end{aligned}
\tag{30}
$$

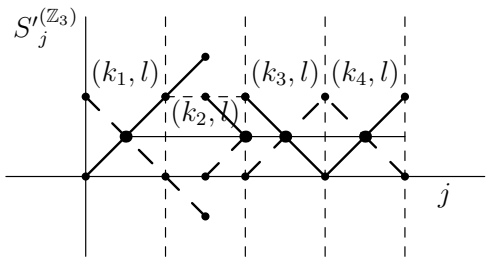

Fig. 2. Ternary random walk of (2102) with four balancing indices at epochs $k_1 = 1, k_2 = 4, k_3 = 5, k_4 = 7$ (indicated by larger dots) and $l = 1$. The solid line represents the random walk of the original word, while the dashed line that of its complement. The random walk of the complement has been shifted to start at the level where the random walk of the original word ends. The vertical dashed lines mark the ternary symbol boundaries.

based on the random variable $X_i^{(\mathbb{Z}_3)}$, $1 \leq i \leq m$, where $P(X_i^{(\mathbb{Z}_3)} = 0) = P(X_i^{(\mathbb{Z}_3)} = 1) = P(X_i^{(\mathbb{Z}_3)} = 2) = 1/3$, $X_i^{(\mathbb{Z}_3)}$ is decomposed such that $X_i^{(\mathbb{Z}_3)} = X_i^{(\mathbb{Z}_2,1)} + X_i^{(\mathbb{Z}_2,2)}$ and $\lceil x \rceil$ is the least integer greater than or equal to x. Clearly, a ternary sequence $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_3^m$ of length m corresponding to a random walk $S_j^{(\mathbb{Z}_3)}$, $1 \leq j \leq 2m$, where $X_i^{(\mathbb{Z}_3)}$ corresponds to x_i for $1 \leq i \leq m$, is balanced if, and only if, $S_{2m}^{(\mathbb{Z}_3)} = m$. As in the binary case, we introduce $S'^{(\mathbb{Z}_3)}_{2m}$, which is defined as in (30) with the replacement of $X_i^{(\mathbb{Z}_2,l)}$ by $X_i'^{(\mathbb{Z}_2,l)}$, $l = 1, 2$, where $X_i'^{(\mathbb{Z}_2,l)} = -1$ if $X_i^{(\mathbb{Z}_2,l)} = 0$ and $X_i'^{(\mathbb{Z}_2,l)} = +1$ if $X_i^{(\mathbb{Z}_2,l)} = 1$.

The ternary random walk of (2102) from Example 4 and its complement are depicted in Fig. 2. Four balancing indices can be identified: $k_1 = 1, k_2 = 4, k_3 = 5, k_4 = 7$ corresponding to $k' = 1, 4, 5, 7$ from Example 4. The complementation stages, denoted by k', from Example 4, are equivalent to complementing each random walk segment from Fig. 2 one at a time. Thus, for $k' = 1$, complementing the first segment results in $1 = 0 + 1$ as the first symbol, giving (1102). For $k' = 2$, complementing the first two segments results in $0 = 0 + 0$ as the first symbol, giving (0102). For $k' = 3$, we move to the next symbol. Complementing the first segment results in $0 = 0 + 0$ as the second symbol, giving (0002). For $k' = 4$, since it corresponds to the symbol '1', we have to uncomplement the previous segment before complementing the next segment, resulting in $2 = 1 + 1$ as the second symbol, giving (0202). The same process is continued for the last two ternary symbols. It should be obvious that this representation of the complementation process for the ternary case shows close affinity with Knuth's binary balancing process. Furthermore, note the parallels of the ternary random walk with the binary random walk (cf. Fig. 1), in particular, the manner in which the number of balancing indices can be identified by locating the first balancing epoch k_1 and thereafter, determining the number of times the random walk of the original word (solid line) intersects the random walk of its appropriately shifted complement (dashed line).

We would be amiss if we did not make explicit the following subtle point. In the binary case, knowledge of the balancing index k, $1 \leq k \leq m$, which indicates the number of initial bits complemented, is sufficient to decode the original user word. It may, therefore, seem logical that with the generalized ternary random walk, knowledge of the balancing index k, $0 \leq k \leq 2m + 1$, is sufficient to decode the original ternary user word by complementing the first k values of the ternary random walk corresponding to the received word. However, this is *not* true. The reason is that the complementation procedure is different for symbols $\{0, 2\}$ with respect to that of $\{1\}$ [cf. (27)–(29)], so that in order to properly decode the received word (i.e., to reverse the complementation procedure), it is necessary to know whether the corresponding original symbol belonged to $\{0, 2\}$ or $\{1\}$. This information cannot be extracted from the received word, so that unique decoding is not possible. However, this is not of any consequence for our purposes, as we do not need an invertible complementation scheme, but rather a complementation scheme which is equivalent to that of Pelusi *et al.* [16] in the sense that the words resulting from the complementation process are the same as those resulting from the application of $\boldsymbol{\beta}_3$ from (19), irrespective of the order of such words. If this is the case, then the number of balancing indices is the same under both complementation schemes for any user word. Since we wish to determine the amount of auxiliary data, which is dependent on the probability distribution of the number of balancing indices, this is of consequence for our purposes.

