# Pearson Codes

Jos H. Weber[*]
Kees A. Schouhamer Immink[†]
and
Simon R. Blackburn[‡]

September 25, 2015

## Abstract

The Pearson distance has been advocated for improving the error performance of noisy channels with unknown gain and offset. The Pearson distance can only fruitfully be used for sets of $q$-ary codewords, called *Pearson codes*, that satisfy specific properties. We will analyze constructions and properties of optimal Pearson codes. We will compare the redundancy of optimal Pearson codes with the redundancy of prior art $T$-constrained codes, which consist of $q$-ary sequences in which $T$ pre-determined reference symbols appear at least once. In particular, it will be shown that for $q \leq 3$ the 2-constrained codes are optimal Pearson codes, while for $q \geq 4$ these codes are not optimal.

**Key words:** flash memory, digital optical recording, Non-Volatile Memory, NVM, Pearson distance.

[*]Jos H. Weber is with Delft University of Technology, Delft, The Netherlands. E-mail: j.h.weber@tudelft.nl.

[†]Kees A. Schouhamer Immink is with Turing Machines Inc, Willemskade 15b-d, 3016 DK Rotterdam, The Netherlands. E-mail: immink@turing-machines.com.

[‡]Simon R. Blackburn is with the Department of Mathematics, Royal Holloway University of London, Egham, Surrey TW20 0EX, United Kingdom. E-mail: S.Blackburn@rhul.ac.uk

# 1   Introduction

In non-volatile memories, such as floating gate memories, the data is represented by stored charge, which can leak away from the floating gate. This leakage may result in a shift of the offset or *threshold voltage* of the memory cell. The amount of leakage depends on the time elapsed between writing and reading the data. As a result, the offset between different groups of cells may be very different so that prior art automatic offset or gain control, which estimates the mismatch from the previously received data, can not be applied. Methods to solve these difficulties in Flash memories have been discussed in, for example, [4], [5], [6], [7]. In optical disc media, such as the popular Compact Disc, DVD, and Blu-ray disc, 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. Fingerprints and scratches will result in rapidly varying 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 [2], and thus improvements to the art are welcome.

Immink & Weber [3] showed that detectors that use the Pearson distance offer immunity to offset and gain mismatch. The Pearson distance can only be used for a set of codewords with special properties, called a *Pearson set* or *Pearson code*. Let $\mathcal{S}$ be a codebook of chosen $q$-ary codewords $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ over the $q$-ary alphabet $\mathcal{Q} = \{0, 1, \ldots, q - 1\}$, $q \geq 2$, where $n$, the *length* of $\boldsymbol{x}$, is a positive integer. Note that the alphabet symbols are to be treated as being just integers rather than elements of $\mathbb{Z}_q$. A Pearson code with maximum possible size given the parameters $q$ and $n$ is said to be *optimal*.

In Section 2, we set the stage with a description of Pearson distance detection and the properties of the constrained codes used in conjunction with it. Section 3 gives a description of $T$-constrained codes, a type of code described in the prior art [3], used in conjunction with the Pearson distance detector, while Section 4 offers a general construction of optimal Pearson codes and a computation of their cardinalities. The rates of $T$-constrained codes will be compared with optimal rates of Pearson codes. In Section 5, we will describe our conclusions.

## 2 Preliminaries

We use the shorthand notation $a\boldsymbol{v}+b = (av_1+b, av_2+b, \ldots, av_n+b)$. In [3], the authors suppose a situation where the sent codeword, $\boldsymbol{x}$, is received as the vector $\boldsymbol{r} = a(\boldsymbol{x} + \boldsymbol{\nu}) + b$, $r_i \in \mathbb{R}$. 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: $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$, where $\nu_i \in \mathbb{R}$ 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 all $n$ symbols.

