List Decodability of Symbol-Pair Codes

Shu Liu*, Chaoping Xing† and Chen Yuan‡

Abstract

We investigate the list decodability of symbol-pair codes in the present paper. Firstly, we show that list decodability of every symbol-pair code does not exceed the Gilbert-Varshamov bound. On the other hand, we are able to prove that with high probability, a random symbol-pair code can be list decoded up to the Gilbert-Varshamov bound. Our second result of this paper is to derive the Johnson-type bound, i.e., a lower bound on list decoding radius in terms of minimum distance. Finally, we present a list decoding algorithm of Reed-Solomon codes beyond the Johnson-type bound.

1 Introduction

The high-density data storage technologies aim at designing the high-capacity storages at a relatively low cost. To achieve this goal, the theory of symbol-pair coding [2] was proposed to handle channels that output pairs of overlapping symbols, rather than one symbol at a time. Such channels, so called *symbol-pair read channels*, introduce a new metric called *pair distance*. It was showed that the pair error correcting capability of a code is larger than the error correcting capability of the same code in the Hamming metric. Cassuo and Litsyn [3] gave an asymptotic lower bound on coding rates. This lower bound also indicates the existence of symbol-pair codes with higher rate than the codes in Hamming distance provided that both codes have the same relative distance. Chee et al. [4] established a Singleton-type bound and showed the existence of symbol-pair codes meeting this bound. Following this direction, several works contributed to the constructions of symbol-pair codes meeting this bound [5] and [13].

In this paper, we focus on the list decoding of symbol-pair codes. This concept of list decoding was first introduced by Elias [8] and Wozencraft [18]. Unlike the unique decoding algorithm, list decoding algorithm outputs a list of candidate codewords so as to tolerate and correct more errors. One of the key issues in coding theory is to explicitly construct codes with large list decoding radius. Since there are too many works concerned with this topic, we refer the reader to [9] for details. Inspired by the list decoding in Hamming metric, we establish the lower bound and upper bound on the list decoding radius of symbol-pair codes. We also reveal the differences between the codes in Hamming metric and symbol-pair metric by observing the different behaviours of the list decoding of Reed-Solomon codes in both metrics.

*Shu Liu was with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (email: shuliu@uestc.edu.cn)

†Chaoping Xing was with Division of Mathematical Sciences, School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore 637371 (email: xingcp@ntu.edu.sg)

‡Chen Yuan was with Centrum Wiskunde & Informatica, Amsterdam, Netherlands (email: Chen.Yuan@cwi.nl) Part of this work was done when the author was with the School of Physical and Mathematical Science, Nanyang Technological University, Singapore.

Previous results

There are many works dedicated to unique decoding of symbol-pair codes. Cassuto and Blaum [2] presented their decoding algorithm based on the error decoding algorithm in the Hamming metric. Yaakobi, Bruck and Siegel gave two constructions of effective decoding algorithms for linear cyclic codes [20] and [19]. The decoding algorithm utilizing the syndrome of symbol-pair codes was proposed in [15] by Hirotomo, Takita and Morii. They [14] subsequently give an error-trapping decoding algorithm that is required to impose some restrictions on the pair error patterns. There is a decoding algorithm based on linear programming designed for binary linear symbol-pair codes in [16] by Horii, Matsushima and Hirasawa.

Our results

To the best of our knowledge, all known decoding algorithms are designed for the unique decoding of symbol-pair codes. In this paper, we investigate the list decoding of symbol-pair codes. We first establish the Gilbert-Varshamov bound as an upper bound on the list decoding radius for all the symbol-pair codes. On the other hand, we also show that most random symbol-pair codes can be list decoded up to this bound. Then, we derive the Johnson-type bound in terms of minimum distance which indicates that any symbol-pair codes can be list decoded beyond this bound. To show tightness of this bound, we further construct symbol-pair codes that can not be list decoded slightly beyond this bound, while it is an open problem whether there exists any Reed-Solomon code list decodable beyond the Johnson-type bound in Hamming metric. Finally, we give an explicit list decoding algorithm for a family of Reed-Solomon codes beyond this Johnson-type bound.

Organization

This paper is organized as follows. In Section 2, we introduce definitions of symbol-pair codes, the Gilbert-Varshamov bound and some preliminaries on list decoding. In Section 3, we establish an upper bound on the list decoding radius of symbol-pair codes, i.e., the Gilbert-Varshamov bound. In addition, in Section 3 we also show that, with high probability, a random code can be list decoded up to the Gilbert-Varshamov bound. The Johnson-type bound is derived in Section 3 as well. In Section 4, we present an list decoding algorithm of Reed-Solomon codes beyond the Johnson-type bound.

2 Preliminaries

Let q be the finite field with q elements, where q is a power of a prime, and let $\mathbb{F}_q^n$ denote the set of all vectors of length n over $\mathbb{F}_q$. The Hamming weight of $\mathbf{x}$ is denoted by $\mathrm{wt}_{\mathsf{H}}(\mathbf{x})$. A q-ary Hamming metric code $\mathbf{C}$ of length n is a subset of $\mathbb{F}_q^n$. The code $\mathbf{C}$ is called $(\tau n, L)_{\mathsf{H}}$-list decodable if for every word $\mathbf{y} \in \mathbb{F}_q^n$, the intersection of $\mathbf{C}$ with the Hamming ball $\{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{wt}_{\mathsf{H}}(\mathbf{x} - \mathbf{y}) \leq \tau n\}$ has size at most L, here the parameter L is called the list size.

Then, we move to introduce the definitions of symbol-pair codes.