It is readily apparent that any ternary word $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_3^m$ can be decomposed into two binary (not necessarily unique) words

$$\begin{aligned} \mathbf{y}^{(1)} &= (y_1^{(1)}, y_2^{(1)}, \ldots, y_m^{(1)}), \\ \mathbf{y}^{(2)} &= (y_1^{(2)}, y_2^{(2)}, \ldots, y_m^{(2)}), \end{aligned} \tag{31}$$

$\mathbf{y}^{(1)}, \mathbf{y}^{(2)} \in \mathbb{Z}_2^m$, as

$$x_i = \left| \left\{ j : y_i^{(j)} = 1, 1 \leq j \leq 2 \right\} \right|, \; 1 \leq i \leq m, \tag{32}$$

so that

$$\mathbf{x} = \mathbf{y}^{(1)} + \mathbf{y}^{(2)}.$$

Without loss of generality, we assume that $y_i^{(j)} = 1$ for $j = 1, \ldots, x_i$, implying that $y_i^{(j)} = 0$ for $j = x_i + 1, , \ldots, 2$. Then, we define the following complementation function

$$\hat{\phi}(x_i, j) \triangleq \begin{cases} \hat{\phi}^-(x_i, j), & \text{if } 1 \leq j \leq x_i, \\ \hat{\phi}^+(x_i, j), & \text{if } x_i + 1 \leq j \leq 2, \end{cases} \tag{33}$$

where

$$\hat{\phi}^-(x_i, j) \triangleq \sum_{l=1}^{j} \bar{y}_i^{(l)} + \sum_{l=j+1}^{2} y_i^{(l)}, \tag{34}$$

and

$$\hat{\phi}^+(x_i, j) \triangleq \sum_{l=1}^{x_i} y_i^{(l)} + \sum_{l=x_i+1}^{j} \bar{y}_i^{(l)} + \sum_{l=j+1}^{2} y_i^{(l)}, \tag{35}$$

for $1 \leq i \leq m$ and $1 \leq j \leq 2$. The reader should convince themselves that the complementation procession as defined by $\hat{\phi}$ is equivalent to that defined in (27)–(29). This approach can be easily extended to the general non-binary case where $q > 3$.

By analogy with $\boldsymbol{\beta}_q(\mathbf{x}, k)$ from (19) we define

$$\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k) = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{i-1}, \hat{\phi}(x_i, j), x_{i+1}, \ldots, x_{m-1}, x_m), \tag{36}$$

where $0 \leq k \leq 2m+1$ and $i(k) \triangleq \lceil k/2 \rceil$, $j(k) \triangleq [(k-1) \bmod 2] + 1$ as in [16, Eq. (12)].

Example 5: Consider again the ternary word $\mathbf{x} = (2102) \in \mathbb{Z}_3^4$ from Examples 1 and 4. According to (33),

$$\begin{pmatrix} \hat{\phi}(0,0) & \hat{\phi}(0,1) & \hat{\phi}(0,2) \\ \hat{\phi}(1,0) & \hat{\phi}(1,1) & \hat{\phi}(1,2) \\ \hat{\phi}(2,0) & \hat{\phi}(2,1) & \hat{\phi}(2,2) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix},$$

and therefore (the values of k in square brackets after k' denote the complementation stage of the balancing scheme by Pelusi *et al.* [16] from Example 1 which produce the same ternary word):

$$\begin{aligned}
k' = 0\,(i=0, j=2)\,[k=0]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = (2102), \\
k' = 1\,(i=1, j=1)\,[k=2]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 1) = (\mathbf{1102}), \\
k' = 2\,(i=1, j=2)\,[k=1]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 2) = (0102), \\
k' = 3\,(i=2, j=1)\,[k=4]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 3) = (0002), \\
k' = 4\,(i=2, j=2)\,[k=3]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 4) = (\mathbf{0202}), \\
k' = 5\,(i=3, j=1)\,[k=5]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 5) = (\mathbf{0112}), \\
k' = 6\,(i=3, j=2)\,[k=6]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 6) = (0122), \\
k' = 7\,(i=4, j=1)\,[k=8]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 7) = (\mathbf{0121}), \\
k' = 8\,(i=4, j=2)\,[k=7,9]: &\quad \hat{\boldsymbol{\beta}}_3(\mathbf{x}, 8) = (0120).
\end{aligned}$$

As in Example 4, balanced words, which are shown in bold, are obtained for $k' = 1, 4, 5, 7$. □

V. Auxiliary Data for Ternary Case

In this section, we derive an asymptotically exact expression for the amount of auxiliary data for a variable length modified version of the ternary balancing scheme by Pelusi *et al.* [16]. This is achieved by modifying the ternary random walk from the previous section and recycling known results regarding auxiliary data for the binary alphabet. This result shows that in this scenario, as in the binary case, it is possible to regain almost all of the redundancy loss (as compared to optimal redundancy) of the balancing scheme without auxiliary data.

We start with the modification of the ternary random walk. We have seen in the previous section that $1 \in \mathbb{Z}_3$ leads to discontinuity in the ternary random walk. If $x_i = 1$, $1 \leq i \leq m$, then $S'^{(\mathbb{Z}_3)}_{2i-2} = S'^{(\mathbb{Z}_3)}_{2i}$ for the original word. Since '1' is its own complement, the same logic also applies to the random walk of the complemented word. This means that all symbols '1' can be removed from the ternary random walk without affecting the rest of the random walk for the original and complemented word, meaning that the validity of the complementation process for balancing purposes is retained. Stated alternatively, we can ignore the symbol '1' during the complementation process and only apply the complementation process to balance the sub-word consisting of symbols from $\{0, 2\}$. In particular, consider any $\mathbf{x} \in \{0, 2\}^{m'}$, then, since

$$\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = \mathbf{x}) = 2m' - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 2m') = \bar{\mathbf{x}})$$

and

$$\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k+1)) - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) = \pm 1$$

as $\hat{\phi}(x_i, 2) - \hat{\phi}(x_i, 1) = \bar{x}_i - 1 = \pm 1$ and $\bar{x}_i + \hat{\phi}(x_{i+1}, 1) - \hat{\phi}(x_i, 2) - x_{i+1} = -x_{i+1} + 1 = \pm 1$, it follows that there exists at least one k, $0 \leq k \leq 2m'$, such that $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) = m'$. Then, for any ternary word $\mathbf{x} \in \mathbb{Z}_3^m$ of length m, if $\mathbf{x}_{\{0,2\}}$ denotes the sub-word of $\mathbf{x}$ of length m' consisting only of symbols from $\{0, 2\}$ and $\mathbf{x}_{\{1\}}$ the sub-word of $\mathbf{x}$ of length $m - m'$ consisting of only the symbols $\{1\}$, it follows that for at least one k, $0 \leq k \leq 2m'$, $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}_{\{0,2\}}, k)) + \sigma(\mathbf{x}_{\{1\}}) = m' + m - m' = m$. Therefore any ternary word can be balanced by only balancing the sub-word consisting of symbols from $\{0, 2\}$.