The receiver's ignorance of the channel's momentary gain and offset may lead to massive performance degradation as shown, for example, in [3] when a traditional detector, such as threshold or maximum likelihood detector, 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 [8], which are immune to gain and offset mismatch, have been advocated. The redundancy of these codes, denoted by $r_0$, is given by [8]

$$r_0 \approx \log_q n + \log_q(q^2 - 1)\sqrt{q^2 - 4} + \log_q \frac{\pi}{12\sqrt{15}}. \tag{1}$$

In a recent contribution, Pearson distance detection is advocated since its redundancy is much less than that of balanced codes [3]. The Pearson distance between the vectors $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ is defined as follows. For a vector

$\boldsymbol{x}$, define

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2}$$

and

$$\sigma_{\boldsymbol{x}}^2 = \sum_{i=1}^{n}(x_i - \overline{\boldsymbol{x}})^2. \tag{3}$$

Note that $\sigma_{\boldsymbol{x}}$ is closely related to, but not the same as, the standard deviation of $\boldsymbol{x}$. The *(Pearson) correlation coefficient* is defined by

$$\rho_{\boldsymbol{x},\hat{\boldsymbol{x}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{\boldsymbol{x}})(\hat{x}_i - \overline{\hat{\boldsymbol{x}}})}{\sigma_{\boldsymbol{x}}\sigma_{\hat{\boldsymbol{x}}}}, \tag{4}$$

and the *Pearson distance* is given by

$$\delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 1 - \rho_{\boldsymbol{x},\hat{\boldsymbol{x}}}. \tag{5}$$

The Pearson distance and Pearson correlation coefficient are well-known concepts in statistics and cluster analysis. Note that we have $|\rho_{\boldsymbol{x},\hat{\boldsymbol{x}}}| \leq 1$ by a corollary of the Cauchy-Schwarz Inequality [9, Section IV.4.6], which implies that $0 \leq \delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) \leq 2$.

A minimum Pearson distance detector outputs the codeword

$$\boldsymbol{x}_o = \arg\min_{\hat{\boldsymbol{x}}\in\mathcal{S}} \delta(\boldsymbol{r}, \hat{\boldsymbol{x}}).$$

As the Pearson distance is translation and scale invariant, that is,

$$\delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \delta(a\boldsymbol{x} + b, \hat{\boldsymbol{x}}),$$

we conclude that the Pearson distance between the vectors $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ 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. This virtue implies, however, that the minimum Pearson distance detector cannot be used in conjunction with arbitrary codebooks, since

$$\delta(\boldsymbol{r}, \hat{\boldsymbol{x}}) = \delta(\boldsymbol{r}, \hat{\boldsymbol{y}})$$

4

if $\hat{\boldsymbol{y}} = c_1 \hat{\boldsymbol{x}} + c_2$, $c_1, c_2 \in \mathbb{R}$ and $c_1 > 0$. In other words, since a minimum Pearson detector cannot distinguish between the words $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}} = c_1 \hat{\boldsymbol{x}} + c_2$, the codewords must be taken from a codebook $\mathcal{S} \subseteq \mathcal{Q}^n$ that guarantees unambiguous detection with the Pearson distance metric (5).

It is a well-known property of the Pearson correlation coefficient, $\rho_{\boldsymbol{x}, \hat{\boldsymbol{x}}}$, that

$$\rho_{\boldsymbol{x}, \hat{\boldsymbol{x}}} = 1$$

if and only if

$$\hat{\boldsymbol{x}} = c_1 + c_2 \boldsymbol{x},$$

where the coefficients $c_1$ and $c_2 > 0$ are real numbers [9, Section IV.4.6]. It is further immediate, see (4), that the Pearson distance is undefined for codewords $\boldsymbol{x}$ with $\sigma_{\boldsymbol{x}} = 0$, i.e., for multiples of the all-one vector. We coined the name *Pearson code* for a set of codewords that can be uniquely decoded by a minimum Pearson distance detector. We conclude that codewords in a Pearson code must satisfy two conditions, namely

- *Property A:* If $\boldsymbol{x} \in \mathcal{S}$ then $c_1 + c_2 \boldsymbol{x} \notin \mathcal{S}$ for all $c_1, c_2 \in \mathbb{R}$ with $(c_1, c_2) \neq (0, 1)$ and $c_2 > 0$.

- *Property B:* $\boldsymbol{x} = (c, c, \dots, c) \notin \mathcal{S}$ for all $c \in \mathbb{R}$ .

In the remaining part of this paper, we will study constructions and properties of Pearson codes. In particular, we are interested in Pearson codes that are optimal in the sense of having the largest number of codewords for given parameters $n$ and $q$. We will commence with a description of prior art $T$-constrained codes, a first example of Pearson codes.