Definition 1. *(Symbol-pair Read Vector) Let $\mathbf{x} = [x_0, x_1, \cdots, x_{n-1}]$ be a vector in $\mathbb{F}_q^n$. The symbol-pair read vector of $\mathbf{x}$ is defined as*

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

The pair distance between two vectors in $\mathbb{F}_q^n$ is the Hamming distance between their corresponding pair vectors, where two pairs (a, b) and (c, d) are viewed as different if either $a \neq c$ or $b \neq d$.

Definition 2. *(Pair Distance) Let* $\mathbf{x} = (x_0, x_1, \cdots, x_{n-1})$ *and* $\mathbf{y} = (y_0, y_1, \cdots, y_{n-1})$ *be two vectors in* $\mathbb{F}_q^n$. *The pair distance between* $\mathbf{x}$ *and* $\mathbf{y}$ *is defined as*

$$\begin{aligned} d_{\mathsf{P}}(\mathbf{x}, \mathbf{y}) &= d_{\mathsf{H}}(\pi(\mathbf{x}), \pi(\mathbf{y})) \\ &= |\{0 \leq i \leq n-1 : (x_i, x_{i+1}) \neq (y_i, y_{i+1})\}|. \end{aligned}$$

The pair weight of a vector $\mathbf{x} \in \mathbb{F}_q^n$ is defined as $\mathrm{wt}_{\mathsf{P}}(\mathbf{x}) = d_{\mathsf{P}}(\mathbf{x}, \mathbf{0})$ where $\mathbf{0}$ is the all-zero vector of $\mathbb{F}_q^n$. The minimum pair distance of a code $\mathcal{C} \in \mathbb{F}_q^n$ is defined as

$$d_{\mathsf{P}}(\mathcal{C}) = \min_{\mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}} \{d_{\mathsf{P}}(\mathbf{x}, \mathbf{y})\}.$$

For $\mathbf{x}, \mathbf{y}$ in $\mathbb{F}_q^n$, let $0 < d_{\mathsf{H}}(\mathbf{x}, \mathbf{y}) < n$ be the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$. Then, we have

$$d_{\mathsf{H}}(\mathbf{x}, \mathbf{y}) + 1 < d_{\mathsf{P}}(\mathbf{x}, \mathbf{y}) < 2d_{\mathsf{H}}(\mathbf{x}, \mathbf{y}). \tag{1}$$

In the extreme cases, where $d_{\mathsf{H}}(\mathbf{x}, \mathbf{y})$ equals 0 or n, clearly $d_{\mathsf{H}}(\mathbf{x}, \mathbf{y}) = d_{\mathsf{P}}(\mathbf{x}, \mathbf{y})$.

A code over $\mathbb{F}_q$ of length n with size M and minimum pair distance d_{P} is called an $(n, M, d_{\mathsf{P}})_q$-symbol-pair code. Similar to classical Hamming metric codes, we can define the rate and the relative pair distance of an $(n, M, d_{\mathsf{P}})_q$-symbol-pair code $\mathcal{C}$ by

$$R(\mathcal{C}) = \frac{\log_q |\mathcal{C}|}{n} \quad \textit{and} \quad \delta(\mathcal{C}) = \frac{d_{\mathsf{P}} - 2}{n},$$

In literature, the relative distance of $\mathcal{C}$ is defined by $\frac{d_{\mathsf{P}}}{n}$. However, our definition of relative minimum distance given above will bring us advantage to handle some upper bounds like the Singleton bound.

The minimum pair distance is one of the important parameters for a symbol-pair code. A code $\mathcal{C}$ with minimum pair distance d_{P} can uniquely correct t pair errors if and only if $d_{\mathsf{P}} \geq 2t + 1$ see [2]. Hence, it is desirable to keep minimum pair distance d_{P} as large as possible for a symbol-pair code with fixed n. It has been shown [4] that an $(n, M, d_{\mathsf{P}})_q$-symbol-pair code $\mathcal{C}$ must obey the following version of the Singleton bound.

Lemma 1. *(Singleton Bound) Let* $q \geq 2$ *and* $2 \leq d_{\mathsf{P}} \leq n$. *If* $\mathcal{C}$ *is an* $(n, M, d_{\mathsf{P}})_q$*-symbol-pair code, then*

$$M \leq q^{n - d_{\mathsf{P}} + 2}.$$

An alternative way to state the Singleton bound for a symbol-pair code $\mathcal{C}$ *in term of its rate and relative minimum pair distance is*

$$R(\mathcal{C}) + \delta(\mathcal{C}) \leq 1.$$

An $[n, k, d_{\mathsf{P}}]_q$ symbol-pair code is an $\mathbb{F}_q$-linear code over $\mathbb{F}_q$ of length n, dimension k and minimum pair distance d_{P}.

The symbol-pair ball, as an analog to the Hamming metric ball, is used to count the number of words within a given pair distance.

Definition 3. *(Symbol-pair Ball) For a word $\mathbf{y} \in \mathbb{F}_q^n$ and a nonnegative real number r, the symbol-pair ball centered at $\mathbf{x}$ with radius r is defined by*

$$\mathcal{B}_{\mathsf{P}}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{F}_q^n : d_{\mathsf{P}}(\mathbf{x}, \mathbf{y}) \leq r\}.$$

Proposition 2. *(see in [2]) For any $\mathbf{x} \in \mathbb{F}_q^n$, the symbol-pair ball $\mathcal{B}_{\mathsf{P}}(\mathbf{x}, d)$ has size*

$$|\mathcal{B}_{\mathsf{P}}(\mathbf{x}, d)| = 1 + \sum_{i=1}^{d} \sum_{k=\lceil \frac{i}{2} \rceil}^{i-1} D(n, k, i-k)(q-1)^k, \tag{2}$$

where

$$\begin{aligned} D(n, \ell, w) &= \binom{\ell-1}{w-1}\left[\binom{n-\ell-1}{w} + 2\binom{n-\ell-1}{w-1}\right] + \binom{n-\ell-1}{w-1}\binom{\ell-1}{w} \\ &= \frac{n}{w} \cdot \binom{\ell-1}{w-1}\binom{n-\ell-1}{w-1}. \end{aligned}$$

As in the Hamming metric, the codes in the symbol-pair metric also achieve the following Gilbert-Varshamov Bound.

