Pearson Codes
Jos H. Weber and Kees A. Schouhamer Immink

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 codewords, called Pearson codes, that satisfy specific properties. We will analyze constructions and properties of Pearson codes.

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

I. 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 [4], [5], [6]. In optical disc media, such as the popular Compact Disc, DVD, and Blu-ray disc, the received 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 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 $S$ be a codebook of chosen $q$-ary codewords $x = (x_1, x_2, \ldots, x_n)$ over the $q$-ary alphabet $Q = \{0, 1, \ldots, q - 1\}$, $q \geq 2$, where $n$, the length of $x$, is a positive integer. A Pearson code with maximum possible size given the parameters $q$ and $n$ is said to be optimal.

In Section II, we set the stage with a description of the prior art. Section III gives a description of $T$-constrained sequences, while Section IV offers a general construction of optimal Pearson codes and a computation of their cardinalities. In Section V, we will describe our conclusions.

II. PRELIMINARIES

The Pearson distance between the vectors $x$ and $\hat{x}$ is defined by [3]

$$\delta(x, \hat{x}) = 1 - \rho_{x, \hat{x}},$$

where

$$\rho_{x, \hat{x}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\hat{x}_i - \bar{x})}{\sigma_x \sigma_{\hat{x}}}$$

is the (Pearson) correlation coefficient, and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

and

$$\sigma_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Note that since $|\rho_{x, \hat{x}}| \leq 1$, we infer that $0 \leq \delta(x, \hat{x}) \leq 2$. The codewords $x$ and $\hat{x}$ are taken from a codebook $S \subseteq Q^n$ that must guarantee unambiguous detection with the Pearson distance metric (1). In other words, all pairs of codewords $x$ and $\hat{x} \in S$ satisfy

$$\delta(x, \hat{x}) \neq 0 \text{ for } x \neq \hat{x}, \quad x, \hat{x} \in S.$$

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

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

if and only if

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

where the coefficients $c_1$ and $c_2 > 0$ are real numbers, and we use the shorthand notation $x + b = (x_1 + b, x_2 + b, \ldots, x_n + b)$. It is further immediate, see (2), that the Pearson distance is undefined for codewords $x$ with $\sigma_x = 0$. We conclude that codewords in a Pearson code must satisfy two conditions:

- **Property A:** If $x \in S$ then $c_1 + c_2 x \notin S$ for all $c_1, c_2 \in \mathbb{R}$ with $(c_1, c_2) \neq (0, 1)$ and $c_2 > 0$.
- **Property B:** $x = c \notin S$ 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 which 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 $T$-constrained codes that, as we will see shortly, play a role in the construction of Pearson codes.
III. T-CONSTRAINED CODES

T-constrained codes [1], denoted by \( S_{q,n}(a_1, \ldots, a_T) \), consist of \( q \)-ary \( n \)-length codewords, where \( T, 0 < T \leq q, \) preferred or reference symbols \( a_1, \ldots, a_T \in Q, \) must appear at least once in a codeword. Thus, each codeword, \((x_1, x_2, \ldots, x_n)\), in a \( T \)-constrained code satisfies

\[ \{i : x_i = j\} > 0 \text{ for } j \in \{a_1, \ldots, a_T\}. \]

The number of \( n \)-length \( q \)-ary sequences, \( N_T(q, n) \), where \( T, T \leq q, \) distinct symbols occur at least once in the \( n \)-sequence, equals [1]

\[ N_T(q, n) = \sum_{i=0}^{T} (-1)^i \binom{T}{i} (q - i)^n, \ n \geq T. \]  

(5)

For example, we easily find

\[ N_1(q, n) = q^n - (q - 1)^n \]  

and

\[ N_2(q, n) = q^n - 2(q - 1)^n + (q - 2)^n. \]  

(7)

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 \( S_{2,n}(0) \) is obtained by removing the all-‘1’ word from \( Q^n \), that \( S_{2,n}(1) \) is obtained by removing the all-‘0’ word from \( Q^n \), and that \( S_{2,n}(0,1) \) is obtained by removing both the all-‘1’ and all-‘0’ words from \( Q^n \), where \( Q = \{0,1\} \). Hence, indeed, \( N_1(2, n) = 2^n - 1 \) and \( N_2(2, n) = 2^n - 2 \).

