Decoding of Cyclic Codes over Symbol-Pair Read Channels

Eitan Yaakobi*†, **Jehoshua Bruck***, and **Paul H. Siegel**†
*Electrical Engineering Department, California Institute of Technology, Pasadena, CA 91125, U.S.A
†Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA 92093, U.S.A.
{*yaakobi, bruck*}*@caltech.edu, psiegel@ucsd.edu*

***Abstract*—Symbol-pair read channels, in which the outputs of the read process are pairs of consecutive symbols, were recently studied by Cassuto and Blaum. This new paradigm is motivated by the limitations of the reading process in high density data storage systems. They studied error correction in this new paradigm, specifically, the relationship between the minimum Hamming distance of an error correcting code and the minimum pair distance, which is the minimum Hamming distance between symbol-pair vectors derived from codewords of the code. It was proved that for a linear cyclic code with minimum Hamming distance d_H, the corresponding minimum pair distance is at least $d_H + 3$.**

Our main contribution is proving that, for a given linear cyclic code with a minimum Hamming distance d_H, the minimum pair distance is at least $d_H + \left\lceil \frac{d_H}{2} \right\rceil$. We also describe decoding algorithms, based upon bounded distance decoders for the cyclic code, whose pair-symbol error correcting capabilities reflects the larger minimum pair distance. In addition, we consider the case where a read channel output is a prescribed number, $b > 2$, of consecutive symbols and provide some generalizations of our results. We note that the symbol-pair read channel problem is a special case of the sequence reconstruction problem that was introduced by Levenshtein.

I. Introduction

The traditional approach in information theory to analyze noisy channels is to parse the message into independent information units, called symbols. Even though in many works the error correlation and interference between the symbols is studied, the process of writing and reading each symbol is usually assumed to be performed independently. However, in some of today's storage technologies as well as future ones, this is no longer an accurate assumption and symbols can only be written and read together. This brings us to study a model, recently proposed by Cassuto and Blaum [1], for channels whose outputs are overlapping pairs of symbols.

The rapid progress in high density data storage technologies paved the way for high capacity storage with reduced price. However, since the bit size at high densities is so small, one of the fundamental problems is to successfully read the individual bits recorded on the storage medium; for more details, see [1]. The channel model studied by Cassuto and Blaum [1], and later by Cassuto and Litsyn [2], mimics the reading process of such storage technologies. On each reading operation, the value of two consecutive symbols is read, called a *pair-read symbol*. This new model changes the requirement on the error correction capability of error-correction codes. There is already a significant amount of redundancy as every symbol is read twice. Furthermore, the errors are no longer symbol errors, but, rather, *pair-symbol errors*, where in a pair-symbol error at least one of the symbols is erroneous. The main task now becomes to combat these pair-symbol errors by designing codes with large minimum pair distance.

The works in [1], [2] studied the case of pair-read symbols. However, this model can be easily generalized such that on every read operation, multiple, say $b > 2$, consecutive symbols are read and thus every symbol is read b times. In essence, we receive multiple estimations of the same stored word. This connection brings us to the sequence reconstruction problem, which was introduced by Levenshtein [5]–[7]. In this model, the same codeword is transmitted over multiple channels. Then, a decoder receives all channel outputs, which are guaranteed to be different from each other, and outputs an estimation of the transmitted word. The original motivation did not come from storage systems but rather from other fields, such as molecular biology and chemistry, where the amount of redundancy in the information is too low and thus the only way to combat errors is by repeatedly transmitting the same message. This model is very relevant for storage technologies we described above or any other storage where the stored information is read multiple times. Furthermore, we note that the model by Levenshtein was recently studied and extended in the context of associative memories [11].

In the channel model described by Levenshtein, all channels are (almost) independent from each other as it is only guaranteed that the channel outputs are all different. If the transmitted message $\boldsymbol{c}$ belongs to a code with Hamming distance d_H and the number of errors in every channel can be strictly greater than $\lfloor \frac{d_H-1}{2} \rfloor$, then Levenshtein studied the minimum number of channels that are necessary to construct a successful decoder. This value was studied in [6] for the Hamming metric as well as other distance metrics and was later analyzed for a distance metric over permutations, e.g. [3], [4], and error graphs [8].

For the Hamming distance, the following result was proved in [6]. Assume the transmitted word belongs to a code with minimum Hamming distance d_H and the number of errors, t, in every channel is greater than $\lfloor \frac{d_H-1}{2} \rfloor$. Then, in order to construct a successful decoder, the number of channels has to be greater than

$$\sum_{i=0}^{t-\lceil d_H/2 \rceil} \binom{n-d_H}{i} \sum_{k=i+d_H-t}^{t-i} \binom{d_H}{k}.$$

For example, if $t = \lfloor \frac{d_H-1}{2} \rfloor + 1$, i.e., only one more than the error correction capability, then the number of channels has to be at least $\binom{2t}{t} + 1$. If $t > \lfloor \frac{d_H-1}{2} \rfloor + 1$ then this number is even a function of the message length. This disappointing result is a consequence of the arbitrary errors that may occur in every channel. In practice, especially for storage systems, we can take advantage of the fact that the errors are constrained.