Lemma 3. *(Asymptotic Gilbert-Varshamov Bound, see in [3]) There exists a family of a-ary (n, M, d)-symbol-pair codes with rate $R = \lim_{n\to\infty} \frac{\log_q M}{n}$ and relative pair distance $\delta = \lim_{n\to\infty} \frac{d}{n}$ satisfying*

$$R \geq 1 - \max_{0 \leq \frac{\theta}{2} \leq \beta \leq \theta \leq \delta} \left(\beta H_q\left(\frac{2\beta - \theta}{\beta}\right) + (1-\beta) H_q\left(\frac{\theta - \beta}{1-\beta}\right)\right).$$

Remark 1. *Figure 1 reveals the gap between the Gilbert-Varshamov bound in symbol-pair metric and in Hamming metric when $q = 17$. In other words, the codes attaining this bound in symbol-pair metric achieves better trade-off in terms of rate and relative distance.*

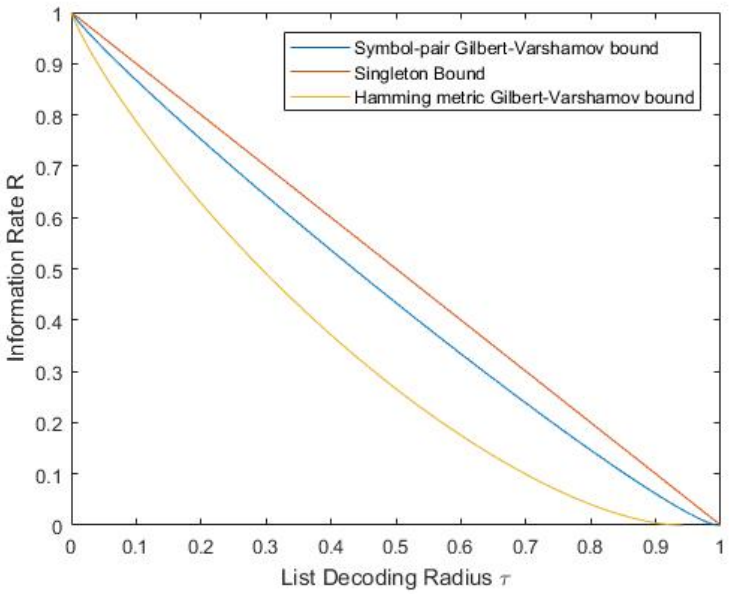

Figure 1: Comparison of the Gilbert-Varshamov bound in Hamming Metric and Symbol-pair Metric.

We now proceed to the definition of list decoding of symbol-pair codes.

Definition 4. *For a real $\tau \in (0,1)$, a symbol-pair code $\mathcal{C} \subseteq \mathbb{F}_q^n$ is said to be $(\tau n, L)_{\mathsf{P}}$-list decodable, if for every $\mathbf{x} \in \mathbb{F}_q^n$, we have*

$$|\mathcal{B}_{\mathsf{P}}(\mathbf{x}, \tau n) \cap \mathcal{C}| \leq L.$$

3 Bounds on the list decoding radius of Symbol-pair Codes

3.1 An upper bound on list decodbility of symbol-pair codes

The Gilbet-Varshamov bound plays a role as an upper bound on the list decoding radius of codes under various metrics, i.e., the Hamming metric codes [11], rank-metric codes [6] and cover-metric codes [17]. It is not surprised that the Gilbert-Varshamov bound is also an upper bound on the list decoding radius of the symbol-pair codes.

In this subsection, we show that list decoding of any symbol-pair code cannot exceed the Gilbert-Varshamov bound. The idea of our proof is based on counting the words in a symbol-pair ball. We firstly estimate the size of a symbol-pair ball.

Lemma 4. *Given a vector $\mathbf{a} \in \mathbb{F}_q^n$, the size of the symbol-pair ball $\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \delta n)$ satisfies*

$$|\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \delta n)| = q^{\kappa_{sp}(\delta)n + o(n)}, \tag{3}$$

where

$$\kappa_{sp}(\delta) = \max_{0 \leq \frac{\theta}{2} \leq \beta \leq \theta \leq \delta} \beta H_q\left(\frac{2\beta - \theta}{\beta}\right) + (1-\beta) H_q\left(\frac{\theta - \beta}{1-\beta}\right), \tag{4}$$

and $H_q(x) = x\log_q(q-1) - x\log_q x - (1-x)\log_q(1-x)$ is the q-ary entropy function.

