Bounds and Constructions of Codes over Symbol-Pair Read Channels

Ohad Elishco[*], **Ryan Gabrys**[†], and **Eitan Yaakobi**[‡]

[*] Massachusetts Institute of Technology, MA, 02139

[†] Spawar Systems Center, San Diego, CA, 92115

[‡] Department of Computer Science, Technion — Israel Institute of Technology, Haifa, 32000 Israel

Emails: ohadeli@mit.edu, ryan.gabrys@navy.mil, yaakobi@cs.technion.ac.il

***Abstract*—Cassuto and Blaum recently studied the symbol-pair channel, a model where every two consecutive symbols are read together. This special structure of channels is motivated by the limitations of the reading process in high density data storage systems, where it is no longer possible to read individual symbols. In this new paradigm, the errors are no longer individual symbol errors, but rather *symbol-pair errors*, where at least one of the symbols is erroneous.**

In this work, we study bounds and construction of codes over the symbol-pair channels. We extend the Johnson bound and the linear programming bound for this channel and show that they improve upon existing bounds. We then propose new code constructions that improve upon existing results that use linear cyclic codes when the pair distance is between four and ten.

I. Introduction

High density data storage systems carry a basic limitation in which the outputs of the reading process are pairs of consecutive symbols instead of individual symbols. The symbol-pair read channel was recently proposed as a model reflecting this limitation and was first studied in [2] and [3]. In those papers the authors studied the fundamental questions arising from pair-symbol readings such as the pair-distance, code constructions, decoding of error-correction codes, and bounds on codes size. These results were later extended in several directions such as cyclic codes, maximum distance separable (MDS) codes, decoding algorithms, and more.

The results in [2] and [3] were extended in [18] for studying linear cyclic codes and their decoding algorithms to correct symbol-pair errors. Several more works presented different decoding algorithms for arbitrary linear codes [8], [9], [13], [15]–[17]. The study of MDS codes for the symbol-pair channel was initiated in [4], and was later extended in several more works and for other non-binary codes; see e.g. [5], [6], [11], [12], [14], [19]. Another generalization of the pair-symbol model was studied in [18] for the b-symbol read channel. Here the assumption is that every $b > 2$ consecutive symbols are read together. This model was further studied for MDS codes in [5], [12].

Assume the stored word is given by the vector $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1})$. The *pair-read vector* is given by

$$\pi(\boldsymbol{x}) = ((x_0, x_1), (x_1, x_2), \ldots, (x_{n-2}, x_{n-1}), (x_{n-1}, x_0)).$$

A *symbol-pair error* is the event where at least one of the symbols in the read pair is in error. The *pair distance* between two words $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted by $d_p(\boldsymbol{x}, \boldsymbol{y})$, is the Hamming distance between their pair-read vectors, that is $d_p(\boldsymbol{x}, \boldsymbol{y}) = d_H(\pi(\boldsymbol{x}), \pi(\boldsymbol{y}))$. Finally, the *minimum pair distance* of a code $\mathcal{C}$ is the minimum pair distance between any two different codewords. Under this paradigm, the ultimate goal is the construction of efficient codes with large minimum pair distance since this is the appropriate figure of merit to study in order to correct symbol-pair errors, that is, a code with minimum pair distance d_p permits the correction of at least $\lfloor \frac{d_p - 1}{2} \rfloor$ symbol-pair errors.

In [18], it was shown that if a linear cyclic code has minimum Hamming distance d_H then its minimum pair distance is at least $d_p \geqslant \lceil 3d_H/2 \rceil$. This work presented also decoding algorithms for such codes. On the other hand, bounds on codes with minimum pair distance were studied in [2], where the authors extended the sphere packing bound for symbol-pair errors.

In this work we improve upon existing results and propose new upper bounds and code constructions of codes for the symbol-pair read channel. Specifically, we show how to extend the Johnson bound and the linear programming bound for this setup and show that the new bounds improve upon the best previously known bound from [2]. We also study code constructions for relatively small minimum pair distance, namely between four and ten. For these cases we show how to improve the result from [18] that uses cyclic linear codes in order to receive codes with better redundancy.

The rest of this paper is organized as follows. In Section II, we review the symbol-pair read channel and list several basic properties that will be used throughout the paper. In Section III, we study bounds on codes correcting symbol-pair errors. Then, in Section IV we present our new code constructions of symbol-pair error-correcting codes, when the minimum pair-distance is between four and ten. Due to the lack of space, some proofs in the paper are omitted.

II. Preliminaries

Let $n \in \mathbb{N}$ (the natural numbers, including 0), and denote by $[n]$ the set $\{0, \ldots, n-1\}$. Let Σ be the binary alphabet and denote by Σ^n the set of all length-n sequences over Σ. We use $\mathbb{F}_q$ to denote the field of size q. For a sequence $\boldsymbol{x} \in \Sigma^n$ denote by $w_H(\boldsymbol{x})$ the Hamming weight of $\boldsymbol{x}$. For two sequences $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$ let $d_H(\boldsymbol{x}, \boldsymbol{y})$ denote the Hamming distance between $\boldsymbol{x}, \boldsymbol{y}$. For a set $\mathcal{C} \subseteq \Sigma^n$, denote by $d_H(\mathcal{C}) \triangleq \min_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \boldsymbol{x} \neq \boldsymbol{y}} \{d_H(\boldsymbol{x}, \boldsymbol{y})\}$ the minimum Hamming distance between any two different sequences in $\mathcal{C}$. We also denote by $\mathbf{0}, \mathbf{1} \in \Sigma^n$ the all zeros and the all ones sequences, respectively.