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

In the symbol-pair read channel, there are in fact two channels. If the stored information is $\boldsymbol{x} = (x_0, \dots, x_{n-1})$, then the *pair-read vector* of $\boldsymbol{x}$ is

$$\pi(\boldsymbol{x}) = [(x_0, x_1), (x_1, x_2), \dots, (x_{n-2}, x_{n-1}), (x_{n-1}, x_0)],$$

and the goal here is to correct a large number of the so-called *symbol-pair errors*. The *pair distance*, $d_p(\boldsymbol{x}, \boldsymbol{y})$, between two pair-read vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the Hamming distance between their pair read vectors, that is, $d_p(\boldsymbol{x}, \boldsymbol{y}) = d_H(\pi(\boldsymbol{x}), \pi(\boldsymbol{y}))$. Accordingly, the *minimum pair distance* of a code $\mathcal{C}$ is defined as $d_p(\mathcal{C}) = \min_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{C},\boldsymbol{x}\neq\boldsymbol{y}}\{d_p(\boldsymbol{x}, \boldsymbol{y})\}$. In [1], it was shown that for a linear cyclic code with minimum Hamming distance d_H, its minimum pair distance, d_p, satisfies $d_p \geqslant d_H + 3$. Our main contribution in this work is proving that

$$d_p \geqslant d_H + \left\lceil \frac{d_H}{2} \right\rceil.$$

According to [1], this permits correction of $\lceil 3d_H/4 \rceil - 1$ symbol-pair errors. Thus, in contrast to Levenshtein's results on independent channels, on the symbol-pair read channel we can correct a large number of symbol-pair errors. In order to establish this result, we explicitly construct a decoder that can correct this number of symbol-pair errors.

The rest of the paper is organized as follows. In Section II, we review the symbol-pair read channel and some basic properties. In Section III, we show that cyclic codes can correct a large number of symbol-pair errors and in Section IV, decoders for such codes are given. Section V generalizes some of the results on the symbol-pair read channel to channels that sense b consecutive symbols on each read, where $b > 2$. Finally, Section VI concludes the paper.

II. Definitions and Basic Properties

In this section, we briefly review the model and definition of the symbol-pair read channel. If a length-n vector is stored in the memory then its pair-read vector is also a length-n vector, while every entry consists of two consecutive symbols of the stored vector. More formally, if $\boldsymbol{x} = (x_0, \dots, x_{n-1}) \in \Sigma^n$ is a length-n vector over some alphabet Σ, then the *symbol-pair read vector* of $\boldsymbol{x}$, denoted by $\pi(\boldsymbol{x})$, is defined to be

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

We will focus in this work on binary vectors, so $\Sigma = \{0, 1\}$. Note that $\pi(\boldsymbol{x}) \in (\Sigma \times \Sigma)^n$, and for $\boldsymbol{x}, \boldsymbol{y} \in \Sigma$, $\pi(\boldsymbol{x} + \boldsymbol{y}) = \pi(\boldsymbol{x}) + \pi(\boldsymbol{y})$. Unless stated otherwise, in this paper, all indices are taken modulo n. The Hamming distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $d_H(\boldsymbol{x}, \boldsymbol{y})$. Similarly, the Hamming weight of a vector $\boldsymbol{x}$ is denoted by $w_H(\boldsymbol{x})$. The *pair distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $d_p(\boldsymbol{x}, \boldsymbol{y})$ and is defined to be

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = d_H(\pi(\boldsymbol{x}), \pi(\boldsymbol{y})).$$

Accordingly, the *pair weight* of $\boldsymbol{x}$ is $w_p(\boldsymbol{x}) = w_H(\pi(\boldsymbol{x}))$. A *symbol-pair error* in the i-th symbol of $\pi(\boldsymbol{x})$ changes at least one of the two bits (x_i, x_{i+1}). Note that the following connection between the pair distance and pair weight holds.

Proposition 1. *For all $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$, $d_p(\boldsymbol{x}, \boldsymbol{y}) = w_p(\boldsymbol{x} + \boldsymbol{y})$.*

A first observation on the connection between the Hamming distance and pair distance was proved in [1].

Proposition 2. *For $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^n$, let $0 < d_H(\boldsymbol{x}, \boldsymbol{y}) < n$ be the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then,*

$$d_H(\boldsymbol{x}, \boldsymbol{y}) + 1 \leqslant d_p(\boldsymbol{x}, \boldsymbol{y}) \leqslant 2d_H(\boldsymbol{x}, \boldsymbol{y}).$$

For a code $\mathcal{C}$, we denote its minimum Hamming distance by $d_H(\mathcal{C})$. The *symbol-pair code* of $\mathcal{C}$ is the code

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

Then, similarly, the *minimum pair distance* of $\mathcal{C}$, $d_p(\mathcal{C})$, is the minimum Hamming distance of $\pi(\mathcal{C})$, i.e.,

$$d_p(\mathcal{C}) = d_H(\pi(\mathcal{C})).$$

From Proposition 2, the following connection between $d_H(\mathcal{C})$ and $d_p(\mathcal{C})$ is established [1]

$$d_H(\mathcal{C}) + 1 \leqslant d_p(\mathcal{C}) \leqslant 2d_H(\mathcal{C}).$$

The goal in constructing codes for the pair-read channel is to achieve high minimum pair distance with respect to the minimum Hamming distance. It was shown in [1] that interleaving two codes with minimum Hamming distance d_H generates a code with the same minimum Hamming distance d_H but with minimum pair distance is $2d_H$. Even though this construction generates codes with the largest possible minimum pair distance with respect to their minimum Hamming distance, it is less attractive as, in general, the interleaving method suffers from a poor Hamming distance relative to its resulting codeword length.

Yet another interesting family of codes that was analyzed in [1] are the linear cyclic codes. It was proved that for a linear cyclic code $\mathcal{C}$ with minimum Hamming distance d_H, its minimum pair distance is at least $d_H + 2$. Using the Hartmann-Tzeng bound, this lower bound was improved to $d_H + 3$, when the code length is a prime number. Our main goal in the next section is to show an improved lower bound on the minimum pair distance of linear cyclic codes.

III. The Pair Distance of Cyclic Codes

The goal of this section is to show that linear cyclic codes have high minimum pair distance. In order to do so, we first give a method to calculate the pair weight of $\boldsymbol{x}$. In fact, a similar property was proved in [1] (Theorem 2) but using a different notation.