Proof. By the equation (2), the size of the symbol-pair ball is

$$|\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \delta n)| \quad = \quad 1 + \sum_{i=1}^{\delta n} \sum_{k=\lceil \frac{i}{2} \rceil}^{i-1} \frac{n}{i-k} \cdot \binom{k-1}{i-k-1} \binom{n-k-1}{i-k-1} (q-1)^k.$$

Let $k = \beta n$ and $i = \theta n$, for some reals $\beta \in (0,1)$ and $\theta \in (0,1)$, we have

$$\binom{k-1}{i-k-1} = 2^{\beta n H_2(\frac{2\beta-\theta}{\beta}) + o(n)}, \ \binom{n-k-1}{i-k-1} = 2^{(1-\beta) n H_2(\frac{\theta-\beta}{1-\beta}) + o(n)},$$

this implies

$$\frac{n}{i-k} \cdot \binom{k-1}{i-k-1} \binom{n-k-1}{i-k-1} (q-1)^k = q^{\beta n H_q\left(\frac{2\beta-\theta}{\beta}\right) + (1-\beta) n H_q\left(\frac{\theta-\beta}{1-\beta}\right) + o(n)}.$$

Thus

$$q^{\kappa_{sp}(\delta)n + o(n)} \leq |\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \delta n)| \leq (\delta n)^2 q^{\kappa_{sp}(\delta)n + o(n)} = q^{\kappa_{sp}(\delta)n + o(n)}.$$

The desired result follows. □

To simplify the notation, we denote $\kappa_{sp}(\delta)$ by κ_{sp} if there is no confusion.

Remark 2. *Lemma 4 simply says that*

$$\lim_{n\to\infty} \frac{\log_q |\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \delta n))|}{n} = \kappa_{sp}.$$

The following theorem shows that the Gilbert-Varshamov bound is an upper bound on the list decoding radius of symbol-pair codes.

Theorem 1. *Assume that a symbol-pair code $\mathcal{C}$ of rate R is $(\tau n, L)_{\mathsf{P}}$-list decodable with list size $L = \mathrm{poly}(n)$. Then, the rate R of $\mathcal{C}$ must obey*

$$R \leq 1 - \kappa_{sp}(\tau) = 1 - \max_{0\leq\frac{\theta}{2}\leq\beta\leq\tau} \left(\beta H_q\left(\frac{2\beta-\theta}{\beta}\right) + (1-\beta)H_q\left(\frac{\theta-\beta}{1-\beta}\right)\right)$$

for all sufficiently large n, where $\kappa_{sp}(\tau)$ is given in (4).

Proof. We prove it by contradiction. Simply denote $\kappa_{sp}(\tau)$ by κ_{sp}. Assume that there exists a symbol-pair code $\mathcal{C}$ of rate R such that $R \geq 1 - \kappa_{sp} + \epsilon$ for some positive constant ϵ. Let L be the upper bound of the list size of this code. Define the set

$$\mathcal{A} = \{(\mathbf{c}, \mathbf{v}) : d_{\mathsf{P}}(\mathbf{c}, \mathbf{v}) \leq \tau n, \mathbf{c} \in \mathcal{C}, \mathbf{v} \in \mathbb{F}_q^n\}.$$

We find two ways to calculate the size of this set. First, for every vector $\mathbf{v}$ in $\mathbb{F}_q^n$, it holds that $|B_{\mathsf{P}}(\mathbf{v}, \tau n) \cap \mathcal{C}| \leq L$. This implies

$$|\mathcal{A}| = \sum_{\mathbf{v}\in\mathbb{F}_q^n} |B_{\mathsf{P}}(\mathbf{v}, \tau n) \cap \mathcal{C}| \leq q^n L.$$

On the other hand, by Lemma 4 we have $|B_{\mathsf{P}}(\mathbf{c}, \tau n)| \geq q^{\kappa_{sp} n - \frac{\epsilon}{2} n}$ for all sufficiently large n. Thus

$$|\mathcal{A}| = \sum_{\mathbf{c}\in\mathcal{C}} |B_{\mathsf{P}}(\mathbf{c}, \tau n)| \geq q^{Rn} q^{\kappa_{sp} n - \frac{\epsilon}{2} n}.$$

Combining them together gives us

$$L \geq q^{Rn + \kappa_{sp} n - \frac{\epsilon}{2} n - n} \geq q^{\frac{\epsilon}{2} n}.$$

A contradiction occurs. □

3.2 List decoding of random symbol-pair codes

In the previous subsection, we show that list decodability of every symbol-pair codes does not exceed the Gilbert-Varshamov bound. In this subsection, we investigate list decodability of random symbol-pair codes. We show that random symbol-pair codes can be list decoded up to the Gilbert-Varshamov bound with high probability. In particular, most symbol-pair codes can be list decoded up to the Gilbert-Varshamov bound with constant list size $O(1/\epsilon)$,

Theorem 2. *For small $\epsilon \in (0,1)$ with a probability at least $1-q^{-n}$, a random symbol-pair code $\mathcal{C} \subseteq \mathbb{F}_q^n$ of rate*

$$R = 1 - \kappa_{sp}(\tau) - \epsilon = 1 - \max_{0 \le \frac{\theta}{2} \le \beta \le \tau} \left(\beta H_q \left(\frac{2\beta - \theta}{\beta} \right) + (1-\beta) H_q \left(\frac{\theta - \beta}{1-\beta} \right) \right) - \epsilon$$

is $(\tau n, O(1/\epsilon))_{\mathsf{P}}$-list decodable for sufficiently large n.

Proof. Put $L = \left\lceil \frac{4}{\epsilon} \right\rceil - 1$. By Lemma 4, for all sufficiently large n, we have $|\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)| \le q^{\kappa_{sp} n + \frac{\epsilon}{2} n}$. Pick a symbol-pair code $\mathcal{C}$ with size q^{Rn} uniformly at random. Let us upper bound the probability that $\mathcal{C}$ is not $(\tau n, L)_{\mathsf{P}}$-list decodable.

