Simple Systematic Pearson Coding

Weber, Jos; Swart, Theo; Schouhamer Immink, KA

DOI
10.1109/ISIT.2016.7541326

Publication date
2016

Document Version
Accepted author manuscript

Published in
Proceedings ISIT 2016

Simple Systematic Pearson Coding

Jos H. Weber*,**
*Delft University of Technology
The Netherlands
j.h.weber@tudelft.nl

Theo G. Swart**
**University of Johannesburg
South Africa
tgswart@uj.ac.za

Kees A. Schouhamer Immink***
***Turing Machines Inc.
The Netherlands
immink@turing-machines.com

***Abstract*—The recently proposed Pearson codes offer immunity against channel gain and offset mismatch. These codes have very low redundancy, but efficient coding procedures were lacking. In this paper, systematic Pearson coding schemes are presented. The redundancy of these schemes is analyzed for memoryless uniform sources. It is concluded that simple coding can be established at only a modest rate loss.**

I. Introduction

Dealing with rapidly varying offset and/or gain is an important issue in signal processing for modern storage and communication systems. For example, methods to solve these difficulties in Flash memories have been discussed in, e.g., [7], [9], and [11]. Also, in optical disc media, the retrieved signal depends on the dimensions of the written features and upon the quality of the light path, which may be obscured by fingerprints or scratches on the substrate, leading to offset and gain variations of the retrieved signal. Automatic gain and offset control in combination with dc-balanced codes are applied albeit at the cost of redundancy [4], and thus improvements to the art are welcome.

Immink and Weber [5] showed that detectors that use the Pearson distance offer immunity to offset and gain mismatch. Use of the Pearson distance demands that the set of codewords satisfies certain special properties. Such sets are called *Pearson codes*. In [10], optimal codes were presented, in the sense of having the largest number of codewords and thus minimum redundancy among all q-ary Pearson codes of fixed length n. However, the important issue of efficient coding procedures was not addressed. In this paper, we present simple systematic Pearson coding schemes, mapping sequences of information symbols generated by a q-ary source to q-ary code sequences. The redundancy of these coding schemes is analyzed for memoryless sources generating q-ary symbols with equal probability.

The remainder of this paper is organized as follows. In Section II, we review the concepts of Pearson detection and q-ary Pearson codes. Then, in Section III, we present our systematic coding schemes and analyze their redundancy. Finally, in Section IV, we draw conclusions.

II. Preliminaries

A. Codes and Redundancies

Let $\mathcal{C}$ be a q-ary code of length n, i.e., $\mathcal{C} \subseteq \mathcal{Q}^n$, where $\mathcal{Q} = \{0, 1, \ldots, q-1\}$ is the code alphabet of size $q \geq 2$. Here the alphabet symbols are to be treated as being real numbers rather than elements of $\mathbb{Z}_q$. The cardinality of the code is denoted by M, i.e., $M = |\mathcal{C}|$. Usually, the redundancy of code $\mathcal{C}$ is then defined as

$$n - \log_q M. \tag{1}$$

Actually, this assumes that all codewords are equally likely to be selected. In a more general setting, an arbitrary probability mass function (PMF) is specified on the codewords. Let the probability that codeword $\boldsymbol{x}_i \in \mathcal{C}$, $1 \leq i \leq M$, is selected for transmission or storage be P_i. Since the average amount of information carried by a codeword is then $-\sum_{i=1}^{M} P_i \log_q P_i$ symbols, the redundancy of code $\mathcal{C}$ with PMF $\{P_i\}$ is

$$n + \sum_{i=1}^{M} P_i \log_q P_i. \tag{2}$$

In case $P_i = 1/M$ for all i, then (2) reduces to (1).