The key idea to notice is that if $x_i = 1$ then there are two non-zero symbols in $\pi(\boldsymbol{x})$, the i-th and $(i-1)$-st symbol. However if $x_{i-1} = 1$, then the $(i-1)$-st symbol is already non-zero as a result of $x_{i-1} = 1$. Hence, if $(x_{i-1}, x_i) = (0, 1)$ we have two non-zero symbols in $\pi(\boldsymbol{x})$ as a result of x_i and if $(x_{i-1}, x_i) = (1, 1)$ we have only a single non-zero symbol in $\pi(\boldsymbol{x})$. Therefore, in order to determine the weight of $\pi(\boldsymbol{x})$, one needs to count the number of $(0, 1)$ sequences in the vector $\boldsymbol{x}$, which we now show how to calculate.

For $\boldsymbol{x} = (x_0, x_1, \dots, x_{n-1})$, we define

$$\boldsymbol{x}' = (x_0 + x_1, x_1 + x_2, \dots, x_{n-1} + x_0).$$

The next lemma shows how to calculate the pair weight of a vector $\boldsymbol{x}$.

Lemma 3. *For any $\boldsymbol{x} \in \Sigma^n$, $w_p(\boldsymbol{x}) = w_H(\boldsymbol{x}) + w_H(\boldsymbol{x}')/2$.*

Proof: Let

$$S_0 = \{i \ : \ (x_i, x_{i+1}) \neq 0 \ \text{ and } \ x_i = 1\},$$
$$S_1 = \{i \ : \ (x_i, x_{i+1}) \neq 0 \ \text{ and } \ x_i = 0, x_{i+1} = 1\}.$$

Hence, $|S_0| = w_H(\boldsymbol{x})$, $S_0 \cap S_1 = \emptyset$, and $w_p(\boldsymbol{x}) = |S_0| + |S_1|$. For all $0 \leqslant i \leqslant n-1$, $i \in S_1$ if and only if $x_{i+1} = 1$ and $x_i = 0$. Thus, $x_i + x_{i+1} = 1$ or $x'_i = 1$, where $x_i = 0$. Hence, we get

$$|S_1| = |\{i \ : \ x_{i+1} = 1 \ \text{ and } x_i = 0\}| .$$

Note that for any $\boldsymbol{x} \in \Sigma^n$,

$$|\{i : x_{i+1} = 1 \text{ and } x_i = 0\}| = |\{i : x_{i+1} = 0 \text{ and } x_i = 1\}| ,$$

and the sum of the cardinality of the two sets is $w_H(\boldsymbol{x}')$. Hence, $|S_1| = \frac{w_H(\boldsymbol{x}')}{2}$ and

$$w_p(\boldsymbol{x}) = |S_0| + |S_1| = w_H(\boldsymbol{x}) + \frac{w_H(\boldsymbol{x}')}{2}.$$

■

Using the property we proved in Lemma 3, we are now ready to show an improved lower bound on the minimum pair distance of linear cyclic codes.

Theorem 4. *Let $\mathcal{C}$ be a linear and cyclic code of dimension greater than one. Then,*

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

Proof: Let $\boldsymbol{x} = (x_0, \dots, x_{n-1})$ be a codeword in $\mathcal{C}$. Assume that $\boldsymbol{x}$ is not the all-ones vector. Since the code is cyclic then $(x_1, \dots, x_{n-1}, x_0) \in \mathcal{C}$ and thus $\boldsymbol{x}' \in \mathcal{C}$. The weight of $\boldsymbol{x}'$ is even and hence $w_H(\boldsymbol{x}') \geqslant 2\lceil d_H(\mathcal{C})/2 \rceil$. Furthermore, $w_H(\boldsymbol{x}) \geqslant d_H(\mathcal{C})$. Together, these facts imply that

$$w_p(\boldsymbol{x}) = w_H(\boldsymbol{x}) + w_H(\boldsymbol{x}')/2 \geqslant d_H(\mathcal{C}) + \left\lceil \frac{d_H(\mathcal{C})}{2} \right\rceil .$$

We conclude by noting that if $\boldsymbol{x} = \mathbf{1}$, the all-ones vector, then the inequality above is easily verified. This completes the proof. ■

Theorem 4 shows that linear cyclic codes are attractive for symbol-pair read channels as their minimum pair distance is large, allowing the correction of a large number of symbol-pair errors. An interesting problem which thus arises is to construct efficient decoders for these codes.

IV. Decoding

After finding codes with large minimum pair distance we now show an efficient decoder for such codes. Given a linear cyclic code $\mathcal{C}$, with minimum distance $d_H(\mathcal{C}) = 2t+1$, we assume it has a decoder $\mathcal{D}_{\mathcal{C}}$ that can correct up to t errors. We will show how to use this decoder in order to construct a decoder for the code $\pi(\mathcal{C})$ which corrects up to $t_0 = \lfloor \frac{3t+1}{2} \rfloor$ symbol-pair errors.

We define the decoder $\mathcal{D}_{\mathcal{C}}$ as a map $\mathcal{D}_{\mathcal{C}} : \Sigma^n \to \mathcal{C} \cup \{F\}$ and the notation $\mathcal{D}_{\mathcal{C}}(\boldsymbol{y}) = \widehat{\boldsymbol{c}}$ indicates that the decoder's input is a received word $\boldsymbol{y}$ and its output is a decoded codeword $\widehat{\boldsymbol{c}}$ or the decoder failure symbol F. If $\boldsymbol{c} \in \mathcal{C}$ is the transmitted word and $d_H(\boldsymbol{c}, \boldsymbol{y}) \leqslant t$, then it is guaranteed that $\widehat{\boldsymbol{c}} = \boldsymbol{c}$. However, if $d_H(\boldsymbol{c}, \boldsymbol{y}) > t$, then either $\widehat{\boldsymbol{c}} = F$, indicating that more than t errors have occurred, or $\widehat{\boldsymbol{c}}$ is a codeword different from $\boldsymbol{c}$, whose Hamming distance from the received word $\boldsymbol{y}$ is at most t, i.e., $d_H(\widehat{\boldsymbol{c}}, \boldsymbol{y}) \leqslant t$.

