Bounds and Constructions of Codes Over Symbol-Pair Read Channels

Ohad Elishco, *Member, IEEE*, Ryan Gabrys, *Member, IEEE*, and Eitan Yaakobi, *Senior Member, IEEE*

***Abstract*— Cassuto and Blaum recently studied the symbol-pair channel, a model where every two consecutive symbols are read together. This special channel structure 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 not individual symbol errors, but rather *symbol-pair errors*, where at least one of the symbols is erroneous. In this work, we study bounds and constructions of codes over the symbol-pair channel. 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 for pair-distance six, seven, and ten.**

***Index Terms*— Coding theory, codes for storage media, symbol-pair codes.**

I. Introduction

A BASIC limitation of high density data storage systems is that the outputs of the reading process are pairs of consecutive symbols rather than individual symbols. The symbol-pair read channel was recently proposed as a model reflecting this limitation and it was first studied in [2] and [3]. In those papers the authors studied 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 [20], maximum distance separable (MDS) codes [13], decoding algorithms [21] and more.

In [21], the authors presented efficient decoding algorithms that improved the initial lower bounds from [2] and [3] on the minimum pair-distance of linear cyclic codes. Several more works presented different decoding algorithms for arbitrary linear codes; see e.g. [9], [10], [15], [18]–[20]. 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], [12], [13], [17]. Another generalization of the pair-symbol model was studied in [21] 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], [13].

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 pair-read vector 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 goal is to construct 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 [21], it was shown that if a linear cyclic code has minimum Hamming distance d_H then its minimum pair-distance is at least $d_p \geqslant \left\lceil \frac{3d_H}{2} \right\rceil$. This work presents also decoding algorithms for such codes. On the other hand, bounds on codes with minimum pair-distance were studied in [3], 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 for the symbol-pair read channel. Specifically, we show how to extend the Johnson bound and the linear programming bound for this setup, and demonstrate how the new bounds improve upon the best previously known bound from [3]. We also study code constructions for relatively small minimum pair-distance, namely six, seven, and ten. For these cases we show how to improve the result from [21] using cyclic linear codes in order to obtain 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

Manuscript received August 24, 2018; revised August 13, 2019; accepted September 7, 2019. Date of publication September 18, 2019; date of current version February 14, 2020. This work was supported in part by the Israel Science Foundation under grant 1624/14 and by the NISE Program at the Naval Information Warefare Center. This paper was presented in part at the 2018 IEEE International Symposium on Information Theory (ISIT) [7].

O. Elishco was with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. He is now with the Electrical Engineering Department, University of Maryland, College Park, MD 20742 USA (e-mail: ohadeli@mit.edu).

R. Gabrys is with the Spawar Systems Center, San Diego, CA 92115 USA (e-mail: ryan.gabrys@navy.mil).

E. Yaakobi is with the Department of Computer Science, Technion — Israel Institute of Technology, Haifa 32000, Israel (e-mail: yaakobi@cs.technion.ac.il).

Communicated by P. Sadeghi, Associate Editor for Coding Techniques.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2942283

pair-distance is six, seven, and ten. Lastly, in Section V we conclude the paper and discuss the results and future research.

II. Preliminaries

Let $n \in \mathbb{N}$ (the natural numbers, including 0), and denote by $[n]$ the set $\{0, \ldots, n-1\}$. For a prime power q, we denote by $\mathbb{F}_q$ the field of size q. For a sequence $\boldsymbol{x} \in \mathbb{F}_2^n$ denote by $w_H(\boldsymbol{x})$ the Hamming weight of $\boldsymbol{x}$. For two sequences $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$ let $d_H(\boldsymbol{x}, \boldsymbol{y})$ denote the Hamming distance between $\boldsymbol{x}, \boldsymbol{y}$. For a set $\mathcal{C} \subseteq \mathbb{F}_2^n$, 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 $\mathbf{0}, \mathbf{1} \in \mathbb{F}_2^n$ the all zeros and the all ones sequences, respectively.

Definition 1. *Let $\pi : \mathbb{F}_2^n \to (\mathbb{F}_2 \times \mathbb{F}_2)^n$ denote the (cyclic) pair-symbol read representation which is defined as follows. For $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1}) \in \mathbb{F}_2^n$,*

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

For a code $\mathcal{C} \subseteq \mathbb{F}_2^n$, we denote by $\pi(\mathcal{C}) \subseteq \left(\mathbb{F}_2^2\right)^n$ the pair-code generated by $\mathcal{C}$, i.e.,

$$\pi(\mathcal{C}) \triangleq \{\pi(\boldsymbol{c}) \ : \ \boldsymbol{c} \in \mathcal{C}\} .$$

We now define the associated pair-weight and pair-distance. For a sequence $\boldsymbol{x} \in \mathbb{F}_2^n$, 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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$, define the pair-distance as

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

with coordinates taken modulo n. For a set $\mathcal{C} \subseteq \mathbb{F}_2^n$ we define $d_p(\mathcal{C})$ as the minimum pair-distance between two codewords,

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

In the following, we make use of the well known property (see [21])

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = w_p(\boldsymbol{x} - \boldsymbol{y}). \tag{1}$$

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

$$d_p(\mathcal{C}) = \min_{\boldsymbol{y}\in\mathcal{C},\boldsymbol{y}\neq\mathbf{0}} d_p(\mathbf{0}, \boldsymbol{y}) = \min_{\boldsymbol{y}\in\mathcal{C},\boldsymbol{y}\neq\mathbf{0}} w_p(\boldsymbol{y}). \tag{2}$$

Example 1. Let $\boldsymbol{x} = (0110)$, $\boldsymbol{y} = (0101) \in \mathbb{F}_2^4$. We have that

- $w_H(\boldsymbol{x}) = w_H(\boldsymbol{y}) = 2$.
- $d_H(\boldsymbol{x}, \boldsymbol{y}) = 2$.
- $\pi(\boldsymbol{x}) = ((0, 1), (1, 1), (1, 0), (0, 0))$.
- $\pi(\boldsymbol{y}) = ((0, 1), (1, 0), (0, 1), (1, 0))$.
- $w_p(\boldsymbol{x}) = 3$, $w_p(\boldsymbol{y}) = 4$.
- $d_p(\boldsymbol{x}, \boldsymbol{y}) = 3$.

Define $r(\boldsymbol{x}) \triangleq |\{i : \pi(\boldsymbol{x})_i = (0, 1)\}|$. It is evident that $r(\boldsymbol{x})$ is 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 the sequence $\mathbf{1}$ has $r(\mathbf{1}) = 0$. We may consider $r(\boldsymbol{x})$ as the number of runs of consecutive symbols in $\boldsymbol{x}$. Throughout the rest of the paper, we will consider only binary vectors unless mentioned otherwise or it is clear from the context. It is known [3] that for the binary case

$$w_p(\boldsymbol{x}) = w_H(\boldsymbol{x}) + r(\boldsymbol{x}). \tag{3}$$

Moreover, if $\mathcal{C}$ is a linear code, from (2), it is straightforward to verify that

$$d_p(\mathcal{C}) = \min_{\boldsymbol{y}\in\mathcal{C},\boldsymbol{y}\neq 0} \{w_H(\boldsymbol{y}) + r(\boldsymbol{y})\}. \tag{4}$$

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$. We make use of the following lemma from [21] which we include for completeness.

Lemma 2. *[21, Lemma 1] Suppose $\mathcal{C} \subseteq \mathbb{F}_2^n$ is a cyclic linear code with dimension greater than 1 and with $d_H(\mathcal{C}) \geqslant d$. 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

$$d_p(\mathcal{C}) \geqslant \left\lceil \frac{3d_H(\mathcal{C})}{2} \right\rceil. \tag{5}$$

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 [11, Theorem 2.3.8]. In Lemma 3, we first derive a bound on the maximal size of a code with minimum pair-distance at least $2w$ where each codeword has pair-weight w. Afterwards, this result is used in Theorem 5 to derive an upper bound for codes with even pair-distance.

Let $n, d, w \in \mathbb{N}$ and denote by $A_p(n, d)$ the maximal size of a code of length n with pair-distance d, and by $A_p(n, d, w)$ the maximal size of a code of length n with pair-distance at least d where each codeword has pair-weight w. For the specific case in which $d = 2w$ we have the following equality.