B. Pearson Detection

For convenience, we use the shorthand notation $a\boldsymbol{v} + b = (av_1 + b, av_2 + b, \ldots, av_n + b)$. A common assumption is that a transmitted codeword $\boldsymbol{x}$ is received as a vector $\boldsymbol{r} = a(\boldsymbol{x} + \boldsymbol{\nu}) + b$ in $\mathbb{R}^n$. Here a and b are unknown real numbers with a positive, called the *gain* and the (dc-)*offset*, respectively. Moreover, $\boldsymbol{\nu}$ is an additive noise vector, where the $\nu_i \in \mathbb{R}$ are noise samples from a zero-mean Gaussian distribution. Note that both gain and offset do not vary from symbol to symbol, but are the same for the whole block of n symbols. The receiver's ignorance of the channel's momentary gain and offset may lead to massive performance degradation as shown, for example, in [5] when a traditional detector, based on thresholds or the Euclidean distance, is used. In the prior art, various methods have been proposed to overcome this difficulty. In a first method, data reference, or 'training', patterns are multiplexed with the user data in order to 'teach' the data detection circuitry the momentary values of the channel's characteristics such as impulse response, gain, and offset. In a channel with unknown gain and offset, we may use two reference symbol values, where in each codeword, a first symbol is set equal to the lowest signal level and a second symbol equal to the highest signal level. The positions and amplitudes of the two reference symbols are known to the receiver. The receiver can straightforwardly measure the amplitude of the retrieved reference symbols, and normalize the amplitudes of the remaining symbols of the retrieved

codeword before applying detection. Clearly, the redundancy of the method is two symbols per codeword.

In a second prior art method, codes satisfying equal balance and energy constraints [2], which are immune to gain and offset mismatch, have been advocated. However, these codes suffer from a rather high redundancy. In a recent contribution, Pearson distance detection is advocated since its redundancy is much less than that of balanced codes [5]. The Pearson distance between the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as follows. For a vector $\boldsymbol{u}$, define $\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n} u_i$ and $\sigma_{\boldsymbol{u}}^2 = \sum_{i=1}^{n}(u_i - \overline{\boldsymbol{u}})^2$. Note that $\sigma_{\boldsymbol{u}}$ is closely related to, but not the same as, the standard deviation of $\boldsymbol{u}$. The *(Pearson) correlation coefficient* of $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined by

$$\rho_{\boldsymbol{u},\boldsymbol{v}} = \frac{\sum_{i=1}^{n}(u_i - \overline{\boldsymbol{u}})(v_i - \overline{\boldsymbol{v}})}{\sigma_{\boldsymbol{u}}\sigma_{\boldsymbol{v}}}, \tag{3}$$

and the *Pearson distance* between $\boldsymbol{u}$ and $\boldsymbol{v}$ is given by

$$\delta(\boldsymbol{u},\boldsymbol{v}) = 1 - \rho_{\boldsymbol{u},\boldsymbol{v}}. \tag{4}$$

The Pearson distance and Pearson correlation coefficient are well-known concepts in statistics and cluster analysis. Since $|\rho_{\boldsymbol{u},\boldsymbol{v}}| \leq 1$, it holds that $0 \leq \delta(\boldsymbol{u},\boldsymbol{v}) \leq 2$. The Pearson distance is translation and scale invariant, that is, $\delta(\boldsymbol{u},\boldsymbol{v}) = \delta(\boldsymbol{u}, a\boldsymbol{v}+b)$, for any real numbers a and b with $a > 0$.

Upon receipt of a vector $\boldsymbol{r}$, a minimum Pearson distance detector outputs the codeword $\arg\min_{\boldsymbol{x}\in\mathcal{C}} \delta(\boldsymbol{r},\boldsymbol{x})$. Since the Pearson distance is translation and scale invariant, we conclude that the Pearson distance between the received vector and a codeword is independent of the channel's gain or offset mismatch, so that, as a result, the error performance of the minimum Pearson distance detector is immune to gain and offset mismatch, which is a big advantage in comparison to Euclidean distance detectors. However, Pearson distance detectors are more sensitive to noise. Therefore, hybrid minimum Pearson and Euclidean distance detectors have been proposed [6] to deal with channels suffering from both significant noise and gain/offset.