The 2-constrained code \( S_{q,n}(0, q - 1) \) 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 \( S_{q,n}(0,1) \). Note, however, that not all 2-constrained sets are Pearson codes. For example, \( S_{q,n}(0,2) \) 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 \( S_{2,n}(0,1) \) of size \( 2^n - 2 \) is the optimal binary Pearson code. For the ternary case, \( q = 3 \), it can easily be argued that \( S_{3,n}(0,1) \), \( S_{3,n}(0,2) \), and \( S_{3,n}(1,2) \) are all optimal Pearson codes of size \( 3^n - 2^{n+1} + 1 \).

However, for \( q > 3 \) and \( n > 2 \), the 2-constrained sets such as \( S_{q,n}(0,1) \), \( S_{q,n}(0, q - 1) \), and \( S_{q,n}(q - 2, q - 1) \) are not optimal Pearson codes. For example, for \( q = 4, \) it can be easily checked that the set \( S_{4,n}(0,3) \cup S_{3,n}(0,1,2) \) is a Pearson code. Its size equals \( N_2(4, n) + N_3(3, n) = 4^n - 3^n - 2^{n+1} + 3 \), which 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 \( q \).

IV. OPTIMAL PEARSON CODES

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

1) \( m(x) = 0; \)
2) \( M(x) > 0; \)
3) \( \text{GCD}(x) = 1. \)

**Theorem 1:** For any \( n, q \geq 2, P_{q,n} \) is an optimal Pearson code.

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

We will now show that \( P_{q,n} \) is the greatest among all Pearson codes. To that end, let \( S \) be any \( q \)-ary Pearson code of length \( n \). We substitute all \( x \in S \) by \( x - m(x) \) and call the resulting code \( S' \). Then, we replace all words \( x' \in S' \) by \( x' / \text{GCD}(x'). \) All words in the resulting code \( S'' \) satisfy Properties 1)-3), and thus \( S'' \) of size \( |S| \) is a subset of \( P_{q,n} \), which proves that \( P_{q,n} \) is optimal.

From the definitions of \( T \)-constrained sets and \( P_{q,n} \) it follows that

\[ S_{q,n}(0,1) \subseteq P_{q,n} \subseteq S_{q,n}(0). \]  

(8)

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

A. Cardinality

In this subsection, we study the size \( P_{q,n} \) of \( P_{q,n} \). From (8), we have

\[ N_2(q, n) \leq P_{q,n} \leq N_1(q, n). \]  

(9)

From Property B we have the trivial upper bound

\[ P_{q,n} \leq q^n - q. \]  

(10)

which is tight in case \( q = 2 \) as indicated in Section III, i.e.,

\[ P_{2,n} = 2^n - 2. \]  

(11)

In order to present expressions for larger values of \( q, \) we first prove the following lemma.

**Lemma 1:** For any \( n \geq 2 \) and \( q \geq 3, \)

\[ \sum_{i=2}^{q} (P_{i,n} - P_{i-1,n}) = q^n - 2(q - 1)^n + (q - 2)^n, \]  

(12)

where the summation is over all integers \( i \) in the indicated range such that \( i - 1 \) is a divisor of \( q - 1, \) and where \( P_{1,n} = 0. \)
As discussed above, the 2-constrained codes $S$ of $P$ have a recursive expression for $P$. We thus have a recursive expression for $P$, and (13) that $N_2(q, n) = q^n - 2(q - 1)^n + (q - 2)^n$.