Let us introduce another code that will serve us in this decoder construction. The *double-repetition code* of $\mathcal{C}$ is the code

$$\mathcal{C}_2 = \{(\boldsymbol{c}, \boldsymbol{c}) \ : \ \boldsymbol{c} \in \mathcal{C}\}.$$

Note that its length is $2n$ and its minimum Hamming distance satisfies $d_H(\mathcal{C}_2) = 2d_H(\mathcal{C})$. The code $\mathcal{C}_2$ can correct up to $2t$ errors and we assume that it has a decoder $\mathcal{D}_{\mathcal{C}_2} : \Sigma^n \times \Sigma^n \to \Sigma^n \cup \{F\}$ having the same properties as the decoder $\mathcal{D}_{\mathcal{C}}$. Every codeword in $\mathcal{C}_2$ consists of two identical codewords from $\mathcal{C}$ and thus, for simplicity of notation, we assume that the decoder $\mathcal{D}_{\mathcal{C}_2}$ returns only one copy of the decoded codeword from $\mathcal{C}$. We will address at the end of the section the problem of constructing the decoder $\mathcal{D}_{\mathcal{C}_2}$.

Let $\boldsymbol{c} \in \mathcal{C}$ and let $\pi(\boldsymbol{c}) \in \pi(\mathcal{C})$ be its symbol-pair vector. Let $\boldsymbol{y} = \pi(\boldsymbol{c}) + \boldsymbol{e}$ be a received word, where $\boldsymbol{e} \in (\Sigma \times \Sigma)^n$ is the error vector and $w_H(\boldsymbol{e}) \leqslant \lfloor \frac{3t+1}{2} \rfloor = t_0$. We will show a decoder $\mathcal{D}_\pi : (\Sigma \times \Sigma)^n \to \{0,1\}^n$ which receives the word $\boldsymbol{y}$ and returns $\widehat{\boldsymbol{c}}$.

We denote the received vector by

$$\boldsymbol{y} = \left((y_{0,0}, y_{0,1}), (y_{1,0}, y_{1,1}), \dots, (y_{n-1,0}, y_{n-1,1})\right)$$

and define the following three vectors

$$\begin{aligned} \boldsymbol{y}_L &= (y_{0,0}, \dots, y_{n-1,0}), \\ \boldsymbol{y}_R &= (y_{0,1}, \dots, y_{n-1,1}), \\ \boldsymbol{y}_S &= (y_{0,0} + y_{0,1}, \dots, y_{n-1,0} + y_{n-1,1}). \end{aligned}$$

Since the vector $\boldsymbol{y}$ suffers at most t_0 pair-symbol errors, the vectors $\boldsymbol{y}_L$ and $\boldsymbol{y}_R$ each have at most t_0 errors as well. Note that the vector $\boldsymbol{y}_S$ has at most t_0 errors with respect to the codeword $\boldsymbol{c}' = (c_0 + c_1, \dots, c_{n-1} + c_0)$. In general, the knowledge of the codeword $\boldsymbol{c}'$ does not uniquely determine the value of $\boldsymbol{c}$. However, in this scenario it does. This observation, which we will use of in the decoder algorithm, is proved in the following lemma.

Lemma 5. *If the codeword $\boldsymbol{c}' \in \mathcal{C}$ is successfully decoded then we can recover the codeword $\boldsymbol{c}$.*

Proof: The codeword $\boldsymbol{c}$ satisfies $c_i = c_0 + \sum_{j=0}^{i-1} c'_j$. Hence if we define $\widetilde{\boldsymbol{c}} = [\widetilde{c}_0, \dots, \widetilde{c}_{n-1}]$ by $\widetilde{c}_i = \sum_{j=0}^{i-1} c'_j$ then the codeword $\boldsymbol{c}$ is either $\widetilde{\boldsymbol{c}}$ or $\widetilde{\boldsymbol{c}} + \mathbf{1}$, depending on the value of c_0. The distance between $\boldsymbol{y}_L$ and $\boldsymbol{c}$ is at most t_0 and $d_H(\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{c}} + \mathbf{1}) = n$. Hence, if $d_H(\boldsymbol{y}_L, \widetilde{\boldsymbol{c}}) < d_H(\boldsymbol{y}_L, \widetilde{\boldsymbol{c}} + \mathbf{1})$ then $\boldsymbol{c} = \widetilde{\boldsymbol{c}}$ and otherwise $\boldsymbol{c} = \widetilde{\boldsymbol{c}} + \mathbf{1}$. In any case, we can recover the codeword $\boldsymbol{c}$. ■

According to Lemma 5, it is possible to recover the codeword $\boldsymbol{c}$ from the codeword $\boldsymbol{c}'$. By abuse of notation, we denote by $\boldsymbol{c}'^*$ an operator that calculates, as explained in Lemma 5, the codeword $\boldsymbol{c}$ from $\boldsymbol{c}'$, and so $\boldsymbol{c}'^* = \boldsymbol{c}$.

The number of symbol-pair errors in the vector $\boldsymbol{y}$ is at most t_0. Each symbol-pair error corresponds to one or two bit error in the symbol-pair. We let E_1 be the number of single-bit pair-symbol errors and E_2 be the number of double-bit pair-symbol errors, where $E_1 + E_2 \leqslant t_0$. Thus, the number of errors in $\boldsymbol{y}_S$ is E_1 and the number of errors in $(\boldsymbol{y}_L, \boldsymbol{y}_R)$ is $E_1 + 2E_2$. Another property which we will use in the decoder construction is proved in the next lemma.