C. Pearson Codes

Its immunity to gain and offset mismatch implies that the minimum Pearson distance detector cannot be used in conjunction with arbitrary codes, since $\delta(\boldsymbol{r},\boldsymbol{x}) = \delta(\boldsymbol{r},\boldsymbol{y})$ if $\boldsymbol{y} = c_1 + c_2\boldsymbol{x}$, with $c_1, c_2 \in \mathbb{R}$ and c_2 positive. In other words, since a minimum Pearson detector cannot distinguish between the words $\boldsymbol{x}$ and $\boldsymbol{y} = c_1 + c_2\boldsymbol{x}$, the codewords must be taken from a code $\mathcal{C} \subseteq \mathcal{Q}^n$ that guarantees unambiguous detection with the Pearson distance metric (4) accordingly. Furthermore, note that codewords of the format $\boldsymbol{x} = (c, c, \ldots, c)$ should not be used in order to avoid that $\sigma_{\boldsymbol{x}} = 0$, which would lead to an undefined Pearson correlation coefficient. In conclusion, the following condition must be satisfied:

$$\begin{aligned} &\text{If } \boldsymbol{x} \in \mathcal{C} \text{ then } c_1 + c_2\boldsymbol{x} \notin \mathcal{C} \text{ for all } c_1, c_2 \in \mathbb{R} \\ &\qquad\qquad \text{with } (c_1, c_2) \neq (0, 1) \text{ and } c_2 \geq 0. \end{aligned} \tag{5}$$

A code satisfying (5) is called a Pearson code [10]. Known constructions of Pearson codes read as follows.

- The set of all q-ary sequences of length n having at least one symbol '0' and at least one symbol '1'. We denote this code by $\mathcal{T}(n,q)$. It is a member of the class of T-constrained codes [3], consisting of sequences in which T pre-determined reference symbols each appear at least once.
- The set of all q-ary sequences of length n having at least one symbol '0', at least one symbol not equal to '0', and having the greatest common divisor of the sequence symbols equal to '1'. We denote this code by $\mathcal{P}(n,q)$. It is has been shown in [10] that this code is optimal in the sense that it has the largest number of codewords among all q-ary Pearson codes of length n.

Another code which is of interest, though not being a Pearson code, is defined as follows.

- The set of all q-ary sequences of length n having at least one symbol '0'. We denote this code by $\mathcal{Z}(n,q)$. It is also a member of the class of T-constrained codes [3]. Due to the presence of the reference symbol '0' it is resistant against offset mismatch.

Note that

$$\mathcal{T}(n,q) \subseteq \mathcal{P}(n,q) \subseteq \mathcal{Z}(n,q). \tag{6}$$

The cardinalities and redundancies (in the sense of (1)) of these three codes, as derived in [10], are given in Table I, where, for a positive integer d, the Möbius function $\mu(d)$ is defined [1, Chapter XVI] to be 0 if d is divisible by the square of a prime, otherwise $\mu(d) = (-1)^k$ where k is the number of (distinct) prime divisors of d.

III. Systematic Coding

As stated, the Pearson code $\mathcal{P}(n,q)$ is optimal in the sense of having largest cardinality and thus smallest redundancy. However, an easy coding procedure mapping information sequences to code sequences and vice versa is not evident at all. In this section, we propose easy coding procedures, possibly at the expense of a somewhat higher redundancy. We only use code sequences of a fixed length n, but for the information we consider both fixed-length and variable-length sequences. Hence, fixed-to-fixed (FF) as well as variable-to-fixed (VF) length coding schemes are proposed. For the source we make the common assumption that it is memoryless and that all q source symbols appear with equal probability $1/q$. We start by introducing simple coding schemes resistant against offset mismatch only. Then we continue with similar procedures for Pearson coding.

A. Systematic Coding for $\mathcal{Z}(n,q)$

The code $\mathcal{Z}(n,q)$ consists of all q-ary sequence of length n containing at least one symbol '0'. Its cardinality and redundancy are given in Table I. Here, we propose simple coding procedures systematically mapping q-ary information symbols to code sequences $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ in $\mathcal{Z}(n,q)$.

A well-known extremely simple FF-scheme, which we call $\mathcal{Z}_{\mathrm{FF}}(n,q)$, is to fill the code sequence $\boldsymbol{x}$ with $n-1$ information symbols in the subsequence $(x_1, x_2, \ldots, x_{n-1})$ and to set

TABLE I
CARDINALITY AND REDUNDANCY OF THE CODES $\mathcal{T}(n,q)$, $\mathcal{P}(n,q)$, AND $\mathcal{Z}(n,q)$.