## 3   $T$-constrained codes

For integers $T$ satisfying $1 \leq T \leq q$, $T$-constrained codes [1], denoted by $\mathcal{S}_{q,n}(a_1, \dots, a_T)$, consist of $q$-ary codewords of length $n$, where $T$ *preferred* or *reference* symbols $a_1, \dots, a_T \in \mathcal{Q}$ must each appear at least once in a codeword. Thus, each codeword, $(x_1, x_2, \dots, x_n)$, in a $T$-constrained code satisfies

$$|\{i : x_i = j\}| > 0 \text{ for each } j \in \{a_1, \dots, a_T\}.$$

The number of $q$-ary sequences of length $n$, $N_T(q, n)$, where $T$ distinct pre-defined symbols occur at least once in every sequence, equals [1]

$$N_T(q, n) = \sum_{i=0}^{T} (-1)^i \left( {}^T_{T-i} \right) (q-i)^n, \ n \geq T. \tag{6}$$

For example, we easily find for $T = 1$ and $T = 2$ that

$$N_1(q, n) = q^n - (q-1)^n \tag{7}$$

and

$$N_2(q, n) = q^n - 2(q-1)^n + (q-2)^n. \tag{8}$$

Clearly, the number of $T$-constrained sequences is not affected by the choice of the specific $T$ symbols we like to favor.

For the binary case, $q = 2$, we simply find that $\mathcal{S}_{2,n}(0)$ is obtained by removing the all-'1' word from $\mathcal{Q}^n$, that $\mathcal{S}_{2,n}(1)$ is obtained by removing the all-'0' word from $\mathcal{Q}^n$, and that $\mathcal{S}_{2,n}(0, 1)$ is obtained by removing both the all-'1' and all-'0' words from $\mathcal{Q}^n$, where $\mathcal{Q} = \{0, 1\}$. Hence, indeed,

$$N_1(2, n) = 2^n - 1$$

and

$$N_2(2, n) = 2^n - 2.$$

The 2-constrained code $\mathcal{S}_{q,n}(0, q - 1)$ is a Pearson code as it satisfies Properties A and B [3]. There are more examples of 2-constrained sets that are Pearson codes, such as $\mathcal{S}_{q,n}(0, 1)$. Note, however, that not all 2-constrained sets are Pearson codes. For example, $\mathcal{S}_{q,n}(0, 2)$ does not satisfy Property A if $q \geq 5$, since, e.g., both $(0, 1, 2, \ldots, 2)$ and $(0, 2, 4, \ldots, 4) = 2 \times (0, 1, 2, \ldots, 2)$ are codewords

It is obvious from Property B that the code $\mathcal{S}_{2,n}(0, 1)$ of size $2^n - 2$ is the optimal binary Pearson code. For the ternary case, $q = 3$, it can easily be argued that $\mathcal{S}_{3,n}(0, 1)$, $\mathcal{S}_{3,n}(0, 2)$, and $\mathcal{S}_{3,n}(1, 2)$ are all optimal Pearson codes of size $3^n - 2^{n+1} + 1$.

However, for $q > 3$ the 2-constrained sets such as $\mathcal{S}_{q,n}(0, 1)$, $\mathcal{S}_{q,n}(0, q-1)$, and $\mathcal{S}_{q,n}(q - 2, q - 1)$, all of size $N_2(q, n)$, are not optimal Pearson codes, except when $n = 2$. For example, for $q = 4$, it can be easily

checked that the set $\mathcal{S}_{4,n}(0,3) \cup \mathcal{S}_{3,n}(0,1,2)$ is a Pearson code. Its size equals $N_2(4,n) + N_3(3,n) = 4^n - 3^n - 2^{n+1} + 3$, which is larger than $N_2(4,n)$ and actually turns out to be the maximum possible size of any Pearson code for $q = 4$, as shown in the next section, where we will address the problem of constructing optimal Pearson codes for any value of $q$.