978-1-5386-4780-6/18/$31.00©2018 IEEE

Definition 1. *Let $\pi : \Sigma^n \to (\Sigma \times \Sigma)^n$ denote the (cyclic) pair symbol read representation which is defined as follows. For $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1}) \in \Sigma^n$,*

$$\pi(\boldsymbol{x}) \triangleq ((x_0, x_1), (x_1, x_2), \ldots, (x_{n-2}, x_{n-1}), (x_{n-1}, x_0)).$$

We now define the associated pair-weight and pair-distance. For a sequence $\boldsymbol{x} \in \Sigma^n$, define the pair-weight of $\boldsymbol{x}$ as

$$w_p(\boldsymbol{x}) \triangleq w_H(\pi(\boldsymbol{x})) = \left|\left\{j \in [n] \ : \ (x_j, x_{j+1}) \neq (0,0)\right\}\right|$$

with coordinates taken modulo n. Similarly, for sequences $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$, define the pair-distance as

$$d_p(\boldsymbol{x}, \boldsymbol{y}) \triangleq d_H(\pi(\boldsymbol{x}), \pi(\boldsymbol{y})) = \left|\left\{j \in [n] : (x_j, x_{j+1}) \neq (y_j, y_{j+1})\right\}\right|$$

with coordinates taken modulo n. For a set $\mathcal{C} \subseteq \Sigma^n$ we define $d_p(\mathcal{C})$ as the minimum pair-distance between any two different codewords,

$$d_p(\mathcal{C}) \triangleq \min_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \boldsymbol{x} \neq \boldsymbol{y}} \{d_p(\boldsymbol{x}, \boldsymbol{y})\}.$$

Note that for a linear code, $\mathcal{C}$, we obtain

$$d_p(\mathcal{C}) = \min_{\boldsymbol{y} \in \mathcal{C}, \boldsymbol{y} \neq \boldsymbol{0}} \{d_p(\boldsymbol{0}, \boldsymbol{y})\} = \min_{\boldsymbol{y} \in \mathcal{C}, \boldsymbol{y} \neq \boldsymbol{0}} \{w_p(\boldsymbol{y})\}. \quad (1)$$

Example 1. Let Σ be the binary alphabet and let $n = 4$ and consider the sequences $\boldsymbol{x} = (0110), \boldsymbol{y} = (0101) \in \Sigma^n$. We have that $w_H(\boldsymbol{x}) = w_H(\boldsymbol{y}) = 2$, $d_H(\boldsymbol{x}, \boldsymbol{y}) = 2$, $\pi(\boldsymbol{x}) = ((0,1), (1,1), (1,0), (0,0))$, $w_p(\boldsymbol{x}) = 3$, $\pi(\boldsymbol{y}) = ((0,1), (1,0), (0,1), (1,0))$, $w_p(\boldsymbol{y}) = 4$, $d_p(\boldsymbol{x}, \boldsymbol{y}) = 3$.

We define $r(\boldsymbol{x}) \triangleq |\{i : \pi(\boldsymbol{x})_i = (0,1)\}|$ so that $r(\boldsymbol{x})$ is equal to the number of occurrences of the symbol $(0,1)$ in $\pi(\boldsymbol{x})$. It is straightforward to show that $r(\boldsymbol{x}) = |\{i : \pi(\boldsymbol{x})_i = (0,1)\}| = |\{i : \pi(\boldsymbol{x})_i = (1,0)\}|$. Note that according to our definition, the sequence $\boldsymbol{1}$ has $r(\boldsymbol{1}) = \boldsymbol{0}$. It is known [2] that

$$w_p(\boldsymbol{x}) = w_H(\boldsymbol{x}) + r(\boldsymbol{x}). \quad (2)$$

Moreover, if $\mathcal{C}$ is a linear code, from (1), it is straightforward to verify that

$$d_p(\mathcal{C}) = \min_{\boldsymbol{y} \in \mathcal{C}, \boldsymbol{y} \neq 0} \{w_H(\boldsymbol{y}) + r(\boldsymbol{y})\}. \quad (3)$$

In this work we focus on bounds and constructions for codes that, for fixed levels of redundancy, maximize the ***pair-distance*** $d_p(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$. We make use of the following lemma from [18] which we include for completeness.

Lemma 2. *[18, Lemma 1] Suppose $\mathcal{C} \subseteq \mathbb{F}_2^n$ is a cyclic linear code where $d_H(\mathcal{C}) \geqslant d$. Then, for any $\boldsymbol{x} \in \mathcal{C}$, $r(\boldsymbol{x}) \geqslant \lceil \frac{d}{2} \rceil$.*

As a consequence of Lemma 2, it follows that if $\mathcal{C}$ is a linear cyclic code with minimum Hamming distance $d_H(\mathcal{C})$, then the code $\mathcal{C}$ satisfies

$$d_p(\mathcal{C}) \geqslant \left\lceil \frac{3 d_H(\mathcal{C})}{2} \right\rceil. \quad (4)$$

III. Upper Bounds

In this section, we derive a number of new bounds on the maximum size of a code with a prescribed pair-distance. In the first subsection, we consider upper bounds for even pair-distance and then in the following subsection we apply linear programming techniques.