	Cardinality	Redundancy
$\mathcal{T}(n,q)$	$q^n - 2(q-1)^n + (q-2)^n$	$-\log_q\left(1-2\left(\frac{q-1}{q}\right)^n+\left(\frac{q-2}{q}\right)^n\right)$ $\approx \left(2\left(\frac{q-1}{q}\right)^n-\left(\frac{q-2}{q}\right)^n\right)\Big/\ln(q)$
$\mathcal{P}(n,q)$	$\sum_{d=1}^{q-1}\mu(d)\left(\left(\left\lfloor\frac{q-1}{d}\right\rfloor+1\right)^n-\left\lfloor\frac{q-1}{d}\right\rfloor^n-1\right)$ $= q^n-(q-1)^n+O(\lceil q/2\rceil^n)$ as $n\to\infty$	$-\log_q\left(1-\left(\frac{q-1}{q}\right)^n+O\left(\left(\frac{q+1}{2q}\right)^n\right)\right)$ $\approx \left(\left(\frac{q-1}{q}\right)^n+O\left(\left(\frac{q+1}{2q}\right)^n\right)\right)\Big/\ln(q)$
$\mathcal{Z}(n,q)$	$q^n-(q-1)^n$	$-\log_q\left(1-\left(\frac{q-1}{q}\right)^n\right)$ $\approx \left(\frac{q-1}{q}\right)^n\Big/\ln(q)$

$x_n = 0$. Due to the fixed last symbol, which acts as a reference, the redundancy of this method is 1.

Note that while the redundancy of $\mathcal{Z}(n,q)$ is decreasing in n, the redundancy of $\mathcal{Z}_{\rm FF}(n,q)$ remains 1. Next, we propose a systematic VF-scheme, $\mathcal{Z}_{\rm VF}(n,q)$, for which the redundancy decreases in n:

1) Take $n-1$ information from the q-ary source and set these as $(x_1, x_2, \ldots, x_{n-1})$.
2) If $x_i = 0$ for at least one $1 \le i \le n-1$, then choose x_n to be a (new) information symbol, otherwise set $x_n = 0$.

It can easily be seen that the code sequence $\boldsymbol{x}$ is indeed in $\mathcal{Z}(n,q)$ and that the information symbols can be uniquely retrieved from $\boldsymbol{x}$ by checking whether it contains a zero in its first $n-1$ positions: if 'yes', then all n code symbols are information symbols, if 'no', then only the first $n-1$ code symbols are information symbols. Since the number of information symbols may vary from codeword to codeword (being either n or $n-1$), while the length of the codewords is fixed at n, this can be considered a variable-to-fixed length coding procedure. All words in $\mathcal{Z}(n,q)$ can appear as code sequence, but not necessarily with equal probability. This leads to a redundancy as stated in the next theorem.

Theorem 1. *For a memoryless uniform q-ary source, the redundancy of coding scheme $\mathcal{Z}_{\rm VF}(n,q)$ is $(1-1/q)^{n-1}$.*

Proof: This result can be obtained using (2), with the observations that (i) $P_i = (1/q)^{n-1}$ for the $(q-1)^{n-1}$ code sequences $\boldsymbol{x}_i$ with no zeroes among the first $n-1$ symbols and thus with last code symbol equal to zero, and (ii) $P_i = (1/q)^n$ for the other $q(q^{n-1}-(q-1)^{n-1})$ code sequences $\boldsymbol{x}_i$ with at least one zero among the first $n-1$ symbols. Hence, the resulting redundancy is

$$\begin{aligned} & n + \sum_{i=1}^{M} P_i \log_q P_i \\ &= n + (q-1)^{n-1}(1/q)^{n-1}\log_q(1/q)^{n-1} + \\ &\quad q(q^{n-1}-(q-1)^{n-1})(1/q)^n \log_q(1/q)^n \\ &= (1-1/q)^{n-1}. \end{aligned}$$

Another way to derive this result is to observe that the probability of the case that a sequence of $n-1$ information symbols does not contain a zero, leading to one redundant symbol, is equal to $(1-1/q)^{n-1}$, while the opposite case leads to no redundancy at all. The weighted average

$$(1-1/q)^{n-1}\times 1 + (1-(1-1/q)^{n-1})\times 0 = (1-1/q)^{n-1}$$

then gives the redundancy of $\mathcal{Z}_{\rm VF}(n,q)$. ■

TABLE II
$\mathcal{Z}_{\rm VF}(3,2)$ CODING FOR A MEMORYLESS UNIFORM BINARY SOURCE.