The corresponding random walk as a discrete stochastic process is denoted by $S^{(\{0,2\})}_{2m'}$ and is defined as in (30) with the replacement of $X_i^{(\mathbb{Z}_3)}$ by $X_i^{(\{0,2\})}$. Similarly for $S'^{(\{0,2\})}_{2m'}$.

Lemma 2: For any $\mathbf{x} \in \{0, 2\}^{m'}$, $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)$ is balanced, i.e. $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) = m'$, for even m' only if k is even, and for odd m' only if k is odd.

Proof: For any $\mathbf{x} \in \{0, 2\}^{m'}$,

$$\sigma(\mathbf{x}) \in \Big\{2i : i = 0, 1, \ldots, m' - 1, m'\Big\}. \tag{37}$$

Hence $\sigma(\mathbf{x})$ is even, irrespective of whether m' is even or odd. Since $\hat{\phi}(x_i, 2) - \hat{\phi}(x_i, 1) = \bar{x}_i - 1 = \pm 1$, $1 \leq i \leq m'$, and $\bar{x}_i + \hat{\phi}(x_{i+1}, 1) - \hat{\phi}(x_i, 2) - x_{i+1} = -x_{i+1} + 1 = \pm 1$, $1 \leq i \leq m' - 1$, it follows that

$$\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k-1)) = \pm 1, \tag{38}$$

for $1 \leq k \leq 2m'$. By repeated application of (38), it follows that $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = \mathbf{x})$ is even only if k is even and odd only if k is odd.

When $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)$ is balanced, $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) = m'$. If m' is even, then $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = \mathbf{x})$ is even since $\sigma(\mathbf{x})$ is even. Therefore, when m' is even, $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)$ is balanced only if k is even. Similarly, if m' is odd, then $\sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)) - \sigma(\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = \mathbf{x})$ is odd. Hence, when m' is odd, $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, k)$ is balanced only if k is odd. ■

The above lemma demonstrates that for $\mathbf{x} \in \{0, 2\}^m$, m even, the balancing index k can only occur at the symbol boundaries. This is exemplified in Fig. 3, where the random walk and balancing indices for $(2202) \in \{0, 2\}^4$ are illustrated. Here, $\hat{\boldsymbol{\beta}}_3((2202), 2) = (0202)$ and $\hat{\boldsymbol{\beta}}_3((2202), 6) = (0022)$. Then, if we replaced each $x_i = 0 \in \mathbb{Z}_3$ by $x'_i = 0 \in \mathbb{Z}_2$ (↘ by ↘) and each $x_i = 2 \in \mathbb{Z}_3$ by $x'_i = 1 \in \mathbb{Z}_2$ (↗ by ↗), the random walks of $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \{0, 2\}^m$ and $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_m) \in \mathbb{Z}_2^m$ are equivalent. In particular, if $\mathbf{x}$ has v balancing indices at $k_1, k_2, \ldots, k_v$, then $\mathbf{x}'$ has v balancing indices at $k_1/2, k_2/2, \ldots, k_v/2$. We denote this equivalence by $S^{(\{0,2\})}_{2m} \equiv S^{(\mathbb{Z}_2)}_m$. If we extend the notation $V^{(q)}_m$ to $V^{(\{0,2\})}_{2m}$ to denote the number of balancing indices for the random walk $S^{(\{0,2\})}_{2m}$, then it follows that

$$P(V^{(\{0,2\})}_{2m} = v) = P(V_m = v), \tag{39}$$

for $1 \leq v \leq m/2$.

It is assumed that all ternary user words are equiprobable and independent so that $P(x_i = 0) = P(x_i = 1) = P(x_i = 2) = 1/3$. Consider the balancing procedure for $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{Z}_3^m$ which applies $\hat{\phi}(x_i, j)$ from (33)

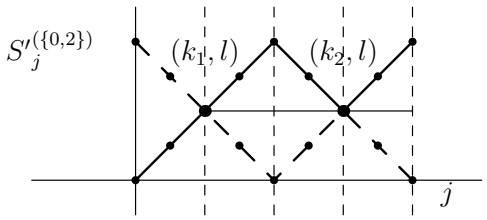

Fig. 3. Random walk of (2202) with two balancing indices at epochs $k_1 = 2, k_2 = 6$ (indicated by larger dots) and $l = 2$. The solid line represents the random walk of the original word, while the dashed line that of its complement. The random walk of the complement has been shifted to start at the level where the random walk of the original word ends. The vertical dashed lines mark the ternary symbol boundaries.

only to symbols $x_i \in \{0, 2\}$, i.e. the balancing procedure is only applied to the sub-word $\mathbf{x}_{\{0,2\}}$. This procedure produces the same balanced words as those by the balancing scheme of Pelusi *et al.* [16] for any user word, except for those words that result during the complementation of $\{1\}$. This means that such a balancing procedure, which results in the same or fewer number of balancing indices for all ternary user words, provides less auxiliary data on average. On the other hand, since this procedure does not include $\{1\}$ during the complementation process, its range for balancing indices k, which is user word dependent, is equal or smaller, allowing for shorter average balanced suffix lengths. The gain in suffix length is offset by the reduction in the amount of auxiliary data.