A. Upper Bounds for Even Pair-Distance

In this subsection, we derive upper bounds for even pair-distance codes using similar logic as in the Johnson bound [10, Theorem 2.3.8]. The main result of this subsection appears in Theorem 3. In order to prove the theorem, we derive a bound on the maximal size of a code where each codeword in the code has pair-weight w and the pair-distance between any two distinct codewords is at least $2w$. This result is then used to prove Theorem 3.

Let $n, d, w \in \mathbb{N}$ and denote by $A_p(n, d, w)$ the maximal size of a code where each codeword in the code has pair-weight w and the pair-distance between any two distinct codewords is at least d. For the specific case in which $d = 2w$ we have proved that $A_p(n, 2w, w) = \left\lfloor \frac{n}{w} \right\rfloor$.

We introduce some additional notation and useful results from [2]. For integers i, j where $i \leqslant j$, let $[i, j] = \{i, i+1, \ldots, j\}$. For integers $n > \ell \geqslant L$, let $D(n, \ell, L)$ be the number of sequences $\boldsymbol{x} \in \Sigma^n$ such that $w_H(\boldsymbol{x}) = \ell$ and $r(\boldsymbol{x}) = L$. It is known [2] that for integers $n > \ell \geqslant L$, $D(n, \ell, L) = \frac{n}{L}\binom{\ell-1}{L-1}\binom{n-\ell-1}{L-1}$.

For a sequence $\boldsymbol{x} \in \Sigma^n$ and for a natural number $t \in \mathbb{N}$, denote by $\mathcal{S}_t(\boldsymbol{x})$ the radius t sphere around $\boldsymbol{x}$, i.e., $\mathcal{S}_t(\boldsymbol{x}) = \{\boldsymbol{y} : d_p(\boldsymbol{x}, \boldsymbol{y}) = t\}$. In particular, from [2] we have

$$|\mathcal{S}_t(\boldsymbol{x})| = \sum_{\ell = \lceil t/2 \rceil}^{t-1} D(n, \ell, t - \ell).$$

Let $\mathcal{B}_t(\boldsymbol{x}) = \{\boldsymbol{y} : d_p(\boldsymbol{x}, \boldsymbol{y}) \leqslant t\}$ be the ball of radius t around $\boldsymbol{x}$. Then,

$$|\mathcal{B}_t(\boldsymbol{x})| = 1 + \sum_{i=1}^{t} |\mathcal{S}_i(\boldsymbol{x})|.$$

Notice from these expressions that the values for $|\mathcal{S}_t(\boldsymbol{x})|$ and $|\mathcal{B}_t(\boldsymbol{x})|$ do not depend on $\boldsymbol{x}$. Consequently, we denote $S_p(n, t) = |\mathcal{S}_t(\boldsymbol{x})|$ and $B_p(n, t) = |\mathcal{B}_t(\boldsymbol{x})|$. Note that for any fixed t, the order of both $S_p(n, t)$ and $B_p(n, t)$ is $\Theta(n^{\lfloor t/2 \rfloor})$. Thus, according to the sphere packing bound [2], the redundancy of a code with minimum symbol-pair distance d_p is at least roughly

$$\left\lfloor \frac{d_p - 1}{4} \right\rfloor \log(n). \quad (5)$$

We may now state the main result of this subsection. Let $A_p(n, d)$ be the maximal size of a code of length n with pair-distance d.

Theorem 3. *Let $n, d \in \mathbb{N}$ where $d \leqslant \frac{n}{4}$. Let $t \in \mathbb{N}$ be such that $d = 2t + 2$, then*

$$A_p(n, d) \leqslant \frac{2^n}{B_p(n, t) + \frac{S_p(n, t+1)}{\lfloor \frac{n}{t+1} \rfloor}}.$$

The proof follows a similar logic as the proof of the Johnson bound [10, Theorem 2.3.8].

Proof: Let $\mathcal{C} \subseteq \Sigma^n$ be a code of size M with $d_p(\mathcal{C}) \geqslant d$ where $d = 2t + 2$. For a sequence $\boldsymbol{x} \in \Sigma^n$ let $d_p(\mathcal{C}, \boldsymbol{x}) = \min_{\boldsymbol{c} \in \mathcal{C}, \boldsymbol{c} \neq \boldsymbol{x}} \{d_p(\boldsymbol{x}, \boldsymbol{c})\}$ and denote $\mathcal{N} = \{\boldsymbol{x} \in \Sigma^n : d_p(\mathcal{C}, \boldsymbol{x}) = t + 1\}$. Clearly,