# 4 Optimal Pearson codes

For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{Q}^n$, let $m(\boldsymbol{x})$ and $M(\boldsymbol{x})$ denote the smallest and largest value, respectively, among the $x_i$. Furthermore, in case $\boldsymbol{x}$ is not the all-zero word, let $\mathrm{GCD}(\boldsymbol{x})$ denote the greatest common divisor of the $x_i$. For integers $n, q \geq 2$, let $\mathcal{P}_{q,n}$ denote the set of all $q$-ary sequences $\boldsymbol{x}$ of length $n$ satisfying the following properties:

1. $m(\boldsymbol{x}) = 0$;

2. $M(\boldsymbol{x}) > 0$;

3. $\mathrm{GCD}(\boldsymbol{x}) = 1$.

**Theorem 1.** *For any $n, q \geq 2$, $\mathcal{P}_{q,n}$ is an optimal Pearson code.*

*Proof.* We will first show that $\mathcal{P}_{q,n}$ is a Pearson code. Property B is satisfied since any word in $\mathcal{P}_{q,n}$ contains at least one '0' and at least one symbol unequal to '0'. It can be shown that Property A holds by supposing that $\boldsymbol{x} \in \mathcal{P}_{q,n}$ and $\hat{\boldsymbol{x}} = c_1 + c_2\boldsymbol{x} \in \mathcal{P}_{q,n}$ for some $c_1, c_2 \in \mathbb{R}$ with $c_2 > 0$. Clearly $c_1 = 0$, since $c_1 \neq 0$ implies that $m(\hat{\boldsymbol{x}}) \neq 0$. Then, since $\hat{\boldsymbol{x}} = c_2\boldsymbol{x}$, we infer that $\mathrm{GCD}(\hat{\boldsymbol{x}}) = c_2 \times \mathrm{GCD}(\boldsymbol{x}) = c_2$. Since, by definition, $\mathrm{GCD}(\hat{\boldsymbol{x}}) = 1$, we have $c_2 = 1$ and conclude $\hat{\boldsymbol{x}} = \boldsymbol{x}$, which proves that also Property A is satisfied. We conclude $\mathcal{P}_{q,n}$ is a Pearson code.

We will now show that $\mathcal{P}_{q,n}$ is the greatest among all Pearson codes. To that end, let $\mathcal{S}$ be any $q$-ary Pearson code of length $n$. We map all $\boldsymbol{x} \in \mathcal{S}$ to $\boldsymbol{x} - m(\boldsymbol{x})$ and call the resulting code $\mathcal{S}'$. Then, we map all words $\boldsymbol{x}'$ in $\mathcal{S}'$ to $\boldsymbol{x}'/\mathrm{GCD}(\boldsymbol{x}')$. Note that both mappings are injective and that all words in the resulting code $\mathcal{S}''$ satisfy Properties 1-3. Hence, $\mathcal{S}''$ of size $|\mathcal{S}|$ is a subset of $\mathcal{P}_{q,n}$, which proves that $\mathcal{P}_{q,n}$ is optimal. $\square$

From the definitions of $T$-constrained sets and $\mathcal{P}_{q,n}$ it follows that

$$\mathcal{S}_{q,n}(0,1) \subseteq \mathcal{P}_{q,n} \subseteq \mathcal{S}_{q,n}(0). \tag{9}$$

In the following subsections, we will consider the cardinality and redundancy of $\mathcal{P}_{q,n}$, and compare these to the corresponding results for $T$-constrained codes.

## 4.1 Cardinality

In this subsection, we study the size $P_{q,n}$ of $\mathcal{P}_{q,n}$. From (9) and the remark following (8), we have

$$N_2(q,n) \le P_{q,n} \le N_1(q,n). \tag{10}$$

From Property B we have the trivial upper bound

$$P_{q,n} \le q^n - q, \tag{11}$$

which is tight in case $q = 2$ as indicated in Section 3, i.e.,

$$P_{2,n} = 2^n - 2. \tag{12}$$

In order to present expressions for larger values of $q$, we first prove the following lemma. We define $P_{1,n} = 0$.

**Lemma 1.** *For any $n \ge 2$ and $q \ge 3$,*

$$\sum_{\substack{i=2 \\ i-1|q-1}}^{q} (P_{i,n} - P_{i-1,n}) = q^n - 2(q-1)^n + (q-2)^n, \tag{13}$$

*where the summation is over all integers $i$ in the indicated range such that $i - 1$ is a divisor of $q - 1$.*

*Proof.* For each $i$ such that $2 \le i \le q$ and $i - 1$ is a divisor of $q - 1$, we define $\mathcal{D}_{i,n}$ as the set of all $i$-ary sequences $\boldsymbol{y}$ of length $n$ satisfying $m(\boldsymbol{y}) = 0$, $M(\boldsymbol{y}) = i - 1$, and $\mathrm{GCD}(\boldsymbol{y}) = 1$. Let $\mathcal{D}$ denote the union of all these disjoint $\mathcal{D}_{i,n}$.