This balancing approach, which is of fixed length, can be reformulated equivalently as the following variable length scheme. Consider a theoretically infinite sequence of ternary symbols $x_1, x_2, \ldots, x_i, \ldots$, where $x_i \in \mathbb{Z}_3$. One may also consider this sequence as a concatenation of an infinite number of fixed length words $\mathbf{x} \in \mathbb{Z}_3^{m''}$. Then, partition this sequence into variable length words such that each word consists of m' symbols $x_i \in \{0, 2\}$, where m' is even. The average length is denoted by m. The effect of this formulation is that the number of possible balancing indices k is fixed per variable length word. We denote by $\hat{H}_{\text{aux}}(3, m)$ the achievable amount of auxiliary data for such a variable length scheme.

Note that, since $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 0) = \mathbf{x}$ and $\hat{\boldsymbol{\beta}}_3(\mathbf{x}, 2m') = \bar{\mathbf{x}}$ for any $\mathbf{x} \in \{0, 2\}^{m'}$ and $\mathbf{x}$ is balanced if, and only if, $\bar{\mathbf{x}}$ is balanced, the range of k can be adjusted to $1 \leq k \leq 2m'$ without affecting the validity of the balancing procedure.

Example 6: Consider some sequence (possibly of infinite length)

$$\mathbf{x} = (2210121110212001020\ldots).$$

Assume that $m' = 4$ and partition $\mathbf{x}$ into variable length segments, such that each segment contains exactly $m' = 4$ symbols from $\{0, 2\}$. The partitioning stops as soon as $m' = 4$ symbols from $\{0, 2\}$ are obtained. If such segments are denoted as $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$, then

$$\begin{aligned}\mathbf{x}_1 &= (221012),\\ \mathbf{x}_2 &= (11102120),\\ \mathbf{x}_3 &= (01020).\end{aligned}$$

Note that $\mathbf{x}_1$ is partitioned as (221012) and not (221012111) and $\mathbf{x}_2$ as (11102120) and not (02120). To encode this sequence, apply the complementation function $\hat{\boldsymbol{\beta}}_3$ to the sub-words $\mathbf{x}_{1,\{0,2\}} = (2202)$, $\mathbf{x}_{2,\{0,2\}} = (0220)$ and $\mathbf{x}_{3,\{0,2\}} = (0020)$ of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, respectively. Then, balanced words are obtained as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_3(\mathbf{x}_{1,\{0,2\}}, 2) &= (0202),\\ \hat{\boldsymbol{\beta}}_3(\mathbf{x}_{2,\{0,2\}}, 8) &= (2002),\\ \hat{\boldsymbol{\beta}}_3(\mathbf{x}_{3,\{0,2\}}, 2) &= (2020).\end{aligned}$$

Since $m' = 4$, balancing indices can only occur at $k = 2, 4, 6, 8$, so that 4 balanced suffixes are needed. The smallest suffix length at which this can be achieved is $p = 4$. Assume that $\boldsymbol{\psi}_3(2) = \mathbf{p}_2 = (1111)$ and $\boldsymbol{\psi}_3(8) = \mathbf{p}_8 = (0211)$. Let $\mathbf{x}'_1$ represent the merger of $\hat{\boldsymbol{\beta}}_3(\mathbf{x}_{1,\{0,2\}}, 2)$ with $\mathbf{x}_{1,\{1\}}$ where the symbols '1' retain the same positions as in $\mathbf{x}_1$. Similarly for $\mathbf{x}'_2$ and $\mathbf{x}'_3$. Then, the encoded sequence is

$$\begin{aligned}&(\mathbf{x}'_1, \boldsymbol{\psi}_3(2), \mathbf{x}'_2, \boldsymbol{\psi}_3(8), \mathbf{x}'_3, \boldsymbol{\psi}_3(2), \ldots)\\ =&(021012\,1111\,11120102\,0211\,21020\,1111\ldots) = \mathbf{y}.\end{aligned}$$

To decode $\mathbf{y}$, partition it into variable length segments, such that each segment contains exactly $m' = 4$ symbols from $\{0, 2\}$. Again, as with encoding, the partitioning stops as soon as $m' = 4$ symbols from $\{0, 2\}$ are obtained. If such segments are denoted as $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots$, and noting that each segment is followed by a suffix of length $p = 4$, then

$$\begin{aligned}\mathbf{y}_1 &= (021012),\\ \mathbf{y}_2 &= (11120102),\\ \mathbf{y}_3 &= (21020).\end{aligned}$$

$\mathbf{y}_1$ is followed by the suffix (1111), $\mathbf{y}_2$ by (0211) and $\mathbf{y}_3$ by (1111). Since $\boldsymbol{\psi}_3^{-1}(1111) = 2$ and $\boldsymbol{\psi}_3^{-1}(0211) = 8$, to decode the sequence, apply the inverse complementation function $\hat{\boldsymbol{\beta}}_3^{-1}$ to the sub-words $\mathbf{y}_{1,\{0,2\}} = (0202)$, $\mathbf{y}_{2,\{0,2\}} = (2002)$ and $\mathbf{y}_{3,\{0,2\}} = (2020)$ of $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$, respectively, to obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_3^{-1}(\mathbf{y}_{1,\{0,2\}}, 2) &= (2202),\\ \hat{\boldsymbol{\beta}}_3^{-1}(\mathbf{y}_{2,\{0,2\}}, 8) &= (0220),\\ \hat{\boldsymbol{\beta}}_3^{-1}(\mathbf{y}_{3,\{0,2\}}, 2) &= (0020).\end{aligned} \tag{40}$$