$$M \cdot B_p(n, t) + |\mathcal{N}| \leqslant 2^n. \quad (6)$$

Consider the set $\mathcal{X} = \{(\boldsymbol{c}, \boldsymbol{x}) \in \mathcal{C} \times \mathcal{N} \ : \ d_p(\boldsymbol{c}, \boldsymbol{x}) = t+1\}$. We first calculate $|\mathcal{X}|$. For any $\boldsymbol{c} \in \mathcal{C}$, denote by $\mathcal{X}_{\boldsymbol{c}} = \{\boldsymbol{x} \in \mathcal{N} \ : \ (\boldsymbol{c}, \boldsymbol{x}) \in \mathcal{X}\}$ and note that $|\mathcal{X}| = \sum_{\boldsymbol{c} \in \mathcal{C}} |\mathcal{X}_{\boldsymbol{c}}|$. For a fixed $\boldsymbol{c} \in \mathcal{C}$, let $\boldsymbol{x} \in \Sigma^n$ be any sequence such that $d_p(\boldsymbol{c}, \boldsymbol{x}) = t+1$. There are exactly $S_p(n, t+1)$ such sequences. Thus, $w_p(\boldsymbol{c} + \boldsymbol{x}) = t+1$ which means that $d_p(\mathcal{C}, \boldsymbol{x}) \leqslant t+1$. We show that for any $\boldsymbol{c}' \in \mathcal{C}, \boldsymbol{c}' \neq \boldsymbol{c}$, we have $d_p(\boldsymbol{c}', \boldsymbol{x}) \geqslant t+1$, which implies $d_p(\mathcal{C}, \boldsymbol{x}) = t+1$. By the triangle inequality we obtain,

$$\begin{aligned} d \leqslant d_p(\boldsymbol{c}, \boldsymbol{c}') &= w_p(\boldsymbol{c} + \boldsymbol{c}') = w_p(\boldsymbol{c} + \boldsymbol{x} + \boldsymbol{c}' + \boldsymbol{x}) \\ &\leqslant w_p(\boldsymbol{c} + \boldsymbol{x}) + w_p(\boldsymbol{x} + \boldsymbol{c}') = t + 1 + w_p(\boldsymbol{c}' + \boldsymbol{x}). \end{aligned}$$

This implies that if $d = 2t+2$ then $w_p(\boldsymbol{c}' + \boldsymbol{x}) = d_p(\boldsymbol{c}, \boldsymbol{x}) \geqslant t+1$. Since $\boldsymbol{c}'$ was arbitrary we obtain $d_p(\mathcal{C}, \boldsymbol{x}) = t+1$. Therefore, for a fixed $\boldsymbol{c} \in \mathcal{C}$, we have that $|\mathcal{X}_{\boldsymbol{c}}| = S_p(n, t+1)$ which implies that

$$|\mathcal{X}| = M \cdot S_p(n, t+1). \tag{7}$$

We now fix $\boldsymbol{x} \in \mathcal{N}$ and consider the set

$$\mathcal{C}'_{\boldsymbol{x}} = \{\boldsymbol{x} + \boldsymbol{c} \ : \ \boldsymbol{c} \in \mathcal{C} \text{ and } d_p(\boldsymbol{x}, \boldsymbol{c}) = t+1\}.$$

Note that $\mathcal{C}'$ is a constant pair-weight code of length n with codewords of pair-weight $t+1$ and minimum pair-distance d. Therefore, for every choice $\boldsymbol{x} \in \mathcal{N}$, $|\mathcal{C}'_{\boldsymbol{x}}| \leqslant A_p(n, 2t+2, t+1)$. This, in turn, implies that

$$|\mathcal{X}| \leqslant |\mathcal{N}| \cdot A_p(n, 2t+2, t+1). \tag{8}$$

Combining the property that $A_p(n, 2w, w) = \left\lfloor \frac{n}{w} \right\rfloor$ with (6), (7) and (8) we obtain

$$M \left(B_p(n,t) + \frac{S_p(n,t)}{\left\lfloor \frac{n}{t+1} \right\rfloor} \right) \leqslant M \cdot B_p(n,t) + |\mathcal{N}| \leqslant 2^n.$$

Since this is true for every code of size M, it also holds for $A_p(n, d)$ which gives the result. ∎

B. Linear Programming Upper Bounds

We now consider the application of linear programming techniques to derive upper bounds on codes under the pair-distance metric. The approach used here is analogous to the approach from [7]. First, we introduce a mapping, which we refer to as $\mathcal{T}_{[m_1,m_2]}$. With additional notations we then establish the linear programming upper bound.

For a linear code $\mathcal{C} \subseteq \Sigma^n$ and for $i, j \in [n+1]$, let

$$A_{i,j} \triangleq \{\boldsymbol{x} \in \mathcal{C} \ : \ w_H(\boldsymbol{x}) = i, \ r(\boldsymbol{x}) = j\},$$

and denote by $a_{i,j} = |A_{i,j}|$. Note that for $i < j$, $A_{i,j} = \emptyset$, and so we may consider only the cases where $j \leqslant i$. Note also that $|\mathcal{C}| = \sum_{0 \leqslant j \leqslant i \leqslant n} a_{i,j}$. For a code $\mathcal{C}$, we denote by $\pi(\mathcal{C}) = \{\pi(\boldsymbol{c}) \ : \ \boldsymbol{c} \in \mathcal{C}\}$.