Lemma 3. *For positive integers n, w,*

$$A_p(n, 2\,w, w) = \left\lfloor \frac{n}{w} \right\rfloor .$$

Proof: Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$ have pair-weight $w_p(\boldsymbol{x}) = w_p(\boldsymbol{y}) = w$ and $d_p(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2w$. Since d_p is a metric, by the triangle inequality, $d_p(\boldsymbol{x}, \boldsymbol{y}) \leqslant d_p(\boldsymbol{x}, \mathbf{0}) + d_p(\mathbf{0}, \boldsymbol{y}) = 2\ w$, which implies that if $d_p(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2w$, then $d_p(\boldsymbol{x}, \boldsymbol{y}) = 2w$. At first, we show that $A_p(n, 2w, w) \geqslant \lfloor \frac{n}{w} \rfloor$. We do so by

constructing a code as follows. Let $\boldsymbol{c}_i$, $0 \leqslant i < \left\lfloor \frac{n}{w} \right\rfloor$, be the sequence generated by appending iw zeroes, $(w-1)$ ones and $n - iw - w + 1$ zeros, i.e., $\boldsymbol{c}_i = 0^{iw}1^{w-1}0^{n-iw-w+1}$ and consider the code $\mathcal{C} = \left\{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{\left\lfloor \frac{n}{w} \right\rfloor - 1}\right\}$. Note that $|\mathcal{C}| = \left\lfloor \frac{n}{w} \right\rfloor$, $\forall \boldsymbol{x} \in \mathcal{C}$, $w_p(\boldsymbol{x}) = w$ and $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $d_p(\boldsymbol{x}, \boldsymbol{y}) = 2w$. Thus, $A_p(n, 2w, w) \geqslant \left\lfloor \frac{n}{w} \right\rfloor$.

In order to show that $A_p(n, 2w, w) \leqslant \left\lfloor \frac{n}{w} \right\rfloor$ we note (using (1)) that $2w \leqslant d_p(\boldsymbol{x}, \boldsymbol{y}) = w_p(\boldsymbol{x} + \boldsymbol{y})$. Since each word has pair-weight w we obtain that $2w = w_p(\boldsymbol{x}) + w_p(\boldsymbol{y}) = r(\boldsymbol{x}) + w_H(\boldsymbol{x}) + r(\boldsymbol{y}) + w_H(\boldsymbol{y})$. Combining the inequalities with the fact that $d_p(\boldsymbol{x}, \boldsymbol{y}) \leqslant w_p(\boldsymbol{x}) + w_p(\boldsymbol{y})$ we obtain

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

This equality means that for every $j \in [n]$, $\boldsymbol{x}_j \cdot \boldsymbol{y}_j = 0$ and $\boldsymbol{x}_j \cdot \boldsymbol{y}_{j+1} = \boldsymbol{y}_j \cdot \boldsymbol{x}_{j+1} = 0$. In other words, it means that $\boldsymbol{x}$ and $\boldsymbol{y}$ cannot have 1 in the same position. Moreover, if $\boldsymbol{x}$ ($\boldsymbol{y}$) has a run of ones that ends in position j, then $\boldsymbol{y}$ ($\boldsymbol{x}$) cannot have a run of ones beginning in position $j + 1$.

Suppose $\mathcal{C}$ is a code of maximal size where all sequences in $\mathcal{C}$ have pair-weight w and $d_p(\mathcal{C}) = 2w$. Let $\mathcal{M} \in \mathbb{F}_2^{M \times n}$ be a matrix which contains as rows the codewords from $\mathcal{C}$. From the previous discussion, notice that each column of $\mathcal{M}$ has at most a single 1 in it. Additionally, the column immediately after the end of any run of ones must not contain any 1. We say that column j of $\mathcal{M}$ is occupied by $\boldsymbol{x}$ if $\boldsymbol{x}$ has a 1 in position j or a run of ones ending in position $j - 1$. We claim that the number of columns of $\mathcal{M}$ occupied by any codeword in $\mathcal{C}$ is $w_H(\boldsymbol{x}) + r(\boldsymbol{x}) = w$. Since each codeword occupies w columns of $\mathcal{M}$, and each column can only be occupied by a single codeword, we can have at most $\left\lfloor \frac{n}{w} \right\rfloor$ codewords in $\mathcal{C}$. ∎

Before continuing, we introduce some additional notation and useful results from [3]. 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 of weight ℓ that have L runs. Recall that a sequence $\boldsymbol{x}$ is said to have L runs if the symbol $(0, 1)$ occurs L times in $\pi(\boldsymbol{x})$ and so a run is sequence of ones or zeros which appear consecutively. The next lemma appears in [3].

Lemma 4. *[3] For integers $n > \ell \geqslant L$,*

$$D(n, \ell, L) = \frac{n}{L} \begin{pmatrix} \ell - 1 \\ L - 1 \end{pmatrix} \begin{pmatrix} n - \ell - 1 \\ L - 1 \end{pmatrix}.$$

For a sequence $\boldsymbol{x} \in \mathbb{F}_q^n$ 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 [3] we have

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

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

$$S_p(n, t) = |\mathcal{S}_t(\boldsymbol{x})|, \tag{6}$$

and

$$B_p(n, t) = |\mathcal{B}_t(\boldsymbol{x})|. \tag{7}$$

Note that for any fixed t, the order of both $S_p(n, t)$ and $B_p(n, t)$ is $\Theta(n^{\lfloor t/2 \rfloor})$. A code with pair distance $2t+1$ has size at most $\frac{2^n}{B_p(n,t)}$ which implies at least $\log B_p(n, t) = \frac{t}{2} \log n$ bits of redundancy. Thus, according to the sphere packing bound [3], the redundancy of a code with minimum symbol-pair distance d_p is at least

$$\left\lfloor \frac{d_p - 1}{4} \right\rfloor \log(n). \tag{8}$$

We may now state the main result of this subsection. The proof follows a similar logic as in the proof of the Johnson bound [11, Theorem 2.3.8].

Theorem 5. *Let $n, d \in \mathbb{N}$ where d is even. Let $t \in \mathbb{N}$ be such that $d = 2t + 2$, then*

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

Proof: Let $\mathcal{C} \subseteq \mathbb{F}_2^n$ be a code of size M with $d_p(\mathcal{C}) \geqslant d$ where $d = 2t + 2$. For a sequence $\boldsymbol{x} \in \mathbb{F}_2^n$, 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 $\mathcal{N} = \{\boldsymbol{x} \in \mathbb{F}_2^n : d_p(\mathcal{C}, \boldsymbol{x}) = t + 1\}$. Clearly,

$$M \cdot B_p(n, t) + |\mathcal{N}| \leqslant 2^n. \tag{9}$$

Consider the set $\mathcal{X} = \left\{(\boldsymbol{c}, \boldsymbol{x}) \in \mathcal{C} \times \mathcal{N} : d_p(\boldsymbol{c}, \boldsymbol{x}) = t + 1\right\}$. 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 $\boldsymbol{x} \in \mathbb{F}_2^n$ 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

$$|\mathcal{X}| = M \cdot S_p(n, t + 1). \tag{10}$$

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

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

Note that $\mathcal{C}'_{\boldsymbol{x}}$ 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

$$|\mathcal{X}| \leqslant |\mathcal{N}| \cdot A_p(n, 2t+2, t+1). \tag{11}$$

Combining Lemma 3 with (9), (10) and (11) we obtain

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

Since this is true for every code of size M, it is also true for the value $A_p(n,d)$, which establishes the theorem's statement. ∎

B. Linear Programming Upper Bound

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 [8]. First, we introduce a mapping, which we refer to as $\mathcal{T}_{[m_1,m_2]}$. In Lemma 6, we show that if the input to $\mathcal{T}_{[m_1,m_2]}$ is a sequence with Hamming weight i and j runs, then the Hamming weight of the output sequence is a function of i and j only. Lemma 7 then provides a necessary condition on the weight enumerator for a code over $(\mathbb{F}_2^2)^n$ and this condition along with the result from Lemma 6 is used in Lemma 8 to derive a set of necessary conditions on the number of codewords in a code with a prescribed pair-distance. Finally, Theorem 9 gives our linear programming upper bound, which contains the statement of Lemma 8 as one of its necessary conditions.

Before continuing, we first introduce some useful notation. For a linear code $\mathcal{C} \subseteq \mathbb{F}_2^n$ and for $i, j \in [n]$, 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 $|\mathcal{C}| = \sum_{0 \leqslant j \leqslant i < n} a_{i,j}$. The following claim states that given $w_H(\boldsymbol{x})$ and $r(\boldsymbol{x})$ we can determine the statistics of the symbols in $\pi(\boldsymbol{x})$.