By merging $\mathbf{y}_{1,\{1\}}$, $\mathbf{y}_{2,\{1\}}$ and $\mathbf{y}_{3,\{1\}}$ with the corresponding decoded words from (40), the decoded sequence is

$$(2210121110212001020\ldots).$$

Clearly, although the balanced words are variable length, unique decoding is achievable. Also, as the length of $\mathbf{x}$ tends to infinity, $m \to m' \cdot 3/2 = 6$, where m is the average user word length. □

Let $p^{(\{0,2\})}$ denote the required suffix length for the case where $\{1\}$ is excluded from the complementation process, so that the complementation procedure is applied only to a sub-word, in $\{0, 2\}^{m'}$, of the variable length word. Then, it follows that

$$m' = B(3, p^{(\{0,2\})}),$$

so that by [20, Lemma 2.3]

$$m' = 3^{p^{(\{0,2\})}} \sqrt{\frac{3}{4\pi p^{(\{0,2\})}}} \left(1 + O\Big(\frac{1}{p^{(\{0,2\})}}\Big)\right),$$

and hence

$$\log_3(m') = p^{(\{0,2\})} - \frac{1}{2}\log_3 p^{(\{0,2\})} - \frac{1}{2}\log_3 \frac{4\pi}{3} + \log_3\left(1 + O\Big(\frac{1}{p^{(\{0,2\})}}\Big)\right). \quad (41)$$

From the above equation, it follows that $p^{(\{0,2\})} = O(\log_3 m')$. Dividing both sides of (41) by $\log_3 m'$ and taking the limit as $m' \to \infty$, it can be seen that the limit of the last three terms on the right-hand side are all zero (the first by L'Hôpital's rule). Therefore,

$$p^{(\{0,2\})} \sim \log_3 m'. \quad (42)$$

Theorem 2:

$$\hat{H}_{\text{aux}}(3,m) \sim \frac{1}{2}\log_3 m - \frac{1}{2}\left(1 + \frac{\gamma}{\ln 3}\right). \quad (43)$$

Proof:

$$\begin{aligned}
\hat{H}_{\text{aux}}(3,m) &= \sum_{v=1}^{m'/2} P(V_{2m'}^{(\{0,2\})} = v)\log_3 v \\
&= \sum_{v=1}^{m'/2} P(V_{m'} = v)\log_3 v \quad (44)\\
&= \frac{1}{\log_2 3}\sum_{v=1}^{m'/2} P(V_{m'} = v)\log_2 v \\
&= \frac{1}{\log_2 3} H_{\text{aux}}(m') \\
&\sim \frac{1}{\log_2 3}\left[\frac{1}{2}\log_2 m' - \frac{1}{2}\Big(1 + \frac{\gamma}{\ln 2}\Big)\right] \quad (45)\\
&= \frac{1}{2}\log_3 m' - \frac{1}{2\log_2 3}\Big(1 + \frac{\gamma}{\ln 2}\Big), \\
&\sim \frac{1}{2}\log_3 \frac{2}{3}m - \frac{1}{2\log_2 3}\Big(1 + \frac{\gamma}{\ln 2}\Big) \quad (46)\\
&= \frac{1}{2}\log_3 m - \frac{1}{2}\Big(1 + \frac{\gamma}{\ln 3}\Big),
\end{aligned}$$

where (44) follows from (39), (45) from (6) and (46) from the fact that $m/m' \to 3/2$ as $m, m' \to \infty$. ∎

This result shows unambiguously that auxiliary data, albeit in a variable length context, is able to recover almost all of the redundancy loss for the ternary case. Furthermore, combining (42) and (43), we obtain the approximation

$$\begin{aligned}
p^{(\{0,2\})} - \hat{H}_{\text{aux}}(3,m) &\approx \log_3 \frac{2}{3}m - \frac{1}{2}\log_3 m + \frac{1}{2}\Big(1 + \frac{\gamma}{\ln 3}\Big) \\
&= \frac{1}{2}\log_3 m + \frac{1}{2}\left(\frac{2\ln 2 + \gamma - \ln 3}{\ln 3}\right). \quad (47)
\end{aligned}$$

An interesting observation can be made by using (5) and noting that (45) is equal to

$$\sqrt{\frac{2}{\pi m'}}\int_{v=0}^{\infty} e^{\frac{-v^2}{2m'}}\log_3 v\,\mathrm{d}v.$$

This quantity is the asymptotically exact expression for the amount of auxiliary data when the complementation process excludes $\{1\}$. Setting $m' = (3/2)m$ and using [18, Eq. (4.333), p. 575] to evaluate the improper integral produces

$$\frac{2}{\sqrt{3\pi m}}\int_{v=0}^{\infty} e^{\frac{-v^2}{3m}}\log_3 v\,\mathrm{d}v = \frac{1}{2}\log_3 m - \frac{1}{2}\left(\frac{2\ln 2 + \gamma - \ln 3}{\ln 3}\right),$$

which is equal to (47) subtracted from $\log_3 m$. This implies that

$$\frac{V_m^{(3)}}{\sqrt{3m/2}} \xrightarrow{d} Y,$$

where $Y \triangleq |Z|$, with $|Z|$ representing the absolute value of Z, for standard normally distributed $Z \sim \mathcal{N}(0,1)$. This leads to the following conjecture.