Info	Codeword $\in \mathcal{Z}(3,2)$	Probability	Redundancy
000	000	1/8	0
001	001	1/8	0
010	010	1/8	0
011	011	1/8	0
100	100	1/8	0
101	101	1/8	0
11	110	1/4	1

As an example, we consider scheme $\mathcal{Z}_{\rm VF}(3,2)$ for a memoryless binary source producing zeroes and ones with equal probability. The seven codewords of $\mathcal{Z}(3,2)$ are then used with probabilities as indicated in Table II, and thus the average redundancy is 1/4. This result can be obtained by applying (2), i.e., $3 + 6\times(1/8)\log_2(1/8) + (1/4)\log_2(1/4) = 1/4$, or by directly applying Theorem 1, i.e, $(1-1/2)^2 = 1/4$. Note that achieving the somewhat lower redundancy $3-\log_2(7) = 0.19$ of the code $\mathcal{Z}(3,2)$ as such would require all seven codewords to be used with probability 1/7, which does not naturally match the source statistics.

In conclusion, the redundancy of $\mathcal{Z}_{\rm VF}(n,q)$ is $(1-1/q)^{n-1}$, while the approximate redundancy of $\mathcal{Z}(n,q)$ is $(1-1/q)^n/\ln q$ as given in Table I. Hence, the redundancy of the proposed VF-scheme $\mathcal{Z}_{\rm VF}(n,q)$ is roughly a factor

$$q\ln(q)/(q-1)$$

higher than the redundancy of $\mathcal{Z}(n,q)$. Note that this factor does not depend on the code length n, but only on the alphabet size q. For the binary case $q=2$ this factor is $2\ln(2) = 1.39$, for the quaternary case $q=4$ it is $(4/3)\ln(4) = 1.85$, while for large values of q it is roughly $\ln(q)$.

B. Systematic Pearson Coding

An extremely simple FF scheme, called $\mathcal{T}_{\mathrm{FF}}(n,q)$, resistant against both offset and gain mismatch, is to fill the first $n-2$ positions in the code sequence $\boldsymbol{x}$ with information symbols and to reserve the last two symbols for reference purposes: $x_{n-1}=0$ and $x_n=1$. The resulting code sequence is in $\mathcal{T}(n,q)$ since it contains at least one '0' and at least one '1'. The redundancy of this scheme is fixed at 2 symbols, but, again, it would be desirable to have a systematic scheme with a redundancy decreasing in the code length, preferably approaching zero for large values of n.

The first VF Pearson scheme, called $\mathcal{T}_{\mathrm{VF}}(n,q)$, we propose is similar to the VF scheme $\mathcal{Z}_{\mathrm{VF}}(n,q)$ presented in the previous subsection. It reads as follows.

1) Take $n-2$ information from the q-ary source and set these as $(x_1,x_2,\ldots,x_{n-2})$.
2) If $x_i=0$ for at least one $1\leq i\leq n-2$, then choose x_{n-1} to be a (new) information symbol, otherwise set $x_{n-1}=0$.
3) If $x_i=1$ for at least one $1\leq i\leq n-1$, then choose x_n to be a (new) information symbol, otherwise set $x_n=1$.

Since any code sequence obtained this way contains at least one '0' and at least one '1', it is a member of $\mathcal{T}(n,q)$. Also, the $n-2$, $n-1$, or n information symbols can easily be retrieved from the code sequence. The redundancy of this scheme is given in the next theorem.