If $\mathcal{C}$ is not $(\tau n, L)_{\mathsf{P}}$-list decodable, there exists a word $\mathbf{a} \in \mathbb{F}_q^n$ and a subset $\mathcal{S} \subseteq \mathcal{C}$ with $|\mathcal{S}| = L+1$ such that $\mathcal{S} \subseteq \mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)$. The probability that codeword $\mathbf{c} \in \mathcal{C}$ is contained in $\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)$ is

$$\Pr[c \in \mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)] = \frac{|\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)|}{q^n} \le q^{\kappa_{sp} n + \frac{\epsilon}{2} n} \cdot q^{-n}. \tag{5}$$

Let $E_{\mathbf{a},\mathcal{S}}$ be the event that all codewords in $\mathcal{S}$ are contained in $\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)$. By Equation (5), we have

$$\Pr[E_{\mathbf{a},S}] \le \left(\frac{|\mathcal{B}_{\mathsf{P}}(\mathbf{a}, \tau n)|}{q^n} \right)^{L+1} \le \left(q^{\kappa_{sp} n + \frac{\epsilon}{2} n} \cdot q^{-n} \right)^{L+1}.$$

Taking the union bound over all q^n choices of $\mathbf{a}$ and $\mathcal{S}$ over any $(L+1)$-subsets of $\mathcal{C}$, we have

$$\begin{aligned}
\sum_{\mathbf{a},\mathcal{S}} \Pr[E_{\mathbf{a},\mathcal{S}}] &\le q^n \cdot \binom{|\mathcal{C}|}{L+1} \cdot \left(q^{\kappa_{sp} n + \frac{\epsilon}{2} n} \cdot q^{-n} \right)^{L+1} \\
&\le q^n \cdot |\mathcal{C}|^{L+1} \cdot q^{(\kappa_{sp} n + \frac{\epsilon}{2} n)(L+1)} \cdot q^{-n(L+1)} \\
&\le q^n \cdot q^{Rn(L+1)} \cdot q^{(\kappa_{sp} n + \frac{\epsilon}{2} n - n)(L+1)} \\
&= q^{n(L+1)\left(\frac{1}{L+1} + R + \kappa_{sp} + \frac{\epsilon}{2} - 1\right)} \\
&\le q^{n(L+1)\left(\frac{\epsilon}{4} + R + \kappa_{sp} + \frac{\epsilon}{2} - 1\right)} \le q^{-n}.
\end{aligned}$$

The last inequality holds since $R = 1 - \kappa_{sp} - \epsilon$. Thus, a symbol-pair code $\mathcal{C}$ with rate R is not $(\tau n, L)_{\mathsf{P}}$-list decodable with probability at most q^{-n}. □

3.3 The Johnson-type bound

The Johnson-type bound in the topic of list decoding usually provides a lower bound on list decoding radius in terms of minimum distance of a code. However, for some metrics such as rank-metric, the Johnson-type bound does not exist. In this section, we show that one has a Johnson-type bound for pair metric. On the hand hand, there is an evidence showing that the Johnson-type bound given in this subsection is tight.

Theorem 3. *(Johnson-type Bound) Any symbol-pair code $\mathcal{C}$ in $\mathbb{F}_q^n$ with relative distance δ is $(\tau n, 2(q^2-1)nd)$-list decodable for*

$$\tau = \frac{q^2-1}{q^2} \left(1 - \sqrt{1 - \frac{q^2 \delta}{q^2 - 1}} \right)$$

Proof. We fix a vector $\mathbf{y} \in \mathbb{F}_q^n$. Assume that $B_{\mathsf{P}}(\mathbf{y}, \tau n) \cap \mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_L\}$ for L. Our goal is to bound the size L. Let $\mathbf{v}_i = \mathbf{c}_i - \mathbf{y}$. Since $d_{\mathsf{P}} \geq d_{\mathsf{P}}(\mathcal{C})$, we have $d_{\mathsf{P}}(\mathbf{v}_i, \mathbf{v}_j) = d_{\mathsf{P}}(\mathbf{c}_i, \mathbf{c}_j) \geq \delta n$ for every pair (i, j) and $\mathrm{wt}_{\mathsf{P}}(\mathbf{v}_i) \leq \tau n$ for every i. We denote $\mathbf{v}_i$ as $(v_{i,1}, \ldots, v_{i,n}) \in \mathbb{F}_q^n$. By the definition of symbol-pair error, we have

$$\begin{aligned}\frac{L(L-1)\delta n}{2} &\leq \sum_{1 \leq i < j \leq L} d_{\mathsf{P}}(\mathbf{v}_i, \mathbf{v}_j) = \sum_{1 \leq i < j \leq L} |\{k : (v_{i,k}, v_{i,k+1}) \neq (v_{j,k}, v_{j,k+1})\}| \\ &= \sum_{k=1}^{n} |\{(i,j) : (v_{i,k}, v_{i,k+1}) \neq (v_{j,k}, v_{j,k+1}), 1 \leq i < j \leq L\}|.\end{aligned}$$

Next, we fix the coordinate pair $(1, 2)$. Let $x_{a,b}$ be the number of pairs (a, b) among the set $\{(v_{i,1}, v_{i,2}) \in \mathbb{F}_q^2 : 1 \leq i \leq L\}$. It is clear that $\sum_{(a,b) \in \mathbb{F}_q^2} x_{a,b} = L$. It follows that

$$\begin{aligned}&|\{(i,j) : (v_{i,1}, v_{i,2}) \neq (v_{j,1}, v_{j,2}), 1 \leq i < j \leq L\}| = \sum_{(a,b) \in \mathbb{F}_q^2} x_{a,b}(L - x_{a,b}) \\ = &\left(L^2 - x_{0,0}^2 - \sum_{(a,b) \in \mathbb{F}_q^2/(0,0)} x_{a,b}^2\right) \leq \left(L^2 - x_{0,0}^2 - \frac{1}{q^2-1}\left(\sum_{(a,b) \in \mathbb{F}_q^2/(0,0)} x_{a,b}\right)\right) \\ = &\left(L^2 - x_{0,0}^2 - \frac{1}{q^2-1}(L - x_{0,0})^2\right)\end{aligned}$$