Let $m_1, m_2 \in [n+1]$ be such that $m_1 + m_2 \leqslant n$ and let $N = \binom{n}{m_1} \cdot \binom{n-m_1}{m_2}$. We introduce a map $\mathcal{T}_{[m_1,m_2]} : \left(\Sigma^2\right)^n \to \left(\Sigma^2\right)^N$ as follows. Let $\mathcal{J}_0, \ldots, \mathcal{J}_{N-1}$ be all the distinct ways of choosing m_1 positions out of n and then choosing additional m_2 positions. For every $0 \leqslant i \leqslant N-1$ we think of $\mathcal{J}_i$ as a pair $\mathcal{J}_i = \{\mathcal{J}_{i,1}, \mathcal{J}_{i,2}\}$ where $\mathcal{J}_{i,t} \subseteq [n]$, $|\mathcal{J}_{i,t}| = m_t$ for $t = 1, 2$ and $\mathcal{J}_{i,1} \cap \mathcal{J}_{i,2} = \emptyset$. That is, we think of $\mathcal{J}_{i,1}$ as all the m_1 positions that were chosen in the first round and of $\mathcal{J}_{i,2}$ as all the m_2 positions that were chosen in the second round. For a sequence $\boldsymbol{z} = ((z_{0,0}, z_{0,1}), (z_{1,0}, z_{1,1}), \ldots, (z_{n-1,0}, z_{n-1,1})) \in \left(\Sigma^2\right)^n$ we define $\mathcal{T}_{[m_1,m_2]}(\boldsymbol{z}) = \boldsymbol{y}$ where $\boldsymbol{y} = (y_0, \ldots, y_{N-1}) \in \left(\Sigma^2\right)^N$ is defined as follows. For $i \in [N]$,

$$y_i = \left(\sum_{\ell_1 \in \mathcal{J}_{i,1}} (z_{\ell_1,0}, z_{\ell_1,1}) + \sum_{\ell_2 \in \mathcal{J}_{i,2}} (0, z_{\ell_2,1}) \right) \bmod 2.$$

Note that if $\boldsymbol{x} = (x_0, \ldots, x_{n-1})$ and $\mathcal{T}_{[m_1,m_2]}(\pi(\boldsymbol{x})) = \boldsymbol{y}$ where $\boldsymbol{y} = (y_0, \ldots, y_{N-1})$, then for $i \in [N]$,

$$y_i = \left(\sum_{\ell_1 \in \mathcal{J}_{i,1}} (x_{\ell_1,0}, x_{\ell_1,1}) + \sum_{\ell_2 \in \mathcal{J}_{i,2}} (0, x_{\ell_2,1}) \right) \bmod 2.$$

We introduce two more notations for the simplicity of writing. Let $m_1, m_2, n \in \mathbb{N}$ be such that $n > 0$, $m_1, m_2 \geqslant 0$ and let $j, i \in \mathbb{N}$. We define

$$\begin{aligned} K_{m_1,m_2}(n,i,j) \triangleq 4 \sum_{s \equiv t+s'+t' \equiv u \,(\bmod\, 2)} & \binom{i-j}{s} \binom{j}{t} \times \\ & \binom{j}{u} \binom{n-i-j}{m_1-s-t-u} \binom{i-j-s}{s'} \binom{j-t}{t'} \times \\ & \binom{n-i-u}{m_2-s'-t'} - \binom{n}{m_1} \binom{n-m_1}{m_2}. \end{aligned}$$

For integers n, m where $0 \leqslant m \leqslant n$, let

$$K_m(n,i) \triangleq \sum_{k=0}^{m} (-1)^k \binom{i}{k} \binom{n-i}{m-k}.$$

The next corollary establishes the statement of this bound.

Corollary 4. *Suppose $\mathcal{C} \subseteq \Sigma^n$ with $d_p(\mathcal{C}) \geqslant d$. Then, $|\mathcal{C}|$ is upper bounded by the following expression*

$$\begin{aligned} \textit{Maximize} \quad & \sum_{1 \leqslant j \leqslant i} a_{i,j} \\ \textit{Subject to:} \quad & 1) \ a_{0,0} = 1 \\ & 2) \ a_{i,j} = 0 \ \textit{ if } \ i + j < d \\ & 3) \ \sum_{i=0}^{n} K_m(n,i) \left(\sum_{j=1}^{i} a_{i,j} \right) \geqslant 0 \\ & \quad 0 \leqslant m \leqslant n \\ & 4) \ K_{m_1,m_2}(n,0,0) a_{0,0} + \sum_{i=1}^{n-1} \sum_{j=1}^{i} K_{m_1,m_2}(n,i,j) a_{i,j} \\ & \quad + K_{m_1,m_2}(n,n,0) a_{n,0} \geqslant 0 \\ & \quad 0 \leqslant m_1 + m_2 \leqslant n. \end{aligned}$$

Our results are highlighted in Table I. Each entry consists of a pair of numbers delimited by a '/' where the first number in the pair represents the result of using our linear programming bound or the bound from Theorem 3 and the second number represents the sphere-packing upper bound from [2].

TABLE I
RESULTS OF COROLLARY 4 AND THEOREM 3 VS. THE SPHERE PACKING BOUND.