Lemma 6. *If $\boldsymbol{c} \in \mathcal{C}$, $\boldsymbol{y} = \pi(\boldsymbol{c}) + \boldsymbol{e}$, and $w_H(\boldsymbol{e}) \leqslant t_0$, then either $\mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S) = \boldsymbol{c}'$ or $\mathcal{D}_{\mathcal{C}_2}((\boldsymbol{y}_L, \boldsymbol{y}_R)) = \boldsymbol{c}$.*

Proof: If $E_1 \leqslant t$ then the decoder $\mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S)$ is successful. Otherwise, $E_1 \geqslant t+1$ and $E_2 \leqslant t_0 - (t+1)$, so the number of errors in $(\boldsymbol{y}_L, \boldsymbol{y}_R)$ satisfies

$$E_1 + 2E_2 \leqslant t_0 + t_0 - (t+1) = 2\left\lfloor \frac{3t+1}{2} \right\rfloor - (t+1) \leqslant 2t,$$

and therefore the decoder $\mathcal{D}_{\mathcal{C}_2}((\boldsymbol{y}_L, \boldsymbol{y}_R))$ is successful. ■

According to the last lemma we know that at least one of the two decoders succeeds. However, we cannot determine easily which one of them does and the main task of the decoder construction for $\pi(\mathcal{C})$ is to find the successful decoder. The decoder's output $\mathcal{D}_{\pi}(\boldsymbol{y}) = \widehat{\boldsymbol{c}}$ is calculated as follows:

Step 1. $\boldsymbol{c}_1 = \mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S)$, $e_1 = d_H(\boldsymbol{c}_1, \boldsymbol{y}_S)$.
Step 2. $\boldsymbol{c}_2 = \mathcal{D}_{\mathcal{C}_2}((\boldsymbol{y}_L, \boldsymbol{y}_R))$, $e_2 = d_H((\boldsymbol{c}_2, \boldsymbol{c}_2), (\boldsymbol{y}_L, \boldsymbol{y}_R))$.
Step 3. If $\boldsymbol{c}_1 = F$ or $w_H(\boldsymbol{c}_1)$ is odd then $\widehat{\boldsymbol{c}} = \boldsymbol{c}_2$.
Step 4. If $e_1 \leqslant \lfloor \frac{t+2}{2} \rfloor$, then $\widehat{\boldsymbol{c}} = \boldsymbol{c}_1^*$.
Step 5. If $e_1 > \lfloor \frac{t+2}{2} \rfloor$, let $e_1 = \lfloor \frac{t+2}{2} \rfloor + a$, $(1 \leqslant a \leqslant \lceil \frac{t}{2} \rceil - 1)$
a) If $e_2 \leqslant t_0 + a$ then $\widehat{\boldsymbol{c}} = \boldsymbol{c}_2$,
b) Otherwise, $\widehat{\boldsymbol{c}} = \boldsymbol{c}_1^*$.

The correctness of the decoder is proved in the next theorem.

Theorem 7. *The decoder output satisfies $\mathcal{D}_{\pi}(\boldsymbol{y}) = \widehat{\boldsymbol{c}} = \boldsymbol{c}$.*

Proof: According to Lemma 6, at least one of the two decoders in Steps 1 and 2 succeeds. Steps 3–5 help to determine which of the two decoders succeeds.

Step 3: Since $\boldsymbol{y}_S$ is a noisy version of the codeword $\boldsymbol{c}'$, in the decoding operation on the first step we try to decode the codeword $\boldsymbol{c}'$. Remember that the weight of $\boldsymbol{c}'$ is even. Hence, if $\boldsymbol{c}_1 = F$ or the Hamming weight of $\boldsymbol{c}_1$ is odd, then this decoding operation fails and thus the decoder in Step 2 succeeds. If we reach Steps 4 and 5 then $w_H(\boldsymbol{c}_1)$ is even.

Step 4: Here we show that if $e_1 \leqslant \lfloor \frac{t+2}{2} \rfloor$, then $E_1 \leqslant \lfloor \frac{t+2}{2} \rfloor$ as well and the decoder in step 1 succeeds. Assume to the contrary that there is a miscorrection in Step 1. Then the word $\boldsymbol{y}_S$ is miscorrected to some codeword of even weight. The weight of the error vector found in Step 1, e_1, is at most $\lfloor \frac{t+2}{2} \rfloor$. Since the minimum distance of the code $\mathcal{C}$ is $2t+1$, the number E_1 of actual errors in $\boldsymbol{y}_S$ satisfies

$$E_1 \geqslant 2t + 2 - \left\lfloor \frac{t+2}{2} \right\rfloor = \left\lceil \frac{3t}{2} \right\rceil + 1 > t_0,$$

which is a contradiction as the number of errors in $\boldsymbol{y}_S$ is at most t_0. Therefore, in this case the decoding operation $\boldsymbol{c}_1 = \mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S) = \boldsymbol{c}'$ succeeds, and according to Lemma 5, we get

$$\widehat{\boldsymbol{c}} = \boldsymbol{c}_1^* = \boldsymbol{c}'^* = \boldsymbol{c}.$$

Step 5: We are left with the case where $e_1 > \lfloor \frac{t+2}{2} \rfloor$. Since $e_1 \leqslant t$, let $e_1 = \lfloor \frac{t+2}{2} \rfloor + a$, where $1 \leqslant a \leqslant \lceil \frac{t}{2} \rceil - 1$.