The inequality above is due to the Cauthy-Schwarz inequality. We can apply this argument to every pair of adjacent coordinates $(k, k+1)$. Let a_k be the number of pairs $(0, 0)$ among the set $\{(v_{i,k}, v_{i,k+1}) \in \mathbb{F}_q^2 : 1 \leq i \leq L\}$. Putting these two formulas together gives us

$$\begin{aligned}\frac{L(L-1)\delta n}{2} &\leq nL^2 - \sum_{k=1}^{n}\left(a_k^2 + \frac{1}{q^2-1}(L - a_k)^2\right) \\ &= \frac{2}{q^2-1}L\sum_{k=1}^{n} a_k - \frac{q^2}{q^2-1}\sum_{k=1}^{n} a_k^2 + n\frac{q^2-2}{q^2-1}L^2 \\ &\leq -\frac{q^2}{n(q^2-1)}\left(\sum_{k=1}^{n} a_k\right)^2 + \frac{2}{q^2-1}L\left(\sum_{k=1}^{n} a_k\right) + n\frac{q^2-2}{q^2-1}L^2\end{aligned}$$

Let $\sum_{k=1}^{n} a_k = Le$ and we then have

$$-\frac{q^2}{n(q^2-1)}L^2e^2 + \frac{2}{q^2-1}L^2e - \frac{L(L-1)\delta n}{2} + n\frac{q^2-2}{q^2-1}L^2 \geq 0.$$

This implies

$$L \leq \frac{2\delta n}{\frac{q^2e^2}{n(q^2-1)} - \frac{2e}{q^2-1} + \delta n - \frac{(q^2-2)n}{q^2-1}}. \tag{6}$$

The condition $\frac{q^2e^2}{n(q^2-1)} - \frac{2e}{q^2-1} + \delta n - \frac{(q^2-2)n}{q^2-1} > 0$ leads to

$$\frac{e}{n} < \frac{1}{q^2} + \frac{q^2-1}{q^2}\sqrt{1 - \frac{q^2\delta}{q^2-1}}.$$

This implies

$$(n-q^2e) > (q^2-1)n\sqrt{1-\frac{q^2\delta}{q^2-1}}.$$

Squaring both sides and observing that $\delta = \frac{d}{n}$ yields

$$(n-q^2e)^2 > (q^2-1)^2n^2 - (q^2-1)q^2nd.$$

Since both sides are integers, we obtain $(n-q^2e)^2 \geq (q^2-1)^2n^2 - (q^2-1)q^2nd + 1$. Observe that (6) is equivalent to

$$L \leq \frac{2(q^2-1)dn}{(n-q^2e)^2 - (q^2-1)^2n^2 + (q^2-1)q^2nd} \leq 2(q^2-1)dn.$$

Then, the desired result follows. □

One may wonder if the Johnson-type Bound derived in this subsection is optimal. We find that the codes in [1] can be used to illustrate that the Johnson-type bound derived in this subsection is at least very close to optimality though we do not have an affirmative answer.

The paper [1] focused on the low-degree linearized polynomials that agrees with a given high-degree linearized polynomials on many coordinates. The following lemma summarize their results. Fix n distinct elements $\alpha_1, \ldots, \alpha_n$. For a polynomial $f(x) \in \mathbb{F}_q[x]$, we denote by $\mathbf{c}_f$ the vector $(f(\alpha_1), \ldots, f(\alpha_n))$. We abuse notations and denote by $d_{\mathsf{P}}(a(x), b(x))$ (and $d_{\mathsf{H}}(a(x), b(x))$, respectively) the symbol-pair distance (and the Hamming distance, respectively) between $\mathbf{c}_a$ and $\mathbf{c}_b$.

Lemma 5 ([1, Theorem 2.1]). *Let ℓ be a prime power and m a positive integer. Put $q = \ell^m$. Let u and v be integers such that $0 \leq u \leq v \leq m$. Then, there is a family $\mathcal{P} \subseteq \mathbb{F}_{\ell^m}[X]$ of* linearized[1] *polynomials of degree ℓ^u and a linearized polynomial $w(x)$ such that*

1. $|\mathcal{P}| \geq \ell^{(u+1)m-v^2}$;
2. *for all $P(x) \in \mathcal{P}$, $d_{\mathsf{H}}(P(x), w(x)) \leq \ell^m - \ell^v$;*
3. $w(x) = x^{\ell^v} + \sum_{i=u+1}^{v-1} a_i x^{\ell^i}$.

Based on this lemma, we have the following result that leads to some symbol-pair codes we need to illustrate optimality of the Johnson-type Bound given in this subsection.

Lemma 6. *Let ℓ be a prime power and m a positive integer. Put $q = \ell^m$. Let u and v be integers such that $0 \leq u \leq v \leq m$. Then, there is a family $\mathcal{P} \subseteq \mathbb{F}_{\ell^m}[X]$ of linearized polynomials of degree ℓ^u and a linearized polynomial $w(x)$ such that*

1. $|\mathcal{P}| \geq \ell^{(u+1)m-v^2}$;
2. *for all $P(x) \in \mathcal{P}$, $d_{\mathsf{P}}(P(x), w(x)) \leq \ell^m - \frac{(\ell-2)}{\ell-1}(\ell^v - 1)$;*
3. $w(x) = x^{\ell^v} + \sum_{i=u+1}^{v-1} a_i x^{\ell^i}$.