Claim 1 *Suppose $\boldsymbol{x} \in \mathbb{F}_2^n$ is such that $w_H(\boldsymbol{x}) = i$ and $r(\boldsymbol{x}) = j$ and denote $\pi(\boldsymbol{x}) = (z_0, z_1, \ldots, z_{n-1}) \in \left(\mathbb{F}_2^2\right)^n$. Then*

1) $|\{\ell \in [n] \ : \ z_\ell = (1,1)\}| = i - j$,
2) $|\{\ell \in [n] \ : \ z_\ell = (1,0)\}| = j$, *and*
3) $|\{\ell \in [n] \ : \ z_\ell = (0,1)\}| = j$.

Proof: Note that under the setup described in the claim, $w_p(\boldsymbol{x}) = i + j$. The fact that $|\{\ell \in [n] \ : \ z_\ell = (1,0)\}| = |\{\ell \in [n] \ : \ z_\ell = (0,1)\}| = j$ follows immediately from the definition of $r(\boldsymbol{x})$ presented in Section II. Since $w_p(\boldsymbol{x}) = i+j$, it follows then that $|\{\ell \in [n] : z_\ell = (1,1)\}| = w_p(\boldsymbol{x}) - 2j = i - j$ as desired. ∎

Let $m_1, m_2 \in [n]$ 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 mapping $\mathcal{T}_{[m_1,m_2]} : \left(\mathbb{F}_2^2\right)^n \to \left(\mathbb{F}_2^2\right)^N$ as follows. Let $\mathcal{J}_0, \ldots, \mathcal{J}_{N-1}$ be 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 $\mathcal{J}_i = \left\{\mathcal{J}_{i,1}, \mathcal{J}_{i,2}\right\}$ where $\mathcal{J}_{i,t} \subseteq [n]$, $|\mathcal{J}_{i,t}| = m_t$ for $t = 1, 2$ and $\mathcal{J}_{i,1} \cap \mathcal{J}_{i,2} = \emptyset$. Meaning that 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 $\boldsymbol{z} = \left((z_{0,0}, z_{0,1}), (z_{1,0}, z_{1,1}), \ldots, (z_{n-1,0}, z_{n-1,1})\right) \in \left(\mathbb{F}_2^2\right)^n$ we define $\mathcal{T}_{[m_1,m_2]}(\boldsymbol{z}) = \boldsymbol{y}$ where $\boldsymbol{y} = (y_0, \ldots, y_{N-1}) \in \left(\mathbb{F}_2^2\right)^N$ is defined as follows. For $i \in [N]$,

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

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 $i \in [N]$,

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

Lemma 6. *Suppose $\boldsymbol{x} \in \mathbb{F}_2^n$, $w_H(\boldsymbol{x}) = i$, $r(\boldsymbol{x}) = j$ where $j \leqslant i$. Then,*

$$w_H\left(\mathcal{T}_{[m_1,m_2]}(\pi(\boldsymbol{x}))\right) = \binom{n}{m_1}\binom{n-m_1}{m_2} - \sum_{s \equiv t+s'+t' \equiv u \,(\bmod\, 2)} \binom{i-j}{s}\binom{j}{t}\binom{j}{u} \times \binom{n-i-j}{m_1-s-t-u}\binom{i-j-s}{s'}\binom{j-t}{t'} \times \binom{n-i-u}{m_2-s'-t'}$$

where $\binom{a}{b} = 0$ whenever $a \leqslant 0, b < 0, b > a$.

Proof: Let $\mathcal{T}_{[m_1,m_2]}(\pi(\boldsymbol{x})) = \boldsymbol{y} = (y_0, \ldots, y_{N-1})$. We would like to calculate the size $|\{\ell \in [N] \ : \ y_\ell = (0,0)\}|$ and obtain the Hamming weight by subtracting this amount from N. From Claim 1, we have that in $\pi(\boldsymbol{x})$ there are $i-j$ positions containing $(1,1)$, j positions containing $(0,1)$ and j positions containing $(1,0)$. Fix $\ell \in [N]$ and let $s = |\left\{k \in \mathcal{J}_{\ell,1} : \pi(\boldsymbol{x})_k = (1,1)\right\}|$, $t = |\{k \in \mathcal{J}_{\ell,1} : \pi(\boldsymbol{x})_k = (0,1)\}|$, $u = |\{k \in \mathcal{J}_{\ell,1} : \pi(\boldsymbol{x})_k = (1,0)\}|$. We have that

$$\sum_{k_1=1}^{s}(1,1) + \sum_{k_2=1}^{t}(0,1) + \sum_{k_3=1}^{u}(1,0) \equiv (0,0) \bmod 2$$

iff $s \equiv t \equiv u \pmod 2$. Now we do the same for the positions in $\mathcal{J}_{\ell,2}$. Let $s' = |\{k \in \mathcal{J}_{\ell,2} \ : \ \pi(\boldsymbol{x})_k = (1,1)\}|$, $t' = |\{k \in \mathcal{J}_{\ell,2} \ : \ \pi(\boldsymbol{x})_k = (0,1)\}|$, $u' = |\{k \in \mathcal{J}_{\ell,2} \ : \ \pi(\boldsymbol{x})_k = (1,0)\}|$. Since the sum on the positions in $\mathcal{J}_{\ell,2}$ contributes only to the $(0,1)$ term, we obtain that

$$\sum_{k_1=1}^{s}(1,1) + \sum_{k_2=1}^{t+s'+t'}(0,1) + \sum_{k_3=1}^{u}(1,0) \equiv (0,0) \bmod 2$$

iff $s \equiv t + s' + t' \equiv u \pmod 2$. Since $\binom{i-j}{s}\binom{j}{t}\binom{j}{u}\binom{n-i-j}{m_1-s-t-u}$ is equal to the number of ways of picking the m_1 positions for the set $\mathcal{J}_{\ell,1}$ (where s positions in the set have value $(1,1)$, t positions have value $(0,1)$, and

u positions have value $(1,0)$), and $\binom{i-j-s}{s'}\binom{j-t}{t'}\binom{n-i-u}{m_2-s'-t'}$ is equal to the number of ways of picking the m_2 positions for the set $\mathcal{J}_{\ell,2}$, we get that $|\{\ell \in [N] \ : \ y_\ell = (0,0)\}|$ can be calculated as

$$\sum_{s,t,s',t',u=1}^{n} \mathbb{1}_{s\equiv t+s'+t'\equiv u \pmod 2}\binom{i-j}{s}\binom{j}{t}\binom{j}{u}\times \binom{n-i-j}{m_1-s-t-u}\cdot\binom{i-j-s}{s'}\binom{j-t}{t'}\times \binom{n-i-u}{m_2-s'-t'}$$

which gives the desired result. ∎

Notice from the previous lemma that $w_H(\mathcal{T}_{[m_1,m_2]}(\pi(\boldsymbol{x})))$ only depends on $w_H(\boldsymbol{x}) = i, r(\boldsymbol{x}) = j$. Therefore, for any $\boldsymbol{x}$, we can characterize $w_H(\mathcal{T}_{[m_1,m_2]}(\pi(\boldsymbol{x})))$ by the following function

$$w_H(i,j) \triangleq \binom{n}{m_1}\binom{n-m_1}{m_2} - \sum_{s\equiv t+s'+t'\equiv u(\bmod 2)}\binom{i-j}{s}\binom{j}{t}\binom{j}{u}\times \binom{n-i-j}{m_1-s-t-u}\cdot\binom{i-j-s}{s'}\times \binom{j-t}{t'}\binom{n-i-u}{m_2-s'-t'}.$$

where i, j are non-negative integers and $j \leqslant i$.

Before stating the main result of this section, we need one more lemma which requires the following notation. For a pair-code $\mathcal{C}_p \in \left(\mathbb{F}_2^2\right)^n$, let a_i denote the number of codewords with pair-weight i, i.e., $a_i = |\{\boldsymbol{x} \in \mathcal{C}_p \ : \ w_p(\boldsymbol{x}) = i\}|$. Considering $\mathcal{C}_p$ as a binary expansion of a code over $\{0,1,2,3\}$ we obtain the following lemma from the Plotkin bound.