Conjecture 1: For $q \geq 2$,

$$\frac{V_m^{(q)}}{\sqrt{qm/2}} \xrightarrow{d} Y.$$

The above conjecture can be stated equivalently as

$$P(V_m^{(q)} \geq y\sqrt{qm/2}) \to 2[1 - \Phi(y)].$$

If this conjecture is true, using a similar argument as in the binary case leads to

$$\begin{aligned}
H_{\text{aux}}(q,m) &\sim \sqrt{\frac{2}{\pi}}\int_{v=0}^{\infty} e^{\frac{-v^2}{2}}\log_q\left(v\sqrt{qm/2}\right)\mathrm{d}v \\
&= \frac{2}{\sqrt{\pi q m}}\int_{u=0}^{\infty} e^{\frac{-u^2}{qm}}\log_q u\,\mathrm{d}u \\
&= \frac{1}{2}\log_q m - \frac{1}{2}\left(\frac{2\ln 2 + \gamma - \ln q}{\ln q}\right), \quad (48)
\end{aligned}$$

where the first equality is obtained by a change of variable $u \triangleq v\sqrt{qm/2}$ and the second follows from [18, Eq. (4.333), p. 575].

Our attempts to prove this conjecture for $q > 2$ by direct means for the fixed length case have been futile thus far. Nevertheless, it is supported by our simulation results (see below), by the conjecture of Pelusi *et al.* [16, §IV] that $p^{(\mathbb{Z}_q)} - H_{\text{aux}}(q,m) = \frac{1}{2}\log_q m + \Theta(\log\log m)$, where $p^{(\mathbb{Z}_q)}$ is the redundancy without auxiliary data, for their balancing construction based on simulations with 10 million samples, and the fact that (48) reduces to (6) for $q = 2$ and the difference between $\log_3 m$ and (47) for $q = 3$.

Following Conjecture 1, $P(V_m^{(q)} = v)$ can be approximated by

$$f_m^{(q)}(v) \triangleq \frac{2}{\sqrt{\pi q m}} e^{\frac{-v^2}{qm}},$$

for finite m. As a means of evaluating the conjecture, we have obtained by simulation the probabilities $P(V_m^{(q)} = v)$ for the code by Pelusi *et al.* [16] by generating all q^m possible user words and determining the number of balancing indices after the application of $\boldsymbol{\beta}_q$ [cf. (15) and (17)]. Fig. 4–Fig. 7 depict the comparison between $P(V_m^{(q)} = v)$ and $f_m^{(q)}(v)$ for $q = 3$ ($m = 16, 20$) and $q = 4$ ($m = 12, 16$). For $q = 3$ and $q = 4$,

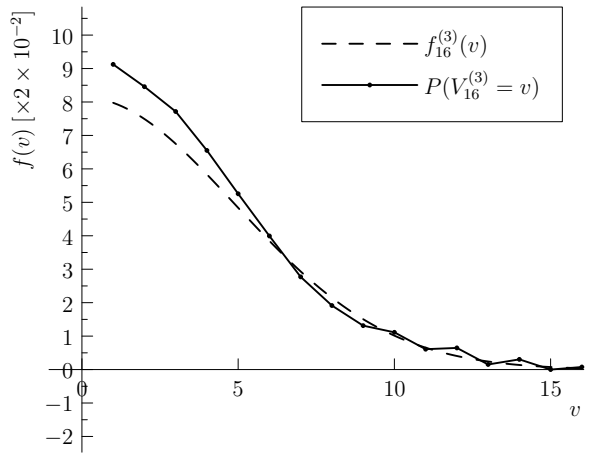

Fig. 4. Plot comparing $P(V_{16}^{(3)} = v)$ and $f_{16}^{(3)}(v)$.

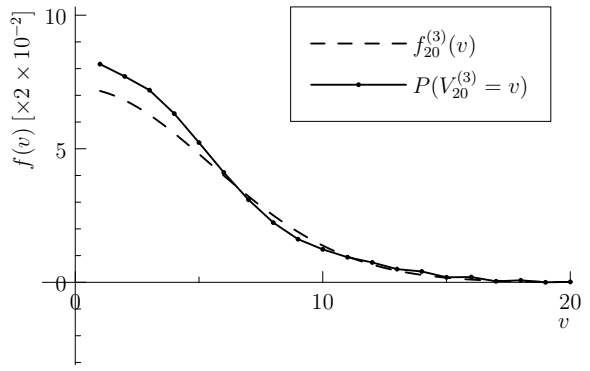

Fig. 5. Plot comparing $P(V_{20}^{(3)} = v)$ and $f_{20}^{(3)}(v)$.

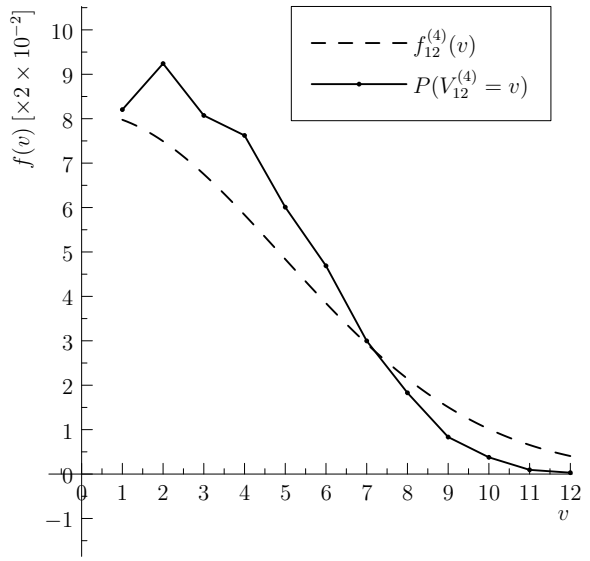

Fig. 6. Plot comparing $P(V_{12}^{(4)} = v)$ and $f_{12}^{(4)}(v)$.

ϕ from [16, Eq. (8)] and [16, Eq. (14)] was used, respectively. These plots show an indication of convergence of $P(V_m^{(q)} = v)$ to $f_m^{(q)}(v)$ with increasing m.