Theorem 2. *For a memoryless uniform q-ary source, the redundancy of coding scheme $\mathcal{T}_{\mathrm{VF}}(n,q)$ is*

$$\left(\frac{2q-1}{q}\right)\left(\frac{q-1}{q}\right)^{n-2}+\left(\frac{1}{q}\right)\left(\frac{q-2}{q}\right)^{n-2}.$$

Proof: The probability that a code sequence $\boldsymbol{x}$ has two redundant symbols is

$$(1-2/q)^{n-2}, \tag{7}$$

which is the probability of having an information sequence of length $n-2$ without zeroes and ones. Further, the probability that $\boldsymbol{x}$ has only a redundant symbol in position $n-1$ is

$$(1-1/q)^{n-2}-(1-2/q)^{n-2}, \tag{8}$$

which is the probability of having an information sequence of length $n-2$ without zeroes but with at least one '1'. The probability that $\boldsymbol{x}$ has only a redundant symbol in position n is

$$\left((1-1/q)^{n-2}-(1-2/q)^{n-2}\right)(1-1/q), \tag{9}$$

where the first multiplicative term is the probability of having an information sequence of length $n-2$ without ones but with at least one '0' and the second multiplicative term is the probability that the information symbol in position $n-1$ is not equal to '1'. Hence, the redundancy is two times the term in (7) plus the terms in (8) and (9), which gives the expression stated in the theorem. ∎

The redundancy of $\mathcal{T}_{\mathrm{VF}}(n,q)$ as stated in Theorem 2 is, for large values of n, a factor

$$\frac{q(2q-1)}{2(q-1)^2}\ln(q)$$

higher than the redundancy of $\mathcal{T}(n,q)$ as stated in Table I. For the binary case $q=2$ this factor is $3\ln(2)=2.08$, for the quaternary case $q=4$ it is $(14/9)\ln(4)=2.16$, while for large values of q it is roughly $\ln(q)$.

The second VF Pearson scheme, called $\mathcal{P}_{\mathrm{VF}}(n,q)$, we propose is based on relaxing the enforcement of having both at least one '0' and at least one '1' in all code sequences to the enforcement that all code sequences $\mathbf{x}$ contain at least one '0' and have the greatest common divisor (GCD) of the x_i equal to one, i.e., $\mathrm{GCD}\{x_1,\ldots,x_n\}=1$. It reads as follows.

1) Take $n-2$ information from the q-ary source and set these as $(x_1,x_2,\ldots,x_{n-2})$.
2) If $x_i=0$ for at least one $1\leq i\leq n-2$, then choose x_{n-1} to be a (new) information symbol, otherwise set $x_{n-1}=0$.
3) If $\mathrm{GCD}\{x_1,\ldots,x_{n-1}\}=1$, then choose x_n to be a (new) information symbol, otherwise set $x_n=1$.

Any code sequence obtained in this way is a member of $\mathcal{P}(n,q)$. Again, the $n-2$, $n-1$, or n information symbols can easily be retrieved from the code sequence. For $q=2$ and $q=3$, the scheme $\mathcal{P}_{\mathrm{VF}}(n,q)$ is the same as $\mathcal{T}_{\mathrm{VF}}(n,q)$, since the condition that a sequence has a GCD of 1 is then equivalent to the condition that a sequence contains a '1'. Therefore, the redundancy is as stated in Theorem 2 in these cases. However, this is not the case if $q\geq 4$, for which we give the redundancy of $\mathcal{P}_{\mathrm{VF}}(n,q)$ in the next theorem. First, we present a lemma, of which the proof is summarized due to lack of space.

Lemma 1. *For any fixed $q\geq 4$, among the q^n q-ary sequences $\boldsymbol{y}$ of length n, there are*

1) $q^n-(q-1)^n+O(\lceil q/2\rceil^n)$ *sequences with* $\mathrm{GCD}(\boldsymbol{y})=1$ *containing at least one '0',*
2) $O(\lceil q/2\rceil^n)$ *sequences with* $\mathrm{GCD}(\boldsymbol{y})\neq 1$ *containing at least one '0',*
3) $(q-1)^n+O(\lfloor (q-1)/2\rfloor^n)$ *sequences with* $\mathrm{GCD}(\boldsymbol{y})=1$ *containing no symbol '0',*
4) $O(\lfloor (q-1)/2\rfloor^n)$ *sequences with* $\mathrm{GCD}(\boldsymbol{y})\neq 1$ *containing no symbol '0'.*

Proof: The first result was proved in [10]. Combining this with the fact that the number of q-ary sequence of length n containing at least one '0' is $q^n-(q-1)^n$ gives the second result.