[1] They did not mention "linearized" in this theorem. Judged from their construction, $\mathcal{P}$ is indeed a family of linearized polynomials.

Proof. Let $[v] = \{\lambda v : \lambda \in \mathbb{F}_\ell^*\}$ and $h = \frac{\ell^m-1}{\ell-1}$. As we know, the set $\mathbb{F}_{\ell^m}^*$ can be partitioned into h disjoint subsets $[v_1], \ldots, [v_h]$. Since the distance of symbol-pair code is greatly affected by the order of its coordinates, we start our proof by arranging the order of coordinates. Given a polynomial $f(x) \in \mathbb{F}_{\ell^m}[X]$, the codeword generated by $f(x)$ is $(f(0), f([v_1]), f([v_2]), \ldots, f([v_h]))$ where $f([v_i]) \triangleq (f(\lambda v_i))_{\lambda \in \mathbb{F}_\ell^*}$. Let $\mathcal{P}$ and $w(x)$ be the family of linearized polynomials and linearized polynomials given by Theorem 5. For any $P(x) \in \mathcal{P}$, let us bound the symbol-pair distance of $P(x)$ and $w(x)$ under the above order of coordinates. By Theorem 5, the linearized polynomial $g_P(x) \triangleq P(x) - w(x)$ has at least ℓ^v roots. Moreover, if $u \in \mathbb{F}_{\ell^m}^*$ subject to $g_P(u) = 0$, then $g_P(\lambda u) = 0$ for every $\lambda \in \mathbb{F}_{\ell^m}^*$. Assume that $u \in [v_i]$ and we have $g_P([v_i]) = \mathbf{0} \in \mathbb{F}_{\ell^m}^{\ell-1}$. It follows that $g_P([v_i])$ contributes $\ell - 2$ pairs of symbols $(0,0) \in \mathbb{F}_{\ell^m}^2$. In summary, the ℓ^v roots of $g_P(x)$ yields at least $(\ell-2)\frac{(\ell^v-1)}{\ell-1}$ pairs of adjacent coordinates whose symbol patterns are $(0,0) \in \mathbb{F}_{\ell^m}^2$. The desired result follows since

$$d_P(P(x), w(x)) = \mathrm{wt}_P(P(x) - w(x)) \leq \ell^m - \frac{(\ell-2)}{\ell-1}(\ell^v - 1).$$

□

Example 1. In this example, we illustrate optimality of the Johnson-type bound given in this subsection.

We follows the parameter setting in [1]. Let ℓ be a prime power and m a positive integer. Put $q = \ell^m$. Lemma 6 yields a symbol-pair code with list decoding radius at most $1 - \frac{\ell-2}{\ell-1}\ell^{v-m}$. The dimension of this code is $K := \ell^u$ and the length of this code is $N := \ell^m$. Setting $u = \delta m$ and $v = \rho m$ gives the list size $|\mathcal{P}| \geq N^{(\delta-\rho^2)\log_\ell N}$ which is super-polynomial in length N for any constant $\delta > \rho^2$. To compare it with our Johnson-type bound, we set $\delta = 1 - \gamma$ and $\rho = 1 - \frac{\gamma}{2} - \frac{\gamma^2}{4}$ for small constant γ. One can check that it satisfies $\delta > \rho^2$ for small constant γ. Let $\ell = \frac{1}{\epsilon}$ and the relative decoding radius then becomes

$$1 - \frac{\ell-2}{\ell-1}\ell^{\rho m - m} = 1 - (1-\epsilon)N^{-\frac{\gamma}{2}-\frac{\gamma^2}{4}}.$$

On the other hand, our Johnson-type bound gives the relative list decoding radius $(1 - \frac{1}{N^2})(1 - N^{-\frac{\gamma}{2}}) \approx 1 - N^{-\frac{\gamma}{2}}$. Thus, the upper bound is very close to the Johnson-type bound for rate $R = N^{-\gamma}$. This implies that the Johnson-type bound given in this subsection is very close to optimality if it is not optimal.

4 List decoding of Reed-Solomon codes beyond the Johnson-type bound

It is well known that any Reed-Solomon codes can be efficiently list decoded up to the Johnson bound for the Hamming metric with the help of famous Guruswami-Sudan list decoding algorithm. On the other hand, some evidence shows that there exist Reed-Solomon codes and subcodes of Reed-Solomon codes that can not be list decoded slightly beyond the Johnson bound for the Hamming metric. Given the importance of Reed-Solomon code in both theory and practice, one would like to clearly understand the limits to the list decoding issue of Reed-Solomon codes. However, we are still far away from this goal anyway for the Hamming metric. It is not even clear whether there exist Reed-Solomon codes that can be list decoded beyond the Johnson bound for the Hamming metric.

On the other hand, one also wonders if Reed-Solomon codes can be list decoded beyond the Johnson bound for the pair metric. In this subsection, we give this question an affirmative answer by showing that Reed-Solomon codes can indeed be list decoded beyond the Johnson-type bound.

The construction comes from the folded Reed-Solomon code. Let us first explain the intuition behind this construction. By the definition of symbol-pair error, each error corresponds to a pair of adjacent coordinates. In our list decoding algorithm, instead of inputting the evaluations index by index, we input the evaluations pair by pair. The question arises whether we can exploit this input to improve our list decoding algorithm. Note that the famous Guruswami-Sudan list decoding algorithm fails to serve our purpose. We turn to the list decoding algorithm of folded Reed-Solomon code in [10] instead. Let γ be a primitive element of $\mathbb{F}_q$.