Lemma 7. *For a code $\mathcal{C}_p \subseteq \left(\mathbb{F}_2^2\right)^n$, we have*

$$\sum_{i=0}^{n-1}(n-\frac{4}{3}i)a_i \geqslant 0.$$

Proof: Similar to the proof of Lemma 3, let $\mathcal{C}_p \subseteq \left(\mathbb{F}_2^2\right)^n$ be a code of size $|\mathcal{C}_p| = M$, and let $\mathcal{M}$ be the codeword matrix which contains as rows all the codewords from $\mathcal{C}_p$. For shorthand, let $m_{i,j}$ denote the element in row i and column j of $\mathcal{M}$. For $j \in [n]$, let $w_{0,j} = |\{k : m_{k,j} = (0,0)\}|$, $w_{1,j} = |\{k : m_{k,j} = (0,1)\}|$, $w_{2,j} = |\{k : m_{k,j} = (1,0)\}|$, and $w_{3,j} = |\{k : m_{k,j} = (1,1)\}|$. Then we have

$$\begin{aligned} M\sum_{i=0}^{n} i a_i &= \sum_{\boldsymbol{c},\boldsymbol{d}} d_H(\boldsymbol{c},\boldsymbol{d}) \\ &= 2\big(\sum_{j=1}^{n} w_{1,j}w_{2,j} + w_{1,j}w_{3,j} + w_{1,j}w_{0,j} \\ &\quad + w_{2,j}w_{3,j} + w_{2,j}w_{0,j} + w_{3,j}w_{0,j}\big) \end{aligned}$$

where $\boldsymbol{c}, \boldsymbol{d} \in \left(\mathbb{F}_2^2\right)^N$. To maximize the above expression we set $w_{i,j} = M/4$ for $i \in \{0,1,2,3\}$ so that $M\sum_{i=0}^{n} i a_i \leqslant \frac{12\ M^2\ n}{16}$ which implies

$$\sum_{i=0}^{n} i a_i \leqslant \frac{12\ Mn}{16} = \frac{3n}{4}\sum_{i=0}^{n} a_i,$$

so that $\frac{3n}{4}\sum_{i=0}^{n} a_i - \sum_{i=0}^{n} i a_i \geqslant 0$ which after simplifying gives the expression in the lemma. ∎

We can now derive the main result of this section by applying the previous lemma to the pair-code $\mathcal{T}_{[m_1,m_2]}(\pi(\mathcal{C}))$. To simplify the notation, let

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

Lemma 8. *For a positive integer n and non-negative integers m_1, m_2 where $0 \leqslant m_1 + m_2 \leqslant n$, we have*

$$\sum_{j\leqslant i} K_{m_1,m_2}(n,i,j)a_{i,j} \geqslant 0.$$

Proof: From Lemma 7 we can write

$$\begin{aligned} 0 &\leqslant \sum_{r=0}^{N}(N-\frac{4}{3}r)a_r \\ &= \sum_{r=0}^{N}\left(\binom{n}{m_1}\binom{n-m_1}{m_2} - \frac{4}{3}r\right)\sum_{w_H(i,j)=r} a_{i,j} \\ &= \sum_{r=0}^{N}\sum_{w_H(i,j)=r}\left(\binom{n}{m_1}\binom{n-m_1}{m_2} - \frac{4}{3}r\right)a_{i,j} \\ &= \sum_{j\leqslant i}\left(\binom{n}{m_1}\binom{n-m_1}{m_2} - \frac{4}{3}w_H(i,j)\right)a_{i,j} \\ &= \sum_{j\leqslant i} a_{i,j}\left(\frac{4}{3}\sum_{s\equiv t+s'+t'\equiv u(\bmod 2)}\binom{i-j}{s}\binom{j}{t}\binom{j}{u}\binom{n-i-j}{m_1-s-t-u}\cdot\binom{i-j-s}{s'}\binom{j-t}{t'}\binom{n-i-u}{m_2-s'-t'} - \frac{1}{3}\binom{n}{m_1}\binom{n-m_1}{m_2}\right), \end{aligned}$$

which simplifies to the expression in the statement of the theorem. ∎

We now may state the linear programming bound. For integers n, m where $0 \leqslant m \leqslant n$, let

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

TABLE I

COMPARISON OF THE RESULTS OF THEOREM 9 AND THEOREM 5 WITH THE SPHERE PACKING BOUND [3]

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

Theorem 9. *Suppose $\mathcal{C} \subseteq \mathbb{F}_2^n$ with $d_p(\mathcal{C}) \geqslant d$. Then, $|\mathcal{C}|$ is upper bounded by the following expression*

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

Proof: The fact that $a_{0,0} = 1$ follows since there is only one codeword of weight zero. Furthermore, if $i + j < d$, then as a result of the minimum pair-distance of the code $\mathcal{C}$, $a_{i,j} = 0$. Since $\sum_{j=1}^{i} a_{i,j}$ represents the set of all codewords of weight i, the third constraint is a direct consequence of the linear programming bound found in [8] for instance. The last constraint follows directly from Lemma 8. ∎

In order to evaluate the new bounds we derived in this section, we compared in Table I our results with the sphere-packing upper bound from [3], for $2 \leqslant n, d_p \leqslant 7$. 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 Johnson bound and the second number represents the sphere-packing upper bound from [3]. For even distances we used our result from Theorem 5 and for odd distances Theorem 9.

IV. CONSTRUCTIONS FOR SMALL MINIMUM PAIR-DISTANCE

In this section, we present new code constructions for the pair-symbol channel. Table II compares between the sizes of the best known code constructions for the pair-symbol channel (in terms of their code lengths) and the best known upper bound on the codes cardinalities. For odd pair distance we used the sphere-packing upper bound from [3] and for even pair distance we used Theorem 5. We have highlighted the contributions of this section by $(*)$. The fourth column Table II, labeled as "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 $d_p = 6, 7, 10$.

Note that the case of $d_p = 2, 3$ are trivial and are known. For the case where $d_p = 2$, since every code has minimum pair-distance at least 2, $A_p(n, 2) = 2^n$. For the case where $d_p = 3$, it can be shown that code with a single parity bit is optimal, which implies $A_p(n, 3) = 2^{n-1}$. For $d_p = 4$, the best known construction is based on two interleaved simple parity codes which results with codes of cardinality $2^n/4$, when n is even. Note that the upper bound from Theorem 5 on the code cardinality in this case is $2^n/3$, and thus this case is still not fully solved (though it is solved for linear codes).

In general, if we were to apply the best known codes 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 $d_H \geqslant \left\lfloor \frac{2d_p+1}{3} \right\rfloor$ (see [21]). Hence, the redundancy of these codes will be roughly

$$\left\lfloor \frac{\left\lfloor \frac{2d_p+1}{3} \right\rfloor - 1}{2} \right\rfloor \log(n) = \left\lfloor \frac{d_p - 1}{3} \right\rfloor \log(n), \qquad (12)$$

which is already close to the lower bound in (8), which states that the redundancy has to be at least $\lfloor (d_p - 1)/4 \rfloor \log(n)$. Our goal in this section is to improve this construction for $d_p = 6, 7, 10$.

Throughout this section we represent codewords also as polynomials. For a word $\boldsymbol{c} = (c_0, \ldots, c_{n-1}) \in \mathbb{F}_q^n$, the polynomial representation $\boldsymbol{c}(x)$ of $\boldsymbol{c}$ (with the indeterminate x) is given by

$$\boldsymbol{c} = \boldsymbol{c}(x) = \sum_{i=0}^{n-1} c_i x^i \in \mathbb{F}_q[x].$$

A. Codes with Minimum Pair-Distance Six

We now show how to construct codes with an even length n and pair-distance equals to 6. Let $\{0, 1, \alpha, \alpha^2\} = \mathbb{F}_4$ and define the map $\Pi : \mathbb{F}_4 \to \mathbb{F}_2^2$ so that $\Pi(0) = (0,0), \Pi(1) = (0,1), \Pi(\alpha) = (1,0), \Pi(\alpha^2) = (1,1)$. Clearly, the map Π is invertible. For a sequence $\boldsymbol{x} = (x_0, x_1, \ldots, x_{\frac{n}{2}-1}) \in \mathbb{F}_4^{\frac{n}{2}}$, let $\Pi(\boldsymbol{x}) = (\Pi(x_0), \ldots, \Pi(x_{\frac{n}{2}-1})) \in \left(\mathbb{F}_2^2\right)^{\frac{n}{2}}$. Note that $\left(\mathbb{F}_2^2\right)^{\frac{n}{2}}$ is isomorphic to $\mathbb{F}_2^n$ and hence we may consider a sequence $\boldsymbol{x} \in \mathbb{F}_4^{\frac{n}{2}}$ as a sequence $\Pi(x) \in \mathbb{F}_2^n$. 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 $\mathcal{Z}$. The following useful claim may be regarded as a slightly generalized version of Theorem 1 in [21], since it holds for all linear codes which do not have to be cyclic.