Assume the decoding in Step 2 fails. According to Lemma 6, the decoding operation $\boldsymbol{c}_1 = \mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S)$ succeeds,

$$E_1 = e_1 = \left\lfloor \frac{t+2}{2} \right\rfloor + a.$$

The value of E_2 satisfies

$$E_2 \leqslant t_0 - \left(\left\lfloor \frac{t+2}{2} \right\rfloor + a \right) = \left\lfloor \frac{3t+1}{2} \right\rfloor - \left\lfloor \frac{t+2}{2} \right\rfloor - a \leqslant t - a.$$

The total number of errors in $(\boldsymbol{y}_L, \boldsymbol{y}_R)$ is

$$E_1 + 2E_2 \leqslant t_0 + t - a = \left\lfloor \frac{5t+1}{2} \right\rfloor - a.$$

Since the decoder $\mathcal{D}_{\mathcal{C}_2}((\boldsymbol{y}_L, \boldsymbol{y}_R))$ fails and the minimum distance of $\mathcal{C}_2$ is $4t+2$, we get that the weight of the error vector in Step 2, e_2, would have to satisfy

$$e_2 \geqslant 4t + 2 - (E_1 + 2E_2) \geqslant 4t + 2 - \left(\left\lfloor \frac{5t+1}{2} \right\rfloor - a \right)$$
$$\geqslant \left\lfloor \frac{3t+1}{2} \right\rfloor + a + 1 = t_0 + a + 1.$$

Hence, we conclude that if $e_2 \leqslant t_0 + a$ then necessarily the decoder in Step 2 succeeds.

Assume the decoding in Step 1 fails. As in Step 4, since $\mathcal{D}_{\mathcal{C}}(\boldsymbol{y}_S)$ fails, the number of errors E_1 in $\boldsymbol{y}_S$ is at least

$$E_1 \geqslant 2t + 2 - \left(\left\lfloor \frac{t+2}{2} \right\rfloor + a \right) = \left\lceil \frac{3t}{2} \right\rceil - (a-1) = t_0 - (a-1).$$

Since $E_1 + E_2 \leqslant t_0$, the value of E_2 satisfies $0 \leqslant E_2 \leqslant a - 1$, and $E_1 + 2E_2$, the total number of errors in $(\boldsymbol{y}_L, \boldsymbol{y}_R)$, satisfies

$$t_0 - (a-1) \leqslant E_1 + 2E_2 = (E_1 + E_2) + E_2 \leqslant t_0 + a - 1.$$

Thus, the decoding operation $\mathcal{D}_{\mathcal{C}_2}((\boldsymbol{y}_L, \boldsymbol{y}_R))$ succeeds, and

$$t_0 - (a-1) \leqslant e_2 \leqslant t_0 + a - 1.$$

Hence, if $e_2 > t_0 + a$ then the decoder in Step 1 succeeds.

That explains a) and b) of Step 5. ■

To complete this section, let us go back to the construction of the decoder $\mathcal{D}_{\mathcal{C}_2}$. This decoder receives two vectors $\boldsymbol{y}_1 = (y_{1,0}, \ldots, y_{1,n-1}), \boldsymbol{y}_2 = (y_{2,0}, \ldots, y_{2,n-1})$. Each is a noisy version of some codeword $\boldsymbol{c} \in \mathcal{C}$, and the goal is to correct a total of $2t$ errors in the two vectors. We define the vector $\widehat{\boldsymbol{y}} = (\widehat{y}_0, \ldots, \widehat{y}_{n-1})$ such that for all $0 \leqslant i \leqslant i - 1$, $\widehat{y}_i = y_{1,i}$ if $y_{1,i} = y_{2,i}$, and otherwise $\widehat{y}_i = ?$ to indicate an erasure. If the number of errors in $\widehat{\boldsymbol{y}}$ is τ and the number of erasures is ρ, then we have $2\tau + \rho \leqslant 2t = d(\mathcal{C}) - 1$, which is within the error and erasure correcting capability of $\mathcal{C}$. We are left only with the problem of defining a decoder that corrects errors and erasures for cyclic codes. For that, we refer to [9], [10].

we extend some of our results on the symbol-pair read channel to the case where more than two symbols are sensed on each read

V. Extensions for Arbitrary b

In this section, we extend some of our results on the symbol-pair read channel to the case where more than two symbols are sensed on each read. For $b \geqslant 3$, the *b-symbol read vector* of $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1}) \in \Sigma^n$ is defined to be

$$\pi_b(\boldsymbol{x}) = [(x_0, \ldots, x_{b-1}), \ldots, (x_{n-1}, x_0, \ldots, x_{b-2})] \in (\Sigma^b)^n.$$

The *b-distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $d_b(\boldsymbol{x}, \boldsymbol{y})$ and is defined to be $d_b(\boldsymbol{x}, \boldsymbol{y}) = d_H(\pi_b(\boldsymbol{x}), \pi_b(\boldsymbol{y}))$. In analogy to Proposition 2, it is possible to show that

$$d_H(\boldsymbol{x}, \boldsymbol{y}) + b - 1 \leqslant d_b(\boldsymbol{x}, \boldsymbol{y}) \leqslant b \cdot d_H(\boldsymbol{x}, \boldsymbol{y}),$$

and $d_b(\boldsymbol{x}, \boldsymbol{y}) = w_b(\boldsymbol{x} + \boldsymbol{y})$.

Our main goal here is to generalize Lemma 3 for arbitrary $b \geqslant 3$. For a vector $\boldsymbol{x}$, let us define the vector $\widehat{\boldsymbol{x}}$ by inverting every sequence of less than $b - 1$ zeros in $\boldsymbol{x}$. More formally, if $(x_i, x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}) = (1, 0, \ldots, 0, 1)$ for some $0 \leqslant i \leqslant n - 1$ and $k \leqslant b - 2$, then $\widehat{x}_j = 1$ for $i + 1 \leqslant j \leqslant i + k$. For all other values of j, $\widehat{x}_j = x_j$.