8

The mapping $\psi$ from $\mathcal{S}_{q,n}(0, q-1)$ to $\mathcal{D}$, defined by dividing $\boldsymbol{x} \in \mathcal{S}_{q,n}(0, q-1)$ by $\mathrm{GCD}(\boldsymbol{x})$, is a bijection. This follows by observing that, on one hand, $\psi(\boldsymbol{x})$ is a unique member of $\mathcal{D}_{(q-1)/\mathrm{GCD}(\boldsymbol{x})+1,n}$, while, on the other hand, any sequence in $\boldsymbol{y} \in \mathcal{D}_{i,n}$ is the image of $((q-1)/(i-1))\boldsymbol{y} \in \mathcal{S}_{q,n}(0, q-1)$ under $\psi$.

Finally, the lemma follows by observing that $|\mathcal{D}_{i,n}| = P_{i,n} - P_{i-1,n}$ and $|\mathcal{S}_{q,n}(0, q-1)| = N_2(q,n) = q^n - 2(q-1)^n + (q-2)^n$. $\qquad\square$

We thus have with (13) a recursive expression for $P_{q,n}$. Starting from the result for $q = 2$ in (12), we can find $P_{q,n}$ for any $n$ and $q$. Expressions for $2 \leq q \leq 8$ of the size of optimal Pearson codes, $P_{q,n}$, are tabulated in Table 1. The next theorem offers a closed formula for the size of optimal Pearson codes, $P_{q,n}$. We start with a definition.

For a positive integer $d$, the Möbius function $\mu(d)$ is defined [10, 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 $d$.

**Theorem 2.** *Let $n$ and $q$ be positive integers. Let $P_{q,n}$ be the cardinality of a $q$-ary Pearson code of length $n$. Then*

$$P_{q,n} = \sum_{d=1}^{q-1} \mu(d) \left( \left( \left\lfloor \frac{q-1}{d} \right\rfloor + 1 \right)^n - \left\lfloor \frac{q-1}{d} \right\rfloor^n - 1 \right). \qquad (14)$$

We use the following well-known theorem (see [10, Section 16.5], for example) in our proof of Theorem 2.

**Theorem 3.** *Let $F : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$ be functions such that*

$$G(x) = \sum_{d=1}^{\lfloor x \rfloor} F(x/d)$$

*for all positive $x$. Then*

$$F(x) = \sum_{d=1}^{\lfloor x \rfloor} \mu(d) G(x/d). \qquad (15)$$

*Proof (of Theorem 2).* For a non-negative real number $x$, define

$$I_x = \{0, 1, \ldots, \lfloor x \rfloor\} = \mathbb{Z} \cap [0, x].$$

Let $V_x$ be the set of vectors of length $n$ with entries in $I_x$ and with at least one zero entry and at least one non-zero entry. Define $G(x) = |V_x|$. To determine $G(x)$, note that there are $|I_x|^n$ length $n$ vectors with entries in $I_x$, and we must exclude the all-zero vector and the $(|I_x| - 1)^n$ vectors with no zero entries. Since $|I_x| = \lfloor x \rfloor + 1$, we find that

$$G(x) = |I_x|^n - (|I_x| - 1)^n - 1 = (\lfloor x \rfloor + 1)^n - \lfloor x \rfloor^n - 1. \qquad (16)$$

For a positive integer $d$, let $V_{x,d}$ be the set of vectors $\mathbf{c} \in V_x$ such that $\mathrm{GCD}(\mathbf{c}) = d$. Since $\mathbf{c} \neq \mathbf{0}$, we see that $1 \leq \mathrm{GCD}(\mathbf{c}) \leq \max_i\{c_i\} \leq \lfloor x \rfloor$ and so $V_x$ can be written as the disjoint union

$$V_x = \bigcup_{d=1}^{\lfloor x \rfloor} V_{x,d}.$$

Moreover, $|V_{x,d}| = |V_{x/d,1}|$, since the map taking $\mathbf{c} \in V_{x,d}$ to $(1/d)\mathbf{c} \in V_{x/d,1}$ is a bijection.