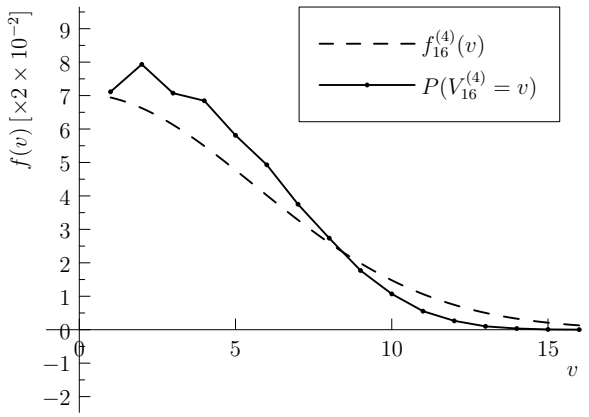

Fig. 7. Plot comparing $P(V_{16}^{(4)} = v)$ and $f_{16}^{(4)}(v)$.

VI. Conclusion

We have investigated various topics related to auxiliary data for non-binary balanced codes. Previous results on this topic have been meager to non-existent. First, we derived an upper bound on the average amount of auxiliary data for non-binary balanced codes and demonstrated that known non-binary generalizations of Knuth's balancing method satisfy this bound. The power of this result lies in its generality and its ability to explain why auxiliary data as applied to Knuth-like balancing schemes, binary and non-binary alike, is able to recover nearly all of the redundancy loss as compared to that of the full balanced set.

The main result is the derivation of an asymptotically exact expression of the average amount of auxiliary data for a variable length manifestation of a modified ternary balanced code by Pelusi *et al.* [16]. The derivation is based on a generalized ternary random walk, which allows for the definition of a complementation process that is dependent on a decomposition of a ternary word to a pair of binary words, and so is analogous to Knuth's binary balancing scheme. A modification of the general complementation procedure is proposed that simplifies various facets of the problem, leading to the main result. This result is in all respects the ternary counterpart to the binary result achieved by Weber and Immink [4].

Finally, we conjecture the convergence in distribution for the probability distribution of the number of balancing indices for the non-binary, $q \geq 3$, balancing construction by Pelusi *et al.* [16]. Based on this conjecture, an asymptotically exact expression is obtained for $H_{\text{aux}}(q, m)$ which reduces to the proven expression when $q = 2$ and to the derived approximation when $q = 3$. While this conjecture is supported by simulation results, a formal proof is still an open problem.

We hope that our contribution to this understudied topic may lead to a comprehensive solution of auxiliary data for non-binary balanced codes.

APPENDIX
ASYMPTOTIC PROBABILITY DISTRIBUTION OF THE NUMBER OF BALANCING INDICES FOR KNUTH'S BINARY METHOD

For the sake of completeness, here we give a formal proof that the probability distribution of the number of balancing indices using Knuth's binary balancing method converges in distribution to a half-normal distribution. This result is used during the proof of Theorem 2. As shown in Section IV, there exists an intimate relationship between the number of balancing indices and the number of returns to the origin of a one-dimensional binary random walk. We exploit this relationship to prove the result.

The binary random walk as a discrete stochastic process S'_m is defined in (25). Let the random variable R_m, m even, denote the number of times that a random walk S'_m returns to the origin, i.e. $|\{j : S'_j = 0, 1 \leq j \leq m\}|$. Then, it is known that [23, p. 96, Prob. 9 & 10]

$$P(R_m = r) = 2^{r-m}\binom{m-r}{m/2}. \tag{49}$$

Consider a sequence of random variables X_m, $m = 1, 2, \ldots$, each with a cumulative distribution function $F_m(x)$. Similarly, the random variable X has a cumulative distribution function $F(x)$. Then, formally, X_m converges in distribution to X ($X_m \xrightarrow{d} X$), if $\lim_{m\to\infty} F_m(x) = F(x)$ for each continuity point x of $F(x)$ [24, p. 8].

To determine the convergence in distribution of the number of balancing indices for Knuth's binary balancing method, we will use known results regarding the convergence in distribution of the number of returns to origin of a random walk. It is known that $R_m/\sqrt{m} \xrightarrow{d} Y$ as $m \to \infty$, where $Y \triangleq |Z|$, with $|Z|$ representing the absolute value of Z, for standard normally distributed $Z \sim \mathcal{N}(0, 1)$ [25, Thm. 7 & Eq. (5.31)] (cf. also [26, Eq. (1)]).

Using (2) and (49), a trite computation shows that

$$\frac{P(R_m = v)}{P(V_m = v)} = 1 - \frac{v}{m}. \tag{50}$$

Furthermore, $R_m/\sqrt{m} \xrightarrow{d} Y$ can be stated equivalently as [25, Eq. (5.31)]

$$P(R_m \geq y\sqrt{m}) \to 2[1 - \Phi(y)],$$

where $y > 0$ and $\Phi(y)$ is standard normal cumulative distribution function. Then, since $P(R_m \geq y\sqrt{m}) = 1 - P(R_m < y\sqrt{m})$, it follows that

$$P(R_m < y\sqrt{m}) \to 2\Phi(y) - 1.$$

This implies that convergence is valid where $v = O(\sqrt{m})$ and y is finite. If $v = O(\sqrt{m})$, from (50), it follows that the ratio of $P(R_m = v)$ to $P(V_m = v)$ tends to 1 as $m \to \infty$. Since a cumulative distribution function for a discrete random variable is fully characterized (including at all continuity points) by its discrete probabilities, it follows that convergence of discrete probabilities implies convergence in distribution. Therefore,

$$P(V_m < y\sqrt{m}) \to 2\Phi(y) - 1,$$

or, equivalently,

$$P(V_m \geq y\sqrt{m}) \to 2[1 - \Phi(y)],$$

and hence $V_m/\sqrt{m} \xrightarrow{d} Y$.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation (NRF) and SENTECH Chair in Broadband Wireless Multimedia Communication.