Using a well-known counting argument from, e.g., Section 16.5 in [1], it follows that the number of sequences of length n with symbols from $\{1,2,\ldots,q-1\}$ and GCD equal to 1 is

$$\sum_{d=1}^{q-1}\mu(d)\lfloor (q-1)/d\rfloor^n=(q-1)^n+O(\lfloor (q-1)/2\rfloor^n),$$

where $\mu(d)$ is the Möbius function already mentioned at the end of Subsection II-C. This proves the third result, which

combined with the fact that the number of q-ary sequence of length n containing no symbol '0' is $(q-1)^n$ also gives the fourth result. ∎

Theorem 3. *For a memoryless uniform q-ary source, with fixed $q \geq 4$, the redundancy of coding scheme $\mathcal{P}_{\rm VF}(n,q)$ is*

$$\left(\frac{q-1}{q}\right)^{n-2} + O\left(\left(\frac{\lceil q/2 \rceil}{q}\right)^{n-2}\right).$$

Proof: The probability that a code sequence $\boldsymbol{x}$ has two redundant symbols is

$$O\left(\left(\frac{\lfloor (q-1)/2 \rfloor}{q}\right)^{n-2}\right), \tag{10}$$

which is the probability of having an information sequence of length $n-2$ without zeroes and with a GCD unequal to 1, as follows from result 4) in Lemma 1. Further, the probability that $\boldsymbol{x}$ has only a redundant symbol in position $n-1$ is

$$\left(\frac{q-1}{q}\right)^{n-2} + O\left(\left(\frac{\lfloor (q-1)/2 \rfloor}{q}\right)^{n-2}\right), \tag{11}$$

which is the probability of having an information sequence of length $n-2$ without zeroes but with a GCD equal to 1, as follows from result 3) in Lemma 1. The probability that $\boldsymbol{x}$ has only a redundant symbol in position n is

$$O\left(\left(\frac{\lceil q/2 \rceil}{q}\right)^{n-2}\right), \tag{12}$$

as follows from result 2) in Lemma 1. Hence, the redundancy is two times the term in (10) plus the terms in (11) and (12), which gives the expression stated in the theorem. ∎

The redundancy of $\mathcal{P}_{\rm VF}(n,q)$ as stated in Theorem 3 is, for fixed $q \geq 4$ and large values of n, a factor

$$\left(\frac{q}{q-1}\right)^2 \ln(q)$$

higher than the redundancy of $\mathcal{P}(n,q)$ as stated in Table I. For the quaternary case $q=4$ this factor is $(16/9)\ln(4) = 2.46$, while for large values of q it is roughly $\ln(q)$. Also, note that, again for fixed $q \geq 4$ and large values of n, the redundancy of $\mathcal{P}_{\rm VF}(n,q)$ is a factor $q/(q-1)$ higher than the redundancy of $\mathcal{Z}_{\rm VF}(n,q)$.

IV. Conclusions

We have presented simple systematic q-ary coding schemes which are resistant against offset as well as gain mismatch or against offset mismatch only. Both coding for fixed and coding for variable length source sequences have been considered, resulting in FF and VF schemes of fixed code block length n, respectively. We analyzed the redundancy of the proposed schemes for memoryless uniform sources. The major findings are summarized in Table III.

The redundancy of the Pearson schemes $\mathcal{T}_{\rm VF}(n,q)$ and $\mathcal{P}_{\rm VF}(n,q)$, resistant against offset as well as gain mismatch, approaches zero for large n, as desired. The redundancy for both schemes is equal if $q = 2,3$ and the redundancy of the former scheme exceeds the redundancy of the the latter scheme by a factor of $(2q-1)/q$ if $q \geq 4$. Furthermore, the redundancy of the Pearson scheme $\mathcal{P}_{\rm VF}(n,q)$ exceeds the redundancy of the $\mathcal{Z}_{\rm VF}(n,q)$ scheme, which offers immunity to offset mismatch only, by a factor of $(2q-1)/(q-1)$ if $q = 2,3$ and by a factor of only $q/(q-1)$ if $q \geq 4$. The schemes $\mathcal{T}_{\rm FF}(n,q)$ and $\mathcal{Z}_{\rm FF}(n,q)$ offer extreme simplicity, using fixed training symbols in fixed positions, at the price of a redundancy which does not decrease with increasing n.