Lemma 8. *For any $\boldsymbol{x} \in \Sigma^n$ and positive integer $b \geqslant 3$*

$$w_b(\boldsymbol{x}) = w_H(\widehat{\boldsymbol{x}}) + (b-1) \cdot \frac{w_H(\widehat{\boldsymbol{x}}')}{2}.$$

Proof: Let us first show that $w_b(\boldsymbol{x}) = w_b(\widehat{\boldsymbol{x}})$. Consider a sequence of $k \leqslant b - 2$ consecutive zeros,

$$(x_i, x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}) = (1, 0, \ldots, 0, 1).$$

The zero bits $x_{i+1}, \ldots, x_k$ appear in the j-th symbol of $\pi_b(\boldsymbol{x})$ for $i - b + 2 \leqslant j \leqslant i + k$. Since $x_i = x_{i+k+1} = 1$, in this range of values of j, $\pi_b(\boldsymbol{x})_j \neq 0$ and $\pi_b(\widehat{\boldsymbol{x}})_j \neq 0$. For all other values of j, the corresponding bits of x_i and $\widehat{x}_i$ are the same and thus $\pi_b(\boldsymbol{x})_j \neq 0$ if and only if $\pi_b(\widehat{\boldsymbol{x}})_j \neq 0$.

Next, we find the value of $w_b(\widehat{x})$, making use of the fact that any run of zeros in $\widehat{x}$ is of length at least $b-1$. Let

$$S_0 = \{i : (\widehat{x}_i, \ldots, \widehat{x}_{i+b-1}) \neq 0, \widehat{x}_i = 1\},$$
$$S_1 = \{i : (\widehat{x}_i, \ldots, \widehat{x}_{i+b-1}) \neq 0, \widehat{x}_i = 0, \widehat{x}_{i+1} = 1\},$$
$$\vdots$$
$$S_{b-2} = \{i : (\widehat{x}_i, \ldots, \widehat{x}_{i+b-1}) \neq 0, \widehat{x}_i = \cdots = \widehat{x}_{i+b-3} = 0, \widehat{x}_{i+b-2} = 1\},$$
$$S_{b-1} = \{i : (\widehat{x}_i, \ldots, \widehat{x}_{i+b-1}) \neq 0, \widehat{x}_i = \cdots = \widehat{x}_{i+b-2} = 0, \widehat{x}_{i+b-1} = 1\}.$$

Therefore, $w_H(\widehat{x}) = |S_0|$; $S_j \cap S_\ell = \emptyset$, for all $0 \leqslant j < \ell \leqslant b-1$; and $w_b(\widehat{x}) = |\cup_{i=0}^{b-1} S_i| = \sum_{i=0}^{b-1} |S_i|$.

Let us show that for all $2 \leqslant \ell \leqslant b-1$, $|S_1| = |S_\ell|$. If $i \in S_1$ then $(\widehat{x}_i, \widehat{x}_{i+1}) = (0,1)$. Since there is no sequence of less than $b-2$ consecutive zeros

$$(\widehat{x}_{i-(b-2)}, \ldots, \widehat{x}_i, \widehat{x}_{i+1}) = (0, \ldots, 0, 1)$$

and thus $i - (\ell - 1) \in S_\ell$. Hence, $|S_\ell| \geqslant |S_1|$. For the opposite inequality, note that if $i \in S_\ell$, then

$$(\widehat{x}_i, \widehat{x}_{i+1}, \ldots, \widehat{x}_{i+\ell-1}, \widehat{x}_{i+\ell}) = (0, \ldots, 0, 1).$$

Therefore $(\widehat{x}_{i+\ell-1}, \widehat{x}_{i+\ell}) = (0,1)$, so $i + \ell - 1 \in S_1$ and we get $|S_1| \geqslant |S_\ell|$. Hence, $|S_1| = |S_\ell|$ for all $2 \leqslant \ell \leqslant b-1$. As in the proof of Lemma 3, $|S_1| = \frac{w_H(\widehat{x}')}{2}$, and finally we get

$$w_b(\widehat{x}) = \sum_{\ell=0}^{b-1} |S_i| = w_H(\widehat{x}) + (b-1) \cdot \frac{w_H(\widehat{x}')}{2}.$$ ■

We would next like to construct codes with large b-distance. As in the case of symbol-pair codes, we define the *b-symbol read code* of $\mathcal{C}$ as the code $\pi_b(\mathcal{C}) = \{\pi_b(\boldsymbol{c}) \ : \boldsymbol{c} \in \mathcal{C}\}$, and the *minimum b-distance* of $\mathcal{C}$, $d_b(\mathcal{C})$, as $d_b(\mathcal{C}) = d_H(\pi_b(\mathcal{C}))$.

The interleaving scheme given in [1] constructs codes $\mathcal{C}$ that satisfy $d_p(\mathcal{C}) = 2d_H(\mathcal{C})$. This construction can be extended for arbitrary b. It can be shown that the interleaving of b codes, all with minimum distance d_H, generates a code $\mathcal{C}$ with minimum Hamming distance $d_H(\mathcal{C}) = d_H$ and minimum b-distance $d_b(\mathcal{C}) = b \cdot d_H(\mathcal{C})$.

A decoder for the interleaving construction works very similarly to the one given for the interleaving scheme for pair-symbols in [1]. The *majority decoder* is a decoder which outputs for every bit its majority value among its b received values, or ? in case of equality between the number of zeros and ones. Then, it is possible to decode every interleaved codeword independently.

The following two lemmas show properties of some special codes. In the first lemma, the code is Σ^n, so its minimum distance is one. The second lemma analyzes the Hamming code.