$n\backslash d_p$	1	2	3	4	5	6	7	8
2	4 / 4	4 / 4	2 / 4	-	-	-	-	-
3	8 / 8	8 / 8	4 / 8	2 / 8	-	-	-	-
4	16 / 16	16 / 16	8 /16	4 / 16	2 / 4	-	-	-
5	32 / 32	32 / 32	16 / 32	9 / 32	4 / 5	2 / 5	-	-
6	64 / 64	64 / 64	32 / 64	21 / 64	8 / 9	4 / 9	2 / 4	-
7	128 / 128	128 / 128	64 / 128	38 / 128	16 / 16	8 / 16	4 / 8	3 / 8

IV. CONSTRUCTIONS FOR SMALL PAIR-DISTANCE

In this section, we present new code constructions for the pair-symbol channel. Table II compares the sizes of the best known code constructions for the pair-symbol channel (in terms of their code lengths) to the sphere-packing upper bound from [2] and to Theorem 3 where we have highlighted the contributions of this section by $(*)$. The fourth column of the table labeled "Hamming Distance of the Code" is a lower bound on the Hamming distance of the code whose size is listed in the third column. As can be seen from Table II, our constructions provide codes that improve upon the state-of-the-art results for the cases where $d_p = 4, 6, 7,$ and 10. Note that the case of $d_p = 2, 3$ are trivial and are fully solved. Every code has minimum pair distance at least 2, so $A_p(n, 2) = 2^n$ and the simple parity code has minimum pair distance 3, and this construction is optimal so $A_p(n, 3) = 2^{n-1}$. In general, if we were to apply the codes from [18] in order to construct codes with minimum pair distance d_p then the Hamming distance d_H of the codes will have to satisfy $\left\lceil \frac{3d_H}{2} \right\rceil \geqslant d_p$, or $d_H \geqslant \left\lfloor \frac{2d_p+1}{3} \right\rfloor$. Hence, the redundancy of these codes will be roughly

$$\left\lfloor \frac{\left\lfloor \frac{2d_p+1}{3} \right\rfloor - 1}{2} \right\rfloor \log(n) = \left\lfloor \frac{d_p - 1}{3} \right\rfloor \log(n),$$

which is already close to the lower bound in (5). Our goal in this section is to improve this construction for $d_p = 4, 6, 7, 10$.

A. Codes with Minimum Pair-Distance Four

We begin with the case of minimum pair distance four. Let $\boldsymbol{w} = (w_1, \ldots, w_n)$ be defined so that $w_i = 1$ if i is odd an $w_i = 2$ otherwise.

Theorem 5. *Let $n \geqslant 4$ be an even integer and let $\mathcal{C}_4(n) = \{\boldsymbol{x} \in \Sigma^n : \sum_{i=1}^n x_i \cdot w_i \equiv 0 \bmod 4\}$. Then, $d_p(\mathcal{C}_4(n)) = 4$.*

The following corollary states the size of a code constructed from Theorem 5.

Corollary 6. *For an even $n \geqslant 4$, $|\mathcal{C}_4(n)| = \frac{2^n}{4}$.*

Suppose $\mathcal{C}_4(n)$ is a code from Theorem 5. If codes from [18] were used to construct a code with pair distance 4, then the code would require Hamming distance at least 3 and so by the sphere packing bound, the cardinality of the code would be at most $\frac{2^n}{n+1}$. Since $\frac{2^n}{n+1} < 2^{n-2}$, the size of $\mathcal{C}_4(n)$ is larger than the codes from [18]. These codes are also close to optimality since by Theorem 3 we have that the upper bound for even n is $\frac{2^n}{3}$.

B. Codes with Minimum Pair-Distance Six

We now show how to construct codes with minimum pair-distance 6. Let $\{0, \alpha, \alpha^2, \alpha^3 = 1\} = \mathbb{F}_4$ and define the map $\Pi : \mathbb{F}_4 \to \mathbb{F}_2^2$ so that $\Pi(0) = (0,0), \Pi(\alpha) = (1,0), \Pi(\alpha^2) = (1,1), \Pi(\alpha^3) = (0,1)$. Clearly, the map Π is invertible. For a sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_{\frac{n}{2}}) \in \mathbb{F}_4^{\frac{n}{2}}$, let $\Pi(\boldsymbol{x}) = (\Pi(x_1), \ldots, \Pi(x_{\frac{n}{2}})) \in \left(\mathbb{F}_2^2\right)^{\frac{n}{2}}$. Note that $\left(\mathbb{F}_2^2\right)^{\frac{n}{2}}$ is isomorphic to $\mathbb{F}_2^n$, and hence, for a sequence $\boldsymbol{x} \in \mathbb{F}_4^{\frac{n}{2}}$, we may consider $\Pi(x) \in \mathbb{F}_2^n$. Similarly for a set $\mathcal{Z} \subseteq \mathbb{F}_4^{\frac{n}{2}}$ let $\Pi(\mathcal{Z})$ be the result of applying the map Π to every element in $\mathcal{Z}$.

We now describe the code using a parity-check matrix. We construct the matrix

$$H = (H^{(0)}, H^{(1)}, \ldots, H^{(m-4)}) \in \mathbb{F}_4^{m \times N}$$

where $N = \frac{4^m - 4^3}{3} - 2m + 6$. For $j \in [m-3]$,

$$H^{(j)} = (\boldsymbol{h}_1^{(j)}, \ldots, \boldsymbol{h}_{4^{m-j}-2}^{(j)}) \in \mathbb{F}_4^{m \times (4^{m-j}-2)}$$

where for every $i \in \{1, \ldots, m-j-2\}$, $\boldsymbol{h}_i^{(j)}$ is a sequence of length m. Note that for all $j_1 \neq j_2 \in [m-3]$, $H^{(j_1)}$ and $H^{(j_2)}$ have different sizes.