Claim 2 *Suppose n is an even integer and let $\mathcal{C} \subseteq \mathbb{F}_4^{\frac{n}{2}}$ be a linear code such that $d_H(\mathcal{C}) \geqslant d$. Then, $d_p(\Pi(\mathcal{C})) \geqslant \lceil \frac{3d}{2} \rceil$.*

Proof: Since $\mathcal{C}$ is a linear code, we can determine the minimum pair-distance of the code $\Pi(\mathcal{C})$ by considering the minimum pair-weight of a sequence in $\Pi(\mathcal{C})$. We show the result by proving that there does not exist a sequence in $\Pi(\mathcal{C})$ with pair-weight less than $\lceil \frac{3d}{2} \rceil$. Let $\boldsymbol{z} = (z_0, \ldots z_{n-1}) = \Pi(\boldsymbol{x}) \in \mathbb{F}_2^n$, $\boldsymbol{x} \in \mathcal{C}$. Let $\mathcal{I} \subseteq [\frac{n}{2}]$ denote the non-zero indices

TABLE II
TABLE OF LARGEST KNOWN CODES AND UPPER BOUNDS

d_p	Upper Bound	Lower Bound on Code Size	Hamming Distance of the Code
4	2^n	2^{n-2} [3]	2 [3]
5	$\frac{2^n}{n+1}$	$\frac{2^n}{n+1}$ [3]	3 [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	$\frac{2^n}{\frac{n(n-1)+(4n+2)\lfloor \frac{n}{4} \rfloor}{2\lfloor \frac{n}{4} \rfloor}}$	$\frac{2^n}{(n+1)^2}$ [21]	5 [21]
9	$\frac{2^n}{\frac{1}{2}(2+3n+n^2)}$	$\frac{2^n}{2(n+1)^2}$	6
10	$\frac{2^n}{\frac{2n(n-3)+(2+3n+n^2)\lfloor \frac{n}{5} \rfloor}{2\lfloor \frac{n}{5} \rfloor}}$	$\frac{2^n}{2(n+1)^2}(*)$	$6(*)$

in $\boldsymbol{x}$ and suppose we partition $\mathcal{I}$ into 3 subsets $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ where for any $k \in \mathcal{I}_1$, we have $x_k = \alpha$, for any $k \in \mathcal{I}_2$, we have $x_k = \alpha^2$, and for any $k \in \mathcal{I}_3$, we have $x_k = 1$. Furthermore, let $i_1 = |\mathcal{I}_1|$, $i_2 = |\mathcal{I}_2|$, and $i_3 = |\mathcal{I}_3|$ where clearly $w_p(\boldsymbol{x}) = i_1 + i_2 + i_3$. From (3) we know that $w_p(\boldsymbol{z}) = w_H(\boldsymbol{z}) + r(\boldsymbol{z})$ where $r(\boldsymbol{z})$ is the number of runs in $\boldsymbol{z}$. Hence, $w_p(\boldsymbol{z}) \geqslant i_1 + 2i_2 + i_3 + r(\boldsymbol{z})$. Since a run of symbols may start with $\Pi(1) = 01$ and may end with $\Pi(\alpha^2) = 10$ we obtain $r(\boldsymbol{z}) \geqslant \left\lceil \frac{i_1+i_3}{2} \right\rceil$. Thus we have

$$w_p(\boldsymbol{z}) \geqslant 2i_2 + \left\lceil \frac{3(i_1+i_3)}{2} \right\rceil.$$

We can minimize the expression by setting $i_2 = 0$ and the lemma follows. ∎

We now describe the code using a parity-check matrix. In the following, we describe how to construct the parity check matrix. We begin by generating a set of matrices

$$H'^{(0)}, H'^{(1)}, \ldots, H'^{(m-4)}.$$

For $j \in [m-3]$, suppose $\alpha_{m-j-1} \in \mathbb{F}_{4^{m-j-1}}$ is a primitive element. Note that every power of α_{m-j-1} can be represented by a sequence over $\mathbb{F}_4$ of length $m-j-1$ which corresponds to the polynomial representation of α_{m-j-1}. We construct the matrix $H'^{(j)} \in \mathbb{F}_4^{m \times 4^{m-j-1}-1}$ as follows:

$$H'^{(j)} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 1 & \cdots & 1 \\ \alpha_{m-j-1} & \alpha_{m-j-1}^2 & \alpha_{m-j-1}^3 & \alpha_{m-j-1}^4 & \cdots & \alpha_{m-j-1}^{4^{m-j-1}-1} \end{bmatrix}, \tag{13}$$

where the first j rows of $H'^{(j)}$ are equal to zero, the $j+1$th row is all ones and the last $m-j-1$ rows are obtained by the sequence representation of α_{m-j-1}^{ℓ} for $\ell \in \{1, \ldots, 4^{m-j-1}-1\}$. Note that under this writing, $\alpha_{m-j-1}^{4^{m-j-1}-1}$ is the (multiplicative) unit of $\mathbb{F}_{4^{m-j-1}}$. Moreover, the 1 in $H'^{(j)}$ corresponds to the (multiplicative) unit in $\mathbb{F}_4$. We now construct a set of matrices $H^{(j)}$, $j \in [m-3]$ by removing one column sequence from $H'^{(j)} = (\boldsymbol{h}_1'^{(j)}, \ldots \boldsymbol{h}_{4^{m-j-1}-1}'^{(j)})$. More precisely, let α be a primitive element in $\mathbb{F}_4$, we remove the column from the matrix $H'^{(j)}$ which equals to

$$\frac{1}{\alpha^3} \cdot \left(\alpha^3 \cdot \boldsymbol{h}_{4^{m-j-1}-1}'^{(j)} + \alpha \cdot \boldsymbol{h}_1'^{(j+1)} \right), \tag{14}$$

where we interpret the matrices cyclically so that $H'^{(m-3)} = H'^{(0)}$. We denote the resulting matrix as $H^{(j)}$ which we index as $H^{(j)} = (\boldsymbol{h}_1^{(j)}, \ldots, \boldsymbol{h}_{4^{m-j}-2}^{(j)})$ for $j \in [m-3]$. Notice that the sequence in (14) is not equal to the first column of $H'^{(j)}$ (for any j) since, according to (13) (considering the last $m-j-1$ components), we have $\boldsymbol{h}_1'^{(j)} = \alpha_{m-j-1}$ and $\alpha^3 \cdot \boldsymbol{h}_{4^{m-j-1}-1}'^{(j)} + \alpha \cdot \boldsymbol{h}_1'^{(j+1)} = \alpha^3 \cdot 1 + \alpha \cdot \alpha_{m-j} \neq \alpha^3 \cdot \alpha_{m-j-1}$ where the last inequality holds since α_{m-j} has a different order than $\alpha^2(\alpha_{m-j-1} - 1)$. Also note that 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. Moreover, for $j_1 \neq j_2 \in [m-3]$, $H^{(j_1)}$ and $H^{(j_2)}$ have different sizes. Next, we form the matrix

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

where $N = \frac{4^m - 4^3}{3} - 2m + 6$.

The matrix H satisfies the following properties which are proved in Appendix:

Lemma 10.

1) *For any non-zero $\boldsymbol{z} \in \mathbb{F}_4^N$, if $H \cdot \boldsymbol{z} = \boldsymbol{0}$, 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 $H^{(j)} \cdot \boldsymbol{z} = 0$, $r(\boldsymbol{z}) \geqslant 2$.*
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 $\mathcal{C}_6 \subseteq \mathbb{F}_4^N$ is defined as

$$\mathcal{C}_6 \triangleq \{\boldsymbol{c} \in \mathbb{F}_4^N : H \cdot \boldsymbol{c} = \boldsymbol{0}\}.$$

The following theorem uses Claim 2 along with Lemma 10 to produce codes with minimum pair-distance six.

Theorem 11. *The code $\mathcal{C}_6$ satisfies $d_p(\Pi(\mathcal{C}_6)) \geqslant 6$.*

Proof: From property 1) of the matrix H, we know $d_H(\mathcal{C}) \geqslant 3$ and so by Claim 2, we have that $d_p(\Pi(\mathcal{C})) \geqslant 5$. We proceed by assuming there exists a codeword $\boldsymbol{y} \in \mathbb{F}_4^N$ where $w_p(\Pi(\boldsymbol{y})) = 5$ and we will arrive at a contradiction.

If $w_p(\Pi(\boldsymbol{y})) = 5$, then since $d_H(\mathcal{C}) = 3$, we have two possible options. Either $w_H(\Pi(\boldsymbol{y})) = 3$ and $r(\Pi(\boldsymbol{y})) = 2$, or $w_H(\Pi(\boldsymbol{y})) = 4$ and $r(\Pi(\boldsymbol{y})) = 1$. We first consider the case in which $w_H(\Pi(\boldsymbol{y})) = 4$ and $r(\Pi(\boldsymbol{y})) = 1$. Thus, the sequence $\Pi(\boldsymbol{y})$ contains 011110 as a subsequence. Under this setup, $w_H(\boldsymbol{y}) \geqslant 3$, and so $\boldsymbol{y}$ contains the substring $1, \alpha^2, \alpha$. Since the symbols $1, \alpha, \alpha^2$ appear consecutively in $\boldsymbol{y}$, it follows that we can invoke property 2) and conclude that $r(\boldsymbol{y}) \geqslant 2$.