Define $F(x) = |V_{x,1}|$, so $F(x)$ is the number of vectors $\mathbf{c} \in V_x$ such that $\mathrm{GCD}(\mathbf{c}) = 1$. Now,

$$G(x) = |V_x| = \sum_{d=1}^{\lfloor x \rfloor} |V_{x,d}| = \sum_{d=1}^{\lfloor x \rfloor} |V_{x/d,1}| = \sum_{d=1}^{\lfloor x \rfloor} F(x/d).$$

So, by Theorem 3, we deduce that (15) holds. Theorem 2 now follows from the fact that $P_{q,n} = F(q - 1)$, by combining (15) and (16). $\qquad \square$

After perusing Table 1, it appears that for $q \geq 4$, $P_{q,n}$ is roughly $q^n - (q - 1)^n$. An intuitive justification is that among the $q^n$ $q$-ary sequences of length $n$ there are $(q - 1)^n$ sequences that do not contain 0, which is the most significant condition to avoid. All this is confirmed by the next corollary.

**Corollary 1.** *For any positive integer $q$, we have that*

$$P_{q,n} = q^n - (q - 1)^n + O(\lceil q/2 \rceil^n)$$

*as $n \to \infty$.*

Table 1: Size of optimal Pearson codes, $P_{q,n}$, for $2 \leq q \leq 8$.

| $q$ | $P_{q,n}$ |
|---|---|
| 2 | $2^n - 2$ |
| 3 | $3^n - 2^{n+1} + 1$ |
| 4 | $4^n - 3^n - 2^{n+1} + 3$ |
| 5 | $5^n - 4^n - 3^n + 2$ |
| 6 | $6^n - 5^n - 3^n - 2^n + 4$ |
| 7 | $7^n - 6^n - 4^n + 2^n + 1$ |
| 8 | $8^n - 7^n - 4^n + 3$ |

*Proof.* The $d = 1$ term in the sum on the right hand side of (14) is $q^n - (q-1)^n$, and the absolute values of remaining terms are each bounded by $\lceil q/2 \rceil^n$, since

$$\lfloor (q-1)/d \rfloor + 1 \leq \lfloor (q-1)/2 \rfloor + 1 = \lceil q/2 \rceil. \qquad \square$$

As discussed above, the 2-constrained codes $\mathcal{S}_{q,n}(0, 1)$ and $\mathcal{S}_{q,n}(0, q-1)$ are Pearson codes. Therefore, it is of interest to compare $P_{q,n}$ with the cardinality $N_2(q, n)$ of 2-constrained codes. For $q \leq 3$, we simply have $\mathcal{S}_{q,n}(0, 1) = \mathcal{P}_{q,n}$, and thus $N_2(q, n) = P_{q,n}$. However, for $q \geq 4$, we infer from (8), i.e., $N_2(q, n) = q^n - 2(q - 1)^n + (q - 2)^n$, and Corollary 1, i.e., $P_{q,n} = q^n - (q - 1)^n + O(\lceil q/2 \rceil^n)$ that $N_2(q, n) < P_{q,n}$, with a possible exception for very small values of $n$. For all $q \geq 2$,

$$P_{q,2} = N_2(q, 2) = 2 \tag{17}$$

and it is not hard to show that

$$P_{q,3} = 6 \sum_{j=1}^{q-1} \phi(j), \tag{18}$$

where $\phi(j)$ is Euler's totient function that counts the totatives of $j$, i.e., the positive integers less than or equal to $j$ that are relatively prime to $j$.

We have computed the cardinalities of $N_1(q, n)$, $N_2(q, n)$, and $P_{q,n}$ by invoking (7), (8), and the expressions in Table 1. Table 2 lists the results of our computations for selected values of $q$ and $n$.