We have computed the cardinalities of the various sets, $N_1(q, n), N_2(q, n), P_{q,n}$ by invoking (6), (7), and the expressions in Table I. Table II lists the results of our computations for selected values of $q$ and $n$.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$P_{q,n}$</th>
<th>$N_1(q, n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$2^n - 2$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$3^n - 2^n + 1$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$4^n - 3^n - 2^n + 3$</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$5^n - 4^n - 3^n + 2$</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$6^n - 5^n - 3^n - 2^n + 4$</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$7^n - 6^n - 4^n + 2^n + 1$</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$8^n - 7^n - 4^n + 3$</td>
<td></td>
</tr>
</tbody>
</table>

**Proof.** For each $i$ such that $2 \leq i \leq q$ and $i - 1$ is a divisor of $q - 1$, we define $D_{i,n}$ as the set of all $i$-ary sequences $y$ of length $n$ satisfying $m(y) = 0, M(y) = i - 1$, and $\text{GCD}(y) = 1$. Let $D$ denote the union of all these $D_{i,n}$. Note that the mapping from $S_{q,n}(0, q - 1)$ to $D$, defined by dividing $x \in S_{q,n}(0, q - 1)$ by $\text{GCD}(x)$, is a bijection. The lemma follows by observing that $|D_{i,n}| = P_{i,n} - P_{i-1,n}$ and $|S_{q,n}(0, q - 1)| = N_2(q, n) = q^n - 2(q - 1)^n + (q - 2)^n$.

We thus have a recursive expression for $P_{q,n}$. Starting from the result for $q = 2$ in (11), we can find an explicit function of $P_{q,n}$ for any $n$ and $q$. The expressions of the size of optimal Pearson codes, $P_{q,n}$, for $2 \leq q \leq 8$ are tabulated in Table I. After perusing the above table, it is clear that for $q \geq 4$, $P_{q,n}$ is roughly $q^n - (q - 1)^n$, which is confirmed by the next theorem.

**Theorem 2:** For any $q \geq 4$,

$$P_{q,n} = q^n - (q - 1)^n + O(\lfloor q/2 \rfloor^n).$$

**Proof.** From (12) and the observation that the largest integer $i < q$ such that $i - 1$ is a divisor of $q - 1$ is at most $\lfloor q/2 \rfloor$, we have

$$P_{q,n} = P_{q-1,n} - \sum_{i=2}^{\lfloor q/2 \rfloor} (P_{i,n} - P_{i-1,n})$$

$$= q^n - 2(q - 1)^n + (q - 2)^n$$

$$= q^n - (q - 1)^n + O(\lfloor q/2 \rfloor^n).$$

B. Redundancy

As usual, the redundancy of a $q$-ary code $C$ of length $n$ is defined by $n - \log_q |C|$. From (6), it follows that the redundancy of a 1-constrained code is

$$r_1 = n - \log_q q^n - (q - 1)^n$$

$$= -\log_q \left(1 - \left(\frac{q - 1}{q}\right)^n\right)$$

$$\approx \left(\frac{q - 1}{q}\right)^n / \ln(q),$$

where the approximation follows from the well-known fact that $\ln(1 + a) \approx a$ when $a$ is close to 0. Similarly, from (7) we infer the redundancy of a 2-constrained code, namely

$$r_2 = n - \log_q (q^n - 2(q - 1)^n + (q - 2)^n)$$

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

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

Since the 2-constrained code $S_{q,n}(0, 1)$ is optimal for $q = 2$, the expression for $r_2$ gives the minimum redundancy for any binary or ternary Pearson code. From Theorem 2, it follows for $q \geq 4$ that the redundancy of optimal Pearson
codes equals
\[
    r_P = n - \log_q \left( q^n - (q-1)^n + O \left( \left( \frac{q+1}{2} \right)^n \right) \right)
    = - \log_q \left( 1 - \left( \frac{q-1}{q} \right)^n + O \left( \frac{g+1}{2q} \right)^n \right)
    \approx \left( \left( \frac{q-1}{q} \right)^n + O \left( \frac{g+1}{2q} \right)^n \right) / \ln(q).
\]
(19)

In conclusion, for sufficiently large $n$, we have
\[
    r_P = r_2 \approx 2r_1
\]
(20)

if $q = 2,3$, while
\[
    r_P \approx r_1 \approx r_2/2
\]
(21)

if $q \geq 4$. Figure 1 shows, as an example, the redundancies $r_1$, $r_2$, and $r_P$ versus $n$ for $q = 8$.

V. 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