Now suppose that $w_H(\Pi(\boldsymbol{y})) = 3$ and $r(\Pi(\boldsymbol{y})) = 2$ and assume that the three non-zero symbols in $\boldsymbol{y}$ are in positions which multiply only columns in $H^{(j)}$ for some $j \in [m-3]$. For instance, if $j = 1$, this assumption implies that all three non-zero symbols in $\boldsymbol{y}$ are contained within the first $4^m - 2$ positions of $\boldsymbol{y}$. Then we have that for some sequence $\boldsymbol{v}$,

$$H^{(j)} \cdot \boldsymbol{v} = \boldsymbol{0},$$

where $H^{(j)}$ contains a row that is all-ones. This implies that the non-zero symbols of $\boldsymbol{v}$ are α, α^2, and 1 which in turn implies the non-zero symbols of $\boldsymbol{y}$ are α, α^2, and 1. Under the map Π, this implies that $w_H(\Pi(\boldsymbol{v})) = w_H(\Pi(\boldsymbol{y})) = 4$. We now would like to apply Property 2). The fact that the conditions are met is proved in Lemma 16 which appears in the appendix. From property 2) of the matrix H, we have $r(\boldsymbol{v}) = r(\boldsymbol{y}) \geqslant 2$ which implies $r(\Pi(\boldsymbol{v})) = r(\Pi(\boldsymbol{y})) \geqslant 2$ and hence $d_p(\Pi(\boldsymbol{y})) \geqslant 6$ and we arrive at a contradiction in this case.

Suppose now that the three non-zero symbols in $\Pi(\boldsymbol{y})$ are in positions which multiply columns in different sub-matrices of H. Suppose first that $\boldsymbol{y}$ contains non-zero components that multiply 3 sub-matrices in H. In other words, we would have

$$\boldsymbol{0} = \gamma_1 \cdot \boldsymbol{h}_{k_1}^{(j_1)} + \gamma_2 \cdot \boldsymbol{h}_{k_2}^{(j_2)} + \gamma_3 \cdot \boldsymbol{h}_{k_3}^{(j_3)}. \tag{15}$$

where $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{F}_4$ and $|\{j_1, j_2, j_3\}| = 3$. Recall that $\boldsymbol{h}_k^{(j)}$ is zero in the first $j-1$ components and 1 in the j-th component so that if $|\{j_1, j_2, j_3\}| = 3$, then (15) cannot hold. Suppose then that $|\{j_1, j_2, j_3\}| = 2$, and without loss of generality, $j_1 = j_2$. By the structure of the sequences $\boldsymbol{h}_{k_1}^{(j_1)}$ (in particular, since $\boldsymbol{h}_{k_1}^{(j_1)}$ has a row which is all-ones), this implies $\gamma_1 = \gamma_2$ and so if $r(\Pi(\boldsymbol{y})) = 2$, then either:

$$\boldsymbol{h}_{2^{m-j_1}-2}^{(j_1)} + \alpha \cdot \boldsymbol{h}_1^{(j_1+1)} = \boldsymbol{h}_{k_2}^{(j_1)}, \tag{16}$$

or

$$\boldsymbol{h}_{2^{m-j_1}-2}^{(j_1)} = \alpha \cdot \boldsymbol{h}_1^{(j_1+1)} + \alpha \cdot \boldsymbol{h}_{k_2}^{(j_1+1)}. \tag{17}$$

Note that (17) cannot be satisfied since $\boldsymbol{h}_{2^{m-j_1}-2}^{(j_1)}$ has 1 in a position in which both $\alpha \cdot \boldsymbol{h}_1^{(j_1+1)}$ and $\alpha \cdot \boldsymbol{h}_{k_2}^{(j_1+1)}$ have zero. As a consequence, we are left with (16). However, this contradicts property 3) of H and so we arrive at a contradiction in this case as well. ∎

We consider the cardinality of the code $\Pi(\mathcal{C}_6)$ and compare it with the previously best known codes. Since $\mathcal{C}_6$ is a linear code with parity check matrix of dimension m, $|\mathcal{C}_6| = \frac{4^N}{4^m}$. Since $N = \frac{4^m - 4^3}{3} - 2m + 6$ where $m \leqslant \log_4(3N + 4^3)$ and since the mapping Π creates codes of length $2N$, setting $n = 2N$ gives:

$$|\mathcal{C}| \geqslant \frac{2^n}{\frac{3n}{2} + 4^3}.$$

The best known construction for codes with $d_p \geqslant 6$ requires a binary cyclic code with Hamming distance at least 4 [21]. Hence the cardinality of that code would be at least $\frac{2^n}{2n}$.

B. Codes with Minimum Pair-Distance Seven

In this subsection we turn to the construction of codes with minimum pair-distance seven. Our point of departure for this construction is a known result from [1] on burst-correcting codes. A code will be called a *b-burst-correcting code* if it can correct any error pattern which is confined to at most b consecutive locations. According to [1], it is known that for every even $m \geqslant 4$, there exists a cyclic 3-burst-correcting code of length $2^m - 1$ with a generator polynomial of the form $(1 + x + x^2)p(x)$, where $p(x)$ is a primitive polynomial. The connection between three-burst-correcting codes and symbol-pair codes with minimum pair distance seven is established in the next theorem.

Theorem 12. *Let $n = 2^m - 1$ where $m \geqslant 4$ is an even integer. Then, there exists a binary cyclic code $\mathcal{C}_7$ of length n such that $d_p(\mathcal{C}_7) \geqslant 7$.*

Proof: Recall from [1] that there exists a cyclic code $\mathcal{C}_7$ of length $2^m - 1$ with a generator polynomial of the form $(1 + x + x^2)p(x)$ where $p(x)$ is primitive that can correct a single burst of errors of length at most 3. Since the minimum distance of the code $\mathcal{C}_7$ is 3, it follows from (5) that any codeword in $\mathcal{C}_7$ with $d_H \geqslant 3$ has at least two runs in it. Since the code is cyclic, it therefore suffices to show that there does not exist a codeword $c(x) \in \mathcal{C}_7$ where we can write $c(x) = B_1(x) + x^\ell B_2(x) \bmod x^n - 1$ for a positive integer ℓ when $(B_1(x), B_2(x)) \in \{(1, 1 + x)(1, 1 + x + x^2), (1 + x, 1 + x)\}$. The result follows by noting that if, on the contrary, $c(x)$ could be written as $c(x) = B_1(x) + x^\ell B_2(x) \bmod x^n - 1$ then the code $\mathcal{C}_7$ could not correct a burst of errors of length 3. This is due to the fact that a linear code is a 3-burst error correcting code iff every codeword (except the zero codeword) is not the sum of two or less cyclic burst of length at most 3 [1]. ∎

For $n = 2^m - 1$, let $\mathcal{C}_7$ be the code constructed according to Theorem 12, so that $d_p(\mathcal{C}) \geqslant 7$. Since the degree of a primitive polynomial is m, this implies that the generator polynomial has degree of $m + 2$ which in turn implies that

$$|\mathcal{C}_7| = \frac{2^n}{4(1 + n)},$$

since $m = \log_2(n + 1)$. If we were to construct a code with minimum pair distance seven using the result from [21], its minimum Hamming distance will have to be at least five. Hence, by the sphere packing bound, this implies that the cardinality of the resulting code will be at most

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

Furthermore, by the sphere packing bound for symbol-pair codes [3], the largest cardinality of a code with minimum pair distance seven is at most $\frac{2^n}{1+2n}$. Therefore, the redundancy of the code $\mathcal{C}_7$ is roughly at most a single bit from optimality.

C. Codes with Minimum Pair-Distance Ten

Now, we turn to constructing codes that have pair-distance 10. We will use the following property of cyclic codes.

Lemma 13. *(c.f., [14]) Suppose $\mathcal{C}$ is a code with a self-reciprocal generator polynomial. Then the codewords in $\mathcal{C}$ are reversible.*

The following lemma will be used to prove our main result.