The matrix H has the following properties:

1) For any non-zero $\boldsymbol{z} \in \mathbb{F}_4^N$, if $H \cdot \boldsymbol{z}^T = \boldsymbol{0}$, then $w_H(\boldsymbol{z}) \geqslant 3$.
2) Suppose $\boldsymbol{z} \in \mathbb{F}_4^{4^{m-j}-2}$ and $w_H(\boldsymbol{z}) = 3$. Then, if $H^{(j)} \cdot \boldsymbol{z}^T = 0$, $r(\boldsymbol{z}) \geqslant 2$.
3) Let α be a primitive element for $\mathbb{F}_4$, for any $j \in [m-3]$,
$$\alpha^3 \cdot \boldsymbol{h}_{4^{m-j}-2}^{(j)} + \alpha \cdot \boldsymbol{h}_1^{(j+1)} \neq \alpha^3 \cdot \boldsymbol{h}_k^{(j)},$$
where $k \in \{1, \ldots, m-j-2\}$ and we assume $\boldsymbol{h}_1^{(m-3)} = \boldsymbol{h}_1^{(0)}$.

The code $\mathcal{C} \subseteq \mathbb{F}_4^N$ is defined as

$$\mathcal{C} \triangleq \{\boldsymbol{c} \in \mathbb{F}_4^N : H \cdot \boldsymbol{c}^T = \boldsymbol{0}\}. \tag{9}$$

The following theorem states the properties of the matrix H to produce codes with minimum pair-distance 6.

Theorem 7. *Let $\mathcal{C}$ be a linear code of length N as defined in (9). Then, $d_p(\Pi(\mathcal{C})) \geqslant 6$.*

We consider the cardinality of the code $\Pi(\mathcal{C})$ and compare it with the previously best known codes from [18]. Since $\mathcal{C}$ is a sub-code of a quaternary Hamming code of length $N = \frac{4^m - 4^3}{3} - 2m + 6$, we know that $\mathcal{C} \leqslant \frac{4^N}{4^m}$ where $m \leqslant \log_4(3N + 4^3)$. Since the mapping Π generates codes

TABLE II
TABLE OF LARGEST KNOWN CODES AND UPPER BOUNDS ((∗) DENOTES OUR CONTRIBUTION)

d_p	Upper Bound	Lower Bound on Code Size	Hamming Distance of the Code
4	$\frac{2^n}{1+\frac{n}{\lfloor \frac{n}{2} \rfloor}}$	$2^{n-2}(*)$	2
5	$\frac{2^n}{n+1}$	$\frac{2^n}{n+1}$ [2]	3
6	$\frac{2^n}{1+n+\frac{n}{\lfloor \frac{n}{3} \rfloor}}$	$\frac{2^n}{1+\frac{3n}{2}+4^3}(*)$	4
7	$\frac{2^n}{1+2n}$	$\frac{2^n}{4(1+n)}(*)$	4
8	$\frac{2^n}{2n+1+\frac{n(n-1)}{2\lfloor \frac{n}{4} \rfloor}}$	$\frac{2^n}{(n+1)^2}$ [18]	5
9	$\frac{2^n}{\frac{1}{2}(2+3n+n^2)}$	$\frac{2^n}{2(n+1)^2}$ [18]	6
10	$\frac{2^n}{\frac{1}{2}(2+3n+n^2)+\frac{n(n-3)}{\lfloor \frac{n}{5} \rfloor}}$	$\frac{2^n}{2(n-1)^2}(*)$	6

of length $2N$, setting $n = 2N$ gives: $|\mathcal{C}| \geqslant \frac{2^n}{\frac{3n}{2}+4^3}$. If the construction from [18] was used to produce a code with $d_p \geqslant 6$, we would need a binary cyclic code with Hamming distance at least 4 and so the cardinality of the code would be at most $\frac{2^n}{2(1+n)}$, which is less than $|\mathcal{C}|$. Note that the upper bound in this case is roughly $\frac{2^n}{n+4}$.

C. Codes with Minimum Pair-Distance Seven

We now turn to the construction of codes capable of correcting 3 pair-symbol errors. Although the proof follows almost directly from [1], we highlight the result by stating it as the following theorem.

Theorem 8. *Let $n = 2^m - 1$ where m is an even integer. Then, the binary cyclic code $\mathcal{C}$ of length n with generator polynomial $g(x) = (1 + x + x^2)p(x)$, where $p(x)$ is a primitive polynomial, satisfies $d_p(\mathcal{C}) \geqslant 7$.*

For $n = 2^m - 1$, where $\mathcal{C}$ is a code constructed according to Theorem 8, we have $d_p(\mathcal{C}) \geqslant 7$, and $|\mathcal{C}| \geqslant \frac{2^n}{4(1+n)}$. If we were to construct a code capable of correcting 3 pair-symbol errors using the techniques from [18], the code would have to have minimum Hamming distance 5, and thus its cardinality will be at most

$$\frac{2^n}{1+n+\binom{n}{2}}.$$

Hence, our codes offer an improvement in codebook size in this case. Furthermore, the redundancy of our codes is at most a single bit from the lower bound on the redundancy.

D. Codes with Minimum Pair-Distance Ten

Now, we turn to construct codes that have minimum pair-distance 10. This is established in the next theorem.