Lemma 9. *If $\mathcal{C} = \Sigma^n$, then for $b \geqslant 3$ the minimum b-distance satisfies $d_b(\mathcal{C}) = b$ and it is possible to correct $\lfloor \frac{b-1}{2} \rfloor$ symbol errors by the majority decoder.*

Proof: Assume $\boldsymbol{x}$ is a non-zero word which is not the all-ones vector. Then, we have $\widehat{\boldsymbol{x}} \neq \mathbf{0}$ and thus $w_H(\widehat{\boldsymbol{x}}) \geqslant 1$ and $w_H(\widehat{\boldsymbol{x}}') \geqslant 2$. According to Lemma 8, we get $w_b(\boldsymbol{x}) \geqslant 1 + (b-1) \cdot \frac{2}{2} = b$. In case $\boldsymbol{x} = \mathbf{1}$, then $w_b(\boldsymbol{x}) = n$ and the inequality above holds as well.

If there are $\lfloor \frac{b-1}{2} \rfloor$ symbol errors, then every bit of the vector $\boldsymbol{x}$ is in error in $\pi_b(\boldsymbol{x})$ at most $\lfloor \frac{b-1}{2} \rfloor$ times. Thus, the majority decoder succeeds. ■

Lemma 10. *If $\mathcal{C}$ is the cyclic Hamming code of length $n = 2^m - 1$ then $d_b(\mathcal{C}) = 2b + 1$, where $b + 2 \leqslant m$.*

Proof: Let $\boldsymbol{x} \in \mathcal{C}$ be a non-zero codeword and assume that $\widehat{\boldsymbol{x}} \neq \mathbf{1}$. Hence, $w_H(\widehat{\boldsymbol{x}}) \geqslant 3$. If $w_H(\widehat{\boldsymbol{x}}') \geqslant 4$, then according to Lemma 8, we get $w_b(\boldsymbol{x}) \geqslant 3 + (b-1) \cdot \frac{4}{2} = 2b + 1$.

Now assume that $w_H(\widehat{\boldsymbol{x}}') = 2$, so $\widehat{\boldsymbol{x}}$ has a single continuous run of ones, and assume it has length ℓ. We notice that $\ell \geqslant m$. Otherwise, the non-zero entries of the codeword $\boldsymbol{x}$ are confined to at most $m-1$ locations. If $g(x)$ is a generator polynomial of degree m for the cyclic Hamming code, then there exists a non-zero polynomial of degree at most $m-1$ which is a multiple of $g(x)$. Thus, we get a contradiction. Therefore, in this case, we get $w_b(\widehat{\boldsymbol{x}}) \geqslant m + (b-1) \cdot \frac{2}{2} = m + b - 1 \geqslant 2b + 1$. To conclude this part of the proof, note that if $\widehat{\boldsymbol{x}} = \mathbf{1}$, then $w_b(\widehat{\boldsymbol{x}}) = n \geqslant 2b + 1$.

To show that the minimum distance is $2b + 1$, we see that if we take a codeword of weight three with two consecutive ones, then its b-weight is exactly $2b + 1$. ■

VI. Conclusion

In this paper, we studied the symbol-pair read channel. After reviewing the channel model and basic properties, we then showed that linear cyclic codes are very effective in correcting symbol-pair errors. The main part of the paper was devoted to the construction of an effective decoding algorithm for such codes. Finally, we extended the model and some of the results to the b-symbol read channel, where $b > 2$.

VII. Acknowledgement

The authors thank to Yuval Cassuto for helpful discussions on the symbol-pair read channels. This research was supported in part by the ISEF Foundation, the Lester Deutsch Fellowship, the University of California Lab Fees Research Program, Award No. 09-LR-06-118620-SIEP, the National Science Foundation under Grant CCF-1116739, the Center for Magnetic Recording Research at UCSD, and the NSF Expeditions in Computing Program under grant CCF-0832824.

References

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

[2] Y. Cassuto and S. Litsyn, "Symbol-pair codes: algebraic constructions and asymptotic bounds," in *Proc. IEEE International Symposium on Information Theory*, pp. 2348–2352, St. Petersburg, Russia, Aug. 2011.

[3] E. Konstantinova, "Reconstruction of signed permutations from their distorted patterns," in *Proc. IEEE International Symposium on Information Theory*, pp. 474–477, Adelaide, Australia, September 2005.

[4] E. Konstantinova, V.I. Levenshtein, and J. Siemons, "Reconstruction of permutations distorted by single transposition errors," arXiv:math/0702191v1, February 2007.

[5] V.I. Levenshtein, "Reconstructing objects from a minimal number of distorted patterns", (in Russian), *Dokl. Acad. Nauk 354*, pp. 593-596; English translation, *Doklady Mathematics* vol. 55 pp. 417–420, 1997.

[6] V.I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. on Information Theory*, vol. 47, no. 1, pp. 2–22, January 2001.

[7] V.I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences", *J. of Combin. Theory, Ser. A*, vol. 93, no. 2, pp. 310–332, 2001.

[8] V.I. Levenshtein and J. Siemons, "Error graphs and the reconstruction of elements in groups," *J. of Combin. Theory, Ser. A*, vol. 116, pp. 795–815, 2009.

[9] E. Orsini and M. Sala, "Correcting errors and erasures via the syndrome variety," *J. of Pure and Applied Algebra*, vol. 200, pp. 191–226, 2005.

[10] H. Shahri and K.K. Tzeng, "On error-and-erasure decoding of cyclic codes," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 489–496, March 1992.

[11] E. Yaakobi and J. Bruck, "On associative memories and the sequences reconstructions problem," *IEEE International Symposium on Information Theory*, Cambridge, MA, July 2012.