Finally, the redundancy of the presented $\mathcal{T}_{\rm VF}(n,q)$, $\mathcal{P}_{\rm VF}(n,q)$, and $\mathcal{Z}_{\rm VF}(n,q)$ schemes is a bit higher than the redundancy of their $\mathcal{T}(n,q)$, $\mathcal{P}(n,q)$, and $\mathcal{Z}(n,q)$ associates. However, note that the low redundancies of these codes as such are only achieved under the assumption that all their codewords are used equally likely, which is hard to realize for memoryless uniform and other practical sources. In contrast, our VF schemes come with natural simple coding mechanisms.

TABLE III
Approximate redundancy of the codes $\mathcal{T}(n,q)$, $\mathcal{P}(n,q)$, and $\mathcal{Z}(n,q)$ and the related FF and VF schemes, for large n and fixed $q \geq 4$.

	Redundancy	Red. FF	Red. VF
$\mathcal{T}(n,q)$	$2\left(\frac{q-1}{q}\right)^n \Big/ \ln(q)$	2	$\frac{2q-1}{q}\left(\frac{q-1}{q}\right)^{n-2}$
$\mathcal{P}(n,q)$	$\left(\frac{q-1}{q}\right)^n \Big/ \ln(q)$		$\left(\frac{q-1}{q}\right)^{n-2}$
$\mathcal{Z}(n,q)$	$\left(\frac{q-1}{q}\right)^n \Big/ \ln(q)$	1	$\left(\frac{q-1}{q}\right)^{n-1}$

References

[1] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers* (Fifth Edition), Oxford University Press, Oxford, 1979.

[2] K. A. S. Immink, "Coding Schemes for Multi-Level Channels with Unknown Gain and/or Offset Using Balance and Energy constraints", *IEEE Int. Symposium on Inform. Theory (ISIT)*, Istanbul, Turkey, July 2013.

[3] K. A. S. Immink, "Coding Schemes for Multi-Level Flash Memories that are Intrinsically Resistant Against Unknown Gain and/or Offset Using Reference Symbols", *Electronics Letters,* vol. 50, pp. 20–22, 2014.

[4] K. A. S. Immink and J. H. Weber, "Very Efficient Balanced Codes", *IEEE Journal on Selected Areas of Communications,* vol. 28, pp. 188–192, 2010.

[5] K. A. S. Immink and J. H. Weber, "Minimum Pearson Distance Detection for Multi-Level Channels with Gain and/or Offset Mismatch", *IEEE Trans. Inform. Theory,* vol. 60, pp. 5966–5974, Oct. 2014.

[6] K. A. S. Immink and J. H. Weber, "Hybrid Minimum Pearson and Euclidean Distance Detection", *IEEE Trans. Commun.,* vol. 63, no. 9, pp. 3290–3298, Sept. 2015.

[7] A. Jiang, R. Mateescu, M. Schwartz, and J. Bruck, "Rank Modulation for Flash Memories", *IEEE Trans. Inform. Theory,* vol. 55, no. 6, pp. 2659–2673, June 2009.

[8] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics* (Third Edition), McGraw-Hill, 1974.

[9] F. Sala, K. A. S. Immink, and L. Dolecek, "Error Control Schemes for Modern Flash Memories: Solutions for Flash Deficiencies", *IEEE Consumer Electronics Magazine,* vol. 4, no.1, pp. 66–73, Jan. 2015.

[10] J. H. Weber, K. A. S. Immink, and S.R. Blackburn, "Pearson Codes", *IEEE Trans. Inform. Theory,* vol. 62, no. 1, pp. 131–135, Jan. 2016.

[11] H. Zhou, A. Jiang, and J. Bruck, "Error-correcting schemes with dynamic thresholds in nonvolatile memories", *IEEE Int. Symposium on Inform. Theory (ISIT)*, St. Petersburg, Russia, July 2011.