Lemma 14. *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 $\{\alpha, \alpha^{-1}\} \subset \mathbb{F}_{2^m}$ where α is a primitive element of $\mathbb{F}_{2^m}$. Then $d_H(\mathcal{C}) \geqslant 4$.*

Proof: Since $g(x)$ has the primitive element α as a root, $d_H(\mathcal{C}) \geqslant 3$ from the BCH bound [16]. Suppose, on the contrary, that $d_H(\mathcal{C}) = 3$. Then there exists a codeword $c(x) \in \mathcal{C}$ where $c(x) = 1 + x^i + x^j$ for $j > i$. Since $g(x)$ is self-reciprocal we know from Lemma 13 that the codewords in $\mathcal{C}$ are reversible, and so there exists a codeword $c'(x) = 1 + x^{j-i} + x^j$. But then $c(x) + c'(x) = x^i + x^{j-i} \in \mathcal{C}$, which is a contradiction since $d_H(\mathcal{C}) \geqslant 3$. ∎

Using the previous lemma, we can now prove the main result of this section.

Theorem 15. *For a positive integer $m > 2$, let $g(x) \in \mathbb{F}_2[x]$ be a generator polynomial for a cyclic code $\mathcal{C}_{10} \subseteq \mathbb{F}_2^{2^m-1}$ 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 $d_p(\mathcal{C}_{10}) \geqslant 10$.*

Proof: Let $n = 2^m - 1$. From the BCH bound [16] we obtain $d_H(\mathcal{C}_{10}) \geqslant 6$ (note that if α (α^{-1}) is a root of the generator polynomial then also all of its conjugates which include α^2 (α^{-2})). Hence, $d_p(\mathcal{C}_{10}) \geqslant 9$ from (5). Suppose, on the contrary that $d_p(\mathcal{C}_{10}) = 9$ and let $c(x) \in \mathcal{C}_{10}$ be a codeword in $\mathcal{C}_{10}$ with pair-weight 9. Then, there are three possibilities to consider. Either a) $c(x)$ is comprised of 3 runs each of length 2, or b) $c(x)$ is comprised of 3 runs where there is a run of length 1, 2, and 3 in $c(x)$, or c) $c(x)$ is comprised of 3 runs where there are 2 runs of length 1 and one run of length 4.

We first consider case a) so assume $c(x)$ is comprised of 3 runs each of length 2. Then we can write $c(x) = (1+x)(1+x^i+x^j)$ for $i \geqslant 3$ and $j - i \geqslant 3$. Since $g(x)$ is the generator polynomial for $\mathcal{C}_{10}$, $g(x)|c(x)$ and so

$$(1+x)(1+x^j+x^j) = q(x) \cdot (x+\alpha^{-1}) \cdot (x+1) \cdot (x+\alpha) \bmod x^n + 1$$

where $q(x) \in \mathbb{F}_2[x]$. This implies $(x+\alpha) \cdot (x+\alpha^{-1}) | (1+x^i+x^j)$ which is a contradiction since from Lemma 14, a code with a generator polynomial $(x+\alpha) \cdot (x+\alpha^{-1})$ has minimum Hamming distance at least 4.

Assume case b) holds and that $c(x) = 1+x+x^2+x^i+x^j+x^{j+1} \notin \mathcal{C}$ where $i \geqslant 4$ and $j-i \geqslant 2$. Then, the reverse of $c(x)$, $c'(x) = 1+x+x^{j-i+1}+x^{j-1}+x^j+x^{j+1} \in \mathcal{C}$. However, then $c(x)+c'(x) \in \mathcal{C}_{10}$ where $c(x)+c'(x) = x^2+x^i+x^{j-i+1}+x^{j-1}$ which is a contradiction since $d_H(\mathcal{C}_{10}) \geqslant 6$.

In the previous paragraph we shows that $\mathcal{C}_{10}$ does not contain any codewords of weight 6 where the first run has length 3 the next run has length 1 and the last run has length 2. Denote the set of sequences of Hamming weight 6 and length n where the first run has length 3 the next run has length 1 and the last run has length 2 as $V(3,1,2)$. In particular, we showed in the previous paragraph that $V(3,1,2) \cap \mathcal{C}_{10} = \emptyset$ and also that $V(2,1,3) \cap \mathcal{C}_{10} = \emptyset$. Since the code $\mathcal{C}_{10}$ is cyclic, we can also conclude that $V(1,2,3) \cap \mathcal{C}_{10} = \emptyset$ and $V(2,3,1) \cap \mathcal{C}_{10} = \emptyset$ from the fact that $V(3,1,2) \cap \mathcal{C}_{10} = \emptyset$ and also that $V(1,3,2) \cap \mathcal{C}_{10} = \emptyset$, $V(3,2,1) \cap \mathcal{C}_{10} = \emptyset$ from $V(2,1,3) \cap \mathcal{C}_{10} = \emptyset$. Thus, $\mathcal{C}_{10}$ cannot contain any codewords of weight 6 comprised of runs of length 1, 2, and 3.

Assume case c) holds and $c(x) = 1 + x^i + x^{i+1} + x^{i+2} + x^{i+3} + x^j \in \mathcal{C}_{10}$ where $i \geqslant 3$ and $j - i \geqslant 5$. Since the codewords of $\mathcal{C}_{10}$ are reversible we know $c'(x) = 1+x^{j-i-3}+x^{j-i-2}+x^{j-i-1}+x^{j-i}+x^j \in \mathcal{C}_{10}$. However, then $c(x) + c'(x) = x^i+x^{i+1}+x^{i+2}+x^{i+3}+x^{j-i-3}+x^{j-i-2}+x^{j-i-1}+x^{j-i} \in \mathcal{C}_{10}$ which is a contradiction since any codeword in $\mathcal{C}_{10}$ has at least 3 runs from Lemma 2. Using similar logic as the previous paragraph, it can be shown that since $\mathcal{C}_{10}$ is a linear, cyclic code with reversible codewords case c) cannot hold. ∎

In order to evaluate the redundancy of the code, we note that since α and α^{-1} are primitive, the degree of the generator polynomial is $2m+1$. Since $m = \log_2(n+1)$ we obtain that the size of the code from Theorem 15 is $\frac{2^n}{2(n+1)^2}$, that is, its redundancy is $2\log_2(n)+1$. This follows since α (α^{-1}) is primitive and hence it corresponds to a polynomial of degree m. This implies that the degree of the generator polynomial of the code $\mathcal{C}_{10}$ is $2m+1$. Note that the best known construction requires the Hamming distance of the code to be at least seven and thus its redundancy is approximately $3\log_2(n)$ [21]. Lastly, according to the sphere packing bound or Theorem 5, the minimum redundancy of a code with minimum pair distance ten is at least $2\log_2(n+1)+c$, for some constant c. Hence, the redundancy of the code $\mathcal{C}_{10}$ is at most a fixed number of bits away from optimality.

V. Conclusions and Discussions

In this work we studied symbol-pair codes for the symbol-pair read channel. First we derived new upper bounds on the code cardinalities. Our new bounds improved upon the best known bound which was based on the sphere packing bound [3]. The new bounds are adaptations of the Johnson bound and the linear programming bound to the symbol-pair read channel. We then showed how to improve the codes from [21] when the minimum pair distance is six, seven, and ten. Our constructions are almost optimal in the sense that the redundancy of our code constructions is at most a constant number of bits away from the lower bound on the redundancy.

There are several new ideas that were introduced in the paper. For the Johnson bound, we derive an upper bound on the

number of codewords of weight w and pair-weight $2w$ based on the idea that symbols and positions "occupy" columns of the matrix M which contains all the codewords in the code. The departure from the traditional result in the Hamming metric here is that codewords need to occupy one additional column (where there are no symbols) to account for the fact that the pair distance is $2w$. For the linear programming bound, the mapping T on the output of the pair channel itself $\pi(\mathbf{x})$ was used. This created some surprising symmetries and we believe the technique may be applicable to other types of channels. One such relatively straightforward extension is to consider the technique on the b-symbol channel. Another potential application of the technique might be to consider it on the error-burst channel whereby codewords should have runs longer than the error-burst. The approach places conditions on the number of runs of the codewords since it uses the output of the pair channel. To the best of the our knowledge, there is no linear programming bound which places constraints on the structure of the ones in the underlying code. The classical LP bound simply places conditions on the number of ones.

An interesting generalization is to consider the q-ary alphabet. There are technical challenges to extending the bounds to the non-binary setup. For the Johnson bound, the main difficulty is in generalizing Theorem 5. Specifically, it requires the calculation of the following value: For a given codeword c with $d_p(c) = 2t + 1$, how many vectors x are there with $d_p(x) = t + 1$ such that $d_p(c - x) = t$? Regarding the LP bound, one of the principal challenges would be to determine an expression for the pair weight in terms of the number of runs of zeros and consecutive non-binary symbols. We believe that some new ideas would be required to make the derivations work. In addition, running the LP bound itself to produce results would be a challenge since, even for $\mathbb{F}_2$, the linear programming bound required a prohibitive amount of time to compute for code lengths beyond 7. The construction of codes with minimum pair distance 6 relies on mapping between $\mathbb{F}_4$ and $\mathbb{F}_2$ and it also relies on some simple properties of the elements of $\mathbb{F}_4$ (see Lemma 10). Hence, it will be interesting to extend this to the non-binary case.