Table 2: $N_2(q,n)$, $P_{q,n}$, and $N_1(q,n)$ for selected values of $q$ and $n$.

| $n$ | $q$ | $N_2(q,n)$ | $P_{q,n}$ | $N_1(q,n)$ |
|---|---|---|---|---|
| 4 | 4 | 110 | 146 | 175 |
| 4 | 5 | 194 | 290 | 369 |
| 4 | 6 | 302 | 578 | 671 |
| 5 | 4 | 570 | 720 | 781 |
| 5 | 5 | 1320 | 1860 | 2101 |
| 5 | 6 | 2550 | 4380 | 4651 |
| 6 | 4 | 2702 | 3242 | 3367 |
| 6 | 5 | 8162 | 10802 | 11529 |
| 6 | 6 | 19502 | 30242 | 31031 |
| 7 | 4 | 12138 | 13944 | 14197 |
| 7 | 5 | 47544 | 59556 | 61741 |
| 7 | 6 | 140070 | 199500 | 201811 |

## 4.2   Redundancy

As usual, the redundancy of a $q$-ary code $\mathcal{C}$ of length $n$ is defined by $n - \log_q |\mathcal{C}|$. From (7), it follows that the redundancy of a 1-constrained code is

$$
\begin{aligned}
r_1 &= n - \log_q(q^n - (q-1)^n) \\
&= -\log_q\left(1 - \left(\frac{q-1}{q}\right)^n\right) \\
&\approx \left(\frac{q-1}{q}\right)^n \bigg/ \ln(q),
\end{aligned}
\tag{19}
$$

for $n$ sufficiently large, where the approximation follows from the well-known fact that $\ln(1+a) \approx a$ when $a$ is close to 0. Similarly, from (8)

we infer the redundancy of a 2-constrained code, namely

$$
\begin{aligned}
r_2 &= n - \log_q(q^n - 2(q-1)^n + (q-2)^n) \\
&= -\log_q\left(1 - 2\left(\frac{q-1}{q}\right)^n + \left(\frac{q-2}{q}\right)^n\right) \\
&\approx \left(2\left(\frac{q-1}{q}\right)^n - \left(\frac{q-2}{q}\right)^n\right) \bigg/ \ln(q)
\end{aligned}
\tag{20}
$$

for $n$ sufficiently large. Since the 2-constrained code $\mathcal{S}_{q,n}(0,1)$ is optimal for $q = 2, 3$, the expression for $r_2$ gives the minimum redundancy for any binary or ternary Pearson code. From Corollary 1, it follows for $q \geq 4$ that the redundancy of optimal Pearson codes equals

$$
\begin{aligned}
r_{\mathrm{P}} &= n - \log_q\left(q^n - (q-1)^n + O\left(\left(\frac{q+1}{2}\right)^n\right)\right) \\
&= -\log_q\left(1 - \left(\frac{q-1}{q}\right)^n + O\left(\left(\frac{q+1}{2q}\right)^n\right)\right) \\
&\approx \left(\left(\frac{q-1}{q}\right)^n + O\left(\left(\frac{q+1}{2q}\right)^n\right)\right) \bigg/ \ln(q).
\end{aligned}
\tag{21}
$$

In conclusion, for sufficiently large $n$, we have

$$
r_P = r_2 \approx 2r_1
\tag{22}
$$

if $q = 2, 3$, while

$$
r_P \approx r_1 \approx r_2/2
\tag{23}
$$

if $q \geq 4$. Figure 1 shows, as an example, the redundancies $r_1$, $r_2$, and $r_P$ versus $n$ for $q = 8$ (the quantity $r_P$ was computed using the expression listed in Table 1). Note that the redundancy $r_2$ decreases while the redundancy of prior art balanced codes, $r_0$, see (1), increases with increasing codeword length $n$. The curve $r_0$ versus $n$ was not plotted in Figure 1 as the redundancy of balanced codes is much higher than that of Pearson codes. For example, an evaluation of (1) shows that the redundancy $r_0 = 2.79$ for $q = 8$ and $n = 10$, while $r_P = 0.147$ for the same parameters.
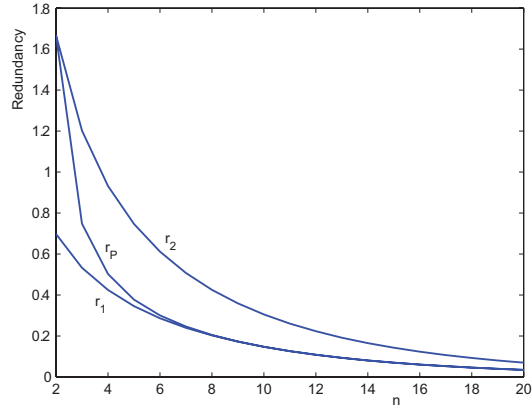
13

Figure 1: Redundancy $r_1$, $r_2$, and $r_P$ versus $n$ for $q = 8$.