REFERENCES

[1] K. A. S. Immink, *Codes for Mass Data Storage Systems*, 2nd ed. Eindhoven, The Netherlands: Shannon Foundation Publishers, 2004.
[2] B. Vasic and E. M. Kurtas, *Coding and Signal Processing for Magnetic Recording Systems*. Boca Raton, FL, USA: CRC Press, 2005.
[3] D. E. Knuth, "Efficient balanced codes," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 1, pp. 51–53, Jan. 1986.
[4] J. H. Weber and K. A. S. Immink, "Knuth's balanced codes revisited," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1673–1679, Apr. 2010.
[5] K. A. S. Immink and J. H. Weber, "Very efficient balanced codes," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 2, pp. 188–192, Feb. 2010.
[6] S. Al-Bassam and B. Bose, "On balanced codes," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 406–408, Mar. 1990.
[7] ——, "Design of efficient balanced codes," *IEEE Trans. Comput.*, vol. 43, no. 3, pp. 362–365, Mar. 1994.
[8] L. G. Tallini, R. M. Capocelli, and B. Bose, "Design of some new efficient balanced codes," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 790–802, May 1996.
[9] L. G. Tallini and B. Bose, "Balanced codes with parallel encoding and decoding," *IEEE Trans. Comput.*, vol. 48, no. 8, pp. 794–814, Aug. 1999.
[10] N. Alon, E. E. Bergmann, D. Coppersmith, and A. M. Odlyzko, "Balancing sets of vectors," *IEEE Trans. Inf. Theory*, vol. 34, no. 1, pp. 128–130, Jan. 1988.
[11] S. Al-Bassam, "Balanced codes," Ph.D. dissertation, Oregon State University, OR, USA, 1990.
[12] L. G. Tallini and U. Vaccaro, "Efficient m-ary balanced codes," *Disc. App. Math.*, vol. 92, no. 1, pp. 17–56, Mar. 1999.
[13] R. Mascella, L. G. Tallini, S. Al-Bassam, and B. Bose, "On efficient balanced codes over the mth roots of unity," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2214–2217, May 2006.
[14] R. Mascella and L. G. Tallini, "Efficient m-ary balanced codes which are invariant under symbol permutation," *IEEE Trans. Comput.*, vol. 55, no. 8, pp. 929–946, Aug. 2006.
[15] T. G. Swart and J. H. Weber, "Efficient balancing of q-ary sequences with parallel decoding," in *Proc. of the 2009 IEEE Int. Symp. on Inf. Theory*, Seoul, Korea, Jun. 28–Jul. 3, 2009, pp. 1564–1568.
[16] D. Pelusi, S. Elmougy, L. G. Tallini, and B. Bose, "m-ary balanced codes with parallel decoding," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3251–3264, Jun. 2015.
[17] J. H. Weber, K. A. S. Immink, P. H. Siegel, and T. G. Swart, "Perspectives on balanced sequences," Sep. 2012, submitted to *IEEE Trans. Inf. Theory*. [Online]. Available: https://arxiv.org/pdf/1301.6484.pdf
[18] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 8th ed., D. Zwillinger and V. Moll, Eds. San Diego, CA, USA: Academic Press-Elsevier, 2015.
[19] Z. Star, "An asymptotic formula in theory of compositions," *Aequationes Mathematicae*, vol. 13, no. 3, pp. 279–284, Oct. 1975.
[20] R. M. Capocelli, L. Gargano, and U. Vaccaro, "Efficient q-ary immutable codes," *Disc. App. Math.*, vol. 33, no. 1-3, pp. 25–41, Nov. 1991.
[21] L. Pezza, L. G. Tallini, and B. Bose, "Variable length unordered codes," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 548–569, Feb. 2012.
[22] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge, UK: Cambridge University Press, 1952.
[23] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York, USA: John Wiley & Sons, 1968, vol. 1.
[24] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. New York, USA: John Wiley & Sons, 1980.
[25] W. Feller, "Fluctuation theory of recurrent events," *Trans. of the Amer. Math. Soc.*, vol. 67, no. 1, pp. 98–119, Sep. 1949.

[26] K. L. Chung and G. A. Hunt, "On the zeros of $\sum_1^n \pm 1$," *Annals of Mathematics*, vol. 50, no. 2, pp. 385–400, Apr. 1949.

Filip Palunčić was born in Belgrade, Serbia. He received the M. Ing. and D. Ing. degrees from the University of Johannesburg, South Africa in 2008 and 2012, respectively. He spent 4 years in industry as a research and development engineer at IDX, a company specializing in industrial communications. During 2016–2017, he was a post-doctoral research fellow at the Sentech group in Broadband Wireless Multimedia Communications, Department of Electrical, Electronic and Computer Engineering, University of Pretoria. He is currently a faculty member in the Department of Electrical, Electronic and Computer Engineering, University of Pretoria.

His research interests include coding techniques (in particular error control coding and constrained coding), information theory, cognitive radio networks and wireless communications.

B. T. Maharaj received his Ph.D. in Engineering in the area of Wireless Communications from the University of Pretoria. He is a full Professor and Dean and currently holds the research position of Sentech Chair in Broadband Wireless Multimedia Communications in the Department of Electrical, Electronic and Computer Engineering at the University of Pretoria. His research interests are in OFDM-MIMO and massive MIMO systems, cognitive radio resource allocation and 5G cognitive radio sensor networks.