APPENDIX
PROOF OF LEMMA 10

Before proving Lemma 10 we highlight a property of the matrices $H'^{(j)}$ which will be useful later. For a sequence $z \in \mathbb{F}_4^\ell$ let $l(z)$ denote the minimum length between the first and the last non-zero symbol in z, considered cyclically. For instance if $z = (0, 1, \alpha, 0, 0, \alpha^2, 0, 0, 0, 0, 0, 0) \in \mathbb{F}_4^{12}$, then $l(z) = \min\{5, 10, 12\} = 5$ since the the minimum is obtained when considering the first non-zero symbol in z to be in position 1 and the last non-zero symbol is in position 5. Note that taking z' to be the 2-cyclic left shift of z, we obtain $l(z') = 12$ and taking z'' to be the 5-cyclic left shift of z we obtain $l(z'') = 10$. Hence, $l(z) = \min\{5, 10, 12\} = 5$.

Lemma 16. *For a non-zero $z \in \mathbb{F}_4^{4^{m-j}-1}$, if $H'^{(j)} \cdot z = \mathbf{0}$, then $l(z) \geqslant 5$ and $w_H(z) \geqslant 3$.*

Proof: The proof follows by noting that $H'^{(j)}$ is a parity check matrix for a cyclic code. This code has $m - j - 1$ length codewords over $\mathbb{F}_4$. Note also that the minimum distance of the code is 3 since every two columns are linearly independent thanks to the row of ones and since α_{m-j-1} is primitive. We also have that the degree of the generator polynomial for the code is at least $m - j \geqslant 4$. Therefore, any codeword polynomial has degree at least 4 which implies $l(\boldsymbol{x}) \geqslant 5$ for any codeword $\boldsymbol{x}$. ∎

From Lemma 16 and noting that $H^{(j)}$ is the result of removing a single column from $H_H^{(j)}$, we can prove Lemma 10

Proof: [Proof of Lemma 10]

1) By the construction of H, it is clear that H has a stairs-like shape. First, it is clear that $w_H(z) \neq 1$. Assume $w_H(z) = 2$ and note that if $H \cdot z = 0$ then z must have 2 non-zero elements in positions which correspond to a specific matrix $H^{(j)}$ for some $j \in [m - 3]$. This is because each matrix $H^{(j)}$ has a row of ones. Moreover, those non-zero elements in z must be equal. Since the next lines contain powers of a primitive element, it results in $H \cdot z \neq 0$. Hence, $w_H(z) \geqslant 3$.
2) This is a direct consequence of Lemma 16. Since if z' is such that $H'^{(j)}z = 0$ then $l(z') \geqslant 5$. This means that if $H^{(j)}z = 0$ then $l(z) \geqslant 4$ and hence $w_H(z) = 3$ implies $r(z) \geqslant 2$.
3) This property follows directly from the construction of the matrix H.

∎

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers, whose comments helped improve the presentation of the paper.

REFERENCES

[1] K. Abdel-Ghaffar, R. McEliece, A. Odlyzko, and H. van Tilborg, "On the existence of optimum cyclic burst- correcting codes," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 6, pp. 768–775, Nov. 1986.
[2] Y. Cassuto and S. Litsyn, "Symbol-pair codes: Algebraic constructions and asymptotic bounds," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT2011)*, Saint Petersburg, Russia, Jul./Aug. 2011, pp. 2348–2352.
[3] Y. Cassuto and M. Blaum, "Codes for symbol-pair read channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 8011–8020, Dec. 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. Inf. Theory*, vol. 59, no. 11, pp. 7259–7267, Nov. 2013.
[5] B. Ding, G. Ge, J. Zhang, T. Zhang, and Y. Zhang, "New constructions of MDS symbol-pair codes," *Des. Codes Cryptogr.*, vol. 86, no. 4, pp. 841–859, Apr. 2018.
[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. Inf. Theory*, vol. 64, no. 4, pp. 2417–2430, Apr. 2018.
[7] O. Elishco, R. Gabrys, and E. Yaakobi, "Bounds and constructions of codes over symbol-pair read channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2505–2509.
[8] J. I. Hall. (2013). *Notes on Coding Theory*. [Online]. Available: http://www.mth.msu.edu/~jhall/classes/codenotes/coding-notes.html
[9] M. Hirotomo, M. Takita, and M. Morii, "Syndrome decoding of symbol-pair codes," in *Proc. IEEE Inf. Theory Workshop*, Hobart, TAS, Australia, Nov. 2014, pp. 162–166.
[10] S. Horii, T. Matsushima, and S. Hirasawa, "Linear programming decoding of binary linear codes for symbol-pair read channel," *IEICE Trans. Fund. Electron., Commun. Comput. Sci.*, vol. E99-A, no. 12, pp. 2170–2178, 2016.

[11] W. Huffman and V. Pless, *Fundamentals of Error Correcting Codes*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[12] X. Kai, S. Zhu, and P. Li, "A construction of new MDS symbol-pair codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5828–5834, Nov. 2015.
[13] S. Li and G. Ge, "Constructions of maximum distance separable symbol-pair codes using cyclic and constacyclic codes," *Des. Codes Cryptogr.*, vol. 84, no. 3, pp. 359–372, Sep. 2017.
[14] O. Milenkovic and N. Kashyap, "On the design of codes for DNA computing," in *Coding and Cryptography* (Lecture Notes in Computer Science), vol. 3969. Berlin, Germany: Springer, 2006, pp. 100–119.
[15] M. Morii, M. Hirotomo, and M. Takita, "Error-trapping decoding for cyclic codes over symbol-pair read channels," in *Proc. Int. Symp. Inf. Theory Appl.*, Monterey, CA, USA, Nov. 2016, pp. 681–685.
[16] R. Roth, *Introduction to Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[17] Z. Sun, S. Zhu, and L. Wang, "The symbol-pair distance distribution of a class of repeated-root cyclic codes over $\mathbb{F}_p^m$," *Cryptogr. Commun.*, vol. 10, no. 4, pp. 643–653, Nov. 2017.
[18] M. Takita, M. Hirotomo, and M. Morii, "A decoding algorithm for cyclic codes over symbol-pair read channels," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. E98-A, no. 12, pp. 2415–2422, 2015.
[19] 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. Fundam. Electron., Commun. Comput. Sci.*, vol. E99-A, no. 12, pp. 2179–2191, 2016.
[20] M. Takita, M. Hirotomo, and M. Morii, "Error-trapping decoding for cyclic codes over symbol-pair read channels," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. E100-A, no. 12, pp. 2578–2584, 2017.
[21] E. Yaakobi, J. Bruck, and P. H. Siegel, "Constructions and decoding of cyclic codes over b-symbol read channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1541–1551, Apr. 2016.

Ohad Elishco (S'12–M'18) received the B.Sc. degree in mathematics, the B.Sc. in electrical engineering, and the M.Sc. and Ph.D. degrees in electrical engineering from the Ben-Gurion University of the Negev, Israel, in 2012, 2013, and 2017, respectively. From 2017 to 2018, he held a post-doctoral researcher position with the Department of Electrical Engineering, Massachusetts Institute of Technology. He is currently a Post-Doctoral Researcher with the Department of Electrical Engineering, University of Maryland at College Park. His research interests are coding and dynamical systems.

Ryan Gabrys received the B.S. degree in mathematics and computer science from the University of Illinois at Urbana-Champaign in 2005 and the Ph.D. degree in electrical engineering from the University of California at Los Angeles, in 2014. He is currently a Scientist with the Naval Information Warfare Center. His research interests broadly lie in the areas of theoretical computer science and electrical engineering, including coding theory, combinatorics, and communication theory.

Eitan Yaakobi (S'07–M'12–SM'17) received the B.A. degree in computer science and mathematics and the M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California at San Diego, in 2011. From 2011 to 2013, he was a Postdoctoral Researcher with the Department of Electrical Engineering, California Institute of Technology and with the Center for Memory and Recording Research, University of California at San Diego. He is currently an Associate Professor with the Computer Science Department, Technion — Israel Institute of Technology. His research interests include information and coding theory with applications to non-volatile memories, associative memories, DNA storage, data storage and retrieval, and private information retrieval. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship in 2010.