# 5    Conclusions

We have studied sets of $q$-ary codewords of length $n$, coined Pearson codes, that can be detected unambiguously by a detector based on the Pearson distance. We have formulated the properties of codewords in Pearson codes. We have presented constructions of optimal Pearson codes and evaluated their cardinalities and redundancies. We conclude that, except for small values of $q$ and/or $n$, the redundancy of optimal Pearson codes is almost the same as the redundancy of 1-constrained codes.

# References

[1] 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.

[2] 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.

[3] 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.

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

[5] F. Sala, R. Gabrys, and L. Dolecek, "Dynamic Threshold Schemes for Multi-Level Non-Volatile Memories", *IEEE Trans. on Commun.,* pp. 2624-2634, Vol. 61, July 2013.

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

[7] 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 (1), pp. 66-73, Jan. 2015.

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

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

[10] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers,* (5th Edition), Oxford University Press, Oxford, 1979.

**Jos H. Weber** (S'87-M'90-SM'00) was born in Schiedam, The Netherlands, in 1961. He received the M.Sc. (in mathematics, with honors), Ph.D., and MBT (Master of Business Telecommunications) degrees from Delft University of Technology, Delft, The Netherlands, in 1985, 1989, and 1996, respectively.

Since 1985 he has been with the Faculty of Electrical Engineering, Mathematics, and Computer Science of Delft University of Technology. Currently, he is an associate professor in the Department of Intelligent Systems. He is the chairman of the WIC (Werkgemeenschap voor Informatie- en Communicatietheorie in the Benelux) and the secretary of the IEEE Benelux Chapter on Information Theory. He was a Visiting Researcher at the University of California at Davis, USA, the University of Johannesburg, South Africa, the Tokyo Institute of Technology, Japan, and EPFL, Switzerland. His main research interests are in the areas of channel and network coding, security, and (quantum) information theory.

**Kees Schouhamer Immink** (M'81-SM'86-F'90) received his PhD degree from the Eindhoven University of Technology. In 1998, he founded Turing Machines Inc, which has been successful in applying the tenets of information theory to digital data storage and transmission. He was from 1994 till 2014 an adjunct professor at the Institute for Experimental Mathematics, Essen-Duisburg University, Germany. In 1998, he founded Turing Machines Inc., an innovative start-up focused on novel signal processing for hard disk drives and solid-state (Flash) memories, where he currently holds the position of president.

Immink designed coding techniques of digital audio and video recording products such as Compact Disc, CD-ROM, CD-Video, Digital Audio Tape recorder, Digital Compact Cassette (DCC), Digital Versatile Disc (DVD), Video Disc Recorder, and Blu-ray Disc. He received a Knighthood in 2000, a personal Emmy award in 2004, the 1996 IEEE Masaru Ibuka Consumer Electronics Award, the 1998 IEEE Edison Medal, the 1999 AES Gold Medal, the 2004 SMPTE Progress Medal, the 2014 Eduard Rhein Prize for Technology , and the 2015 IET Faraday Medal. He received the Golden Jubilee Award for Technological Innovation by the IEEE Information Theory Society in 1998. He was named a fellow of the IEEE, AES, and SMPTE, and was inducted into the Consumer Electronics Hall of Fame, elected into the Royal Netherlands Academy of Sciences, the Royal Holland Society of Sciences, and the (US) National Academy of Engineering. He received an honorary doctorate from the University of Johannesburg in 2014. He served the profession as President of the

Audio Engineering Society inc., New York, in 2003.

**Simon R. Blackburn** (M'12) was born in Beverley, Yorkshire, England in 1968. He received a BSc in Mathematics from Bristol in 1989, and a DPhil in Mathematics from Oxford in 1992. He has worked in the Mathematics Department at Royal Holloway University of London since 1992, and is currently a Professor of Pure Mathematics. His research interests include algebra, combinatorics and associated applications in cryptography and communication theory.