Theorem 9. *For a positive integer $m > 2$, let $g(x) \in \mathbb{F}_2[x]$ be a generator polynomial for a cyclic code $\mathcal{C} \subseteq \mathbb{F}_2^{2^m-1}$ with roots $\{\alpha^{-1}, \alpha^0, \alpha^1\} \subset \mathbb{F}_{2^m}$ where α is a primitive element of $\mathbb{F}_{2^m}$. Then, $d_p(\mathcal{C}) \geqslant 10$.*

The size of a code from Theorem 9 is at least $\frac{2^n}{2(n+1)^2}$, whereas if the construction from [18] were used to construct codes with minimum pair-distance 10, their minimum distance should be at least 7 and so their size will be at most $\frac{2^n}{1+n+\binom{n}{2}+\binom{n}{3}}$.

REFERENCES

[1] K.A.S. Abdel-Ghaffar, R. McEliece, A. Odlyzko, and H.C.A. Van Tilborg, "On the existence of optimum cyclic burst-correcting codes," *IEEE Trans. Inform. Theory*, vol. 32, no. 6, pp. 768–775, Nov. 1986.

[2] Y. Cassuto and M.Blaum, "Codes for symbol-pair read channels," *IEEE Trans. Inform. Theory*, vol. 57, no. 12, pp. 8011–8020, Dec. 2011.

[3] Y. Cassuto and S. Litsyn, "Symbol-pair codes: Algebraic constructions and asymptotic bounds," *Proc. IEEE Int. Symp. Inf. Theory*, St. Petersburg, Russia, pp. 2348–2352, Jul./Aug. 2011.

[4] Y. M. Chee, L. Ji, H.M. Kiah, C. Wang, and J. Yin, "Maximum distance separable codes for symbol-pair read channels," *IEEE Trans. Inform. Theory*, vol. 59, no. 11, pp. 7259–7267, Aug. 2013.

[5] B. Ding, G. Ge, J. Zhang, T. Zhang, and Y. Zhang, "New constructions of MDS symbol-pair codes," *Des. Codes Cryp.*, pp1–17, 2017.

[6] H. Q. Dinh, B. T. Nguyen, A. K. Singh, and S. Sriboonchitta, "On the symbol-pair distance of repeated-root constacyclic codes of prime power lengths," *IEEE Trans. Inform. Theory*, Jul. 2017.

[7] J. I. Hall, *Notes on Coding Theory*, available at http://www.mth.msu.edu/∼jhall/classes/codenotes/coding-notes.html, 2013.

[8] M. Hirotomo, M. Takita, and M. Morii, "Syndrome decoding of symbol-pair codes," *Proc. IEEE Inf. Theory Workshop*, Hobart, TAS, Australia, pp. 162–166, Nov. 2014.

[9] S. Horii, T. Matsushima, and S. Hirasawam, "Linear programming decoding of binary linear codes for symbol-pair read channel," *IEICE Trans. on Fund. of Elec., Comm. and Comp. Sci.*, vol. E99-A, no. 12, pp.2170–2178, 2016.

[10] W.C. Huffman and V. Pless, *Fundamentals of Error Correcting Codes*, Cambridge University Press, 2003.

[11] X. Kai, S. Zhu, and P. Li, "A construction of new MDS symbol-pair codes," *IEEE Trans. Inform. Theory*, vol. 61, no. 11, pp. 5828–5834, Nov. 2015.

[12] S. Li and G. Ge, "Constructions of maximum distance separable symbol-pair codes using cyclic and constacyclic codes," *Des. Codes Cryp.*, vol. 84, pp. 359–371, Sep. 2017.

[13] M. Morii, M. Hirotomo, and M. Takita, "Error-trapping decoding for cyclic codes over symbol-pair read channels," *Int. Symp. on Information Theory and Its App.*, pp. 681–685, Oct. 2016.

[14] Z. Sun, S. Zhu, and L. Wang, "The symbol-pair distance distribution of a class of repeated-root cyclic codes over F_{p^m}," *Cryp. and Comm.*, pp. 116–124, Nov. 2017.

[15] M. Takita, M. Hirotomo, and M. Morii, "A decoding algorithm for cyclic codes over symbol-pair read channels," *IEICE Trans. on Fund. of Elec., Comm. and Comp. Sci.*, vol. E98-A, no. 12, pp.2415–2422, 2015.

[16] M. Takita, M. Hirotomo, and M. Morii, "Algebraic decoding of BCH codes over symbol-pair read channels: Cases of two-pair and three-pair error correction," *IEICE Trans. on Fund. of Elec., Comm. and Comp. Sci.*, vol. E99-A, no. 12, pp.2179–2191, 2016.

[17] M. Takita, M. Hirotomo, and M. Morii, "Error-trapping decoding for cyclic codes over symbol-pair read channels," *IEICE Trans. on Fund. of Elec., Comm. and Comp. Sci.*, vol. E100-A, no. 12, pp.2578–2584, 2017.

[18] E. Yaakobi, J. Bruck, and P. Siegel, "Constructions and decoding of cyclic codes over b-symbol read channels," *IEEE Trans. Inform. Theory*, vol. 62, no. 4, pp. 1541–1551, Apr. 2016.

[19] H. Zhang, "Improvement on minimum distance of symbol-pair codes," *Des. Codes Cryp.*, pp. 116–124, Nov. 2017.