We now consider list decoding of folded Reed-Solomon code. Let γ be a primitive element of $\mathbb{F}_q$. Let $1 \leq k \leq n \leq q-1$. We encodes the polynomial f of degree at most $k-1$ to the codewords $\mathbf{c}_f := (f(1), f(\gamma), \ldots, f(\gamma^{n-1}))$ and

$$\mathbf{c}_f^{(2)} := \begin{pmatrix} f(1) & f(\gamma) & f(\gamma^2) & \cdots & f(\gamma^{n-2})) \\ f(\gamma) & f(\gamma^2) & f(\gamma^3) & \cdots & f(\gamma^{n-1}) \end{pmatrix} \tag{7}$$

Then the Reed-Solomon code $RS[n,k]$ and the folded Reed-Solomon $FRS[n,k]$ are defined by

$$RS[n,k] := \{\mathbf{c}_f : \ f \in \mathbb{F}_q[x], \ \deg(f) \leq k-1\}. \tag{8}$$

and

$$FRS[n-1,k] := \{\mathbf{c}_f^{(2)} : \ f \in \mathbb{F}_q[x], \ \deg(f) \leq k-1\}. \tag{9}$$

respectively. List decoding of folded Reed-Solomon codes were first considered in [10]. The main idea of the following result can be found in [10]. However, for the sake of completeness, let us derive an explicit list decoding algorithm of folded Reed-Solomon codes defined above.

Lemma 7. *The folded Reed-Solomon code $FRS[n-1,k]$ defined in (9) is $(\tau(n-1), q)_{\mathsf{H}}$-list decodable with $\tau = \frac{2}{3} \times \frac{n-2-k}{n-1}$.*

Proof. Assume that $\mathbf{c}_f^{(2)}$ was transmitted and

$$\mathbf{b}^{(2)} := \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_n \\ b_1 & b_2 & b_3 & \cdots & b_n \end{pmatrix}$$

is received with at most τn errors. Thus, $d_{\mathsf{H}}(\mathbf{c}_f^{(2)}, \mathbf{b}^{(2)}) \leq \tau(n-1)$. Put $m = \lceil (n-k)/3 \rceil$. Then one has $3m+k+2 > n-1$. Consider the interpolation polynomial $Q(x, y_1, y_2) := a_0(x) + a_1(x)y_1 + a_2(x)y_2 \in \mathbb{F}_q[x, y_1, y_2]$ with coefficients of $a_i(x)$ to be determined subject to $\deg(a_0) \leq m+k-1$, $\deg(a_1) \leq m$ and $\deg(a_2) \leq m$. Consider the homogenous equation system $a_0(\gamma^{i-1}) + a_1(\gamma^{i-1})a_i + a_2(\gamma^{i-1})b_i = 0$ for $i = 1, 2, \ldots, n-1$. For this equation system, coefficients of $a_i(x)$ are viewed as variables. Thus, there are $3m+k+2$ variables and $n-1$ equations. Hence, there are polynomials $a_0(x), a_1(x), a_2(x) \in \mathbb{F}_q[x]$ with $\deg(a_0) \leq m+k-1$, $\deg(a_1) \leq m$ and $\deg(a_2) \leq m$ that are not all zero such that $a_0(\gamma^{i-1}) + a_1(\gamma^{i-1})a_i + a_2(\gamma^{i-1})b_i = 0$ for $i = 1, 2, \ldots, n-1$. Since $d_{\mathsf{H}}(\mathbf{c}_f^{(2)}, \mathbf{b}^{(2)}) \leq \tau n$, there are at least $n-1-\tau(n-1)$ i's such that $a_0(\gamma^{i-1}) + a_1(\gamma^{i-1})f(\gamma^{i-1}) + a_2(\gamma^{i-1})f(\gamma^i) = 0$. Hence, the polynomial $a_0(x) + a_1(x)f(x) + a_2(x)f(\gamma x)$ has at least $n-1-\tau(n-1)$ roots. On the other hand, $\deg(a_0(x) + a_1(x)f(x) + a_2(x)f(\alpha x)) \leq m+k-1$ and we also have $n-1-\tau(n-1) > m+k-1$, this forces that $a_0(x) + a_1(x)f(x) + a_2(x)f(\gamma x)$ is identical to 0. Note that $x^{q-1} - \gamma$ is irreducible and $x^q \equiv \gamma x \pmod{x^{q-1} - \gamma}$. This gives

$$0 = a_0(x) + a_1(x)f(x) + a_2(x)f(\gamma x) \equiv a_0(x) + a_1(x)f(x) + a_2(x)f^q(x) \pmod{x^{q-1} - \gamma}.$$

In other words, $f(x)$ is a solution of the equation $a_0(x)+a_1(x)z+a_2(x)z^q = 0$ over the field $\mathbb{F}_q[x]/(x^{q-1}-\alpha) \simeq \mathbb{F}_{q^{q-1}}$. Hence, this equation has at most q roots in $\mathbb{F}_q[x]/(x^{q-1}-\gamma)$. Since $\deg(f(x)) < q-1$, the equation $a_0(x) + a_1(x)f(x) + a_2(x)f(\gamma x) = 0$ has at most q roots in $\mathbb{F}_q[x]$. □

By applying Lemma 7 and considering the relation between Hamming distance and pair distance, we immediately obtain the following result.

Theorem 4. *The Reed-Solomon code $RS[n,k]$ over $\mathbb{F}_q$ for any $1 \le k \le n \le q$ is $(\tau n, q)_{\mathsf{P}}$-list decodable with $\tau = \frac{2}{3} \times \frac{n-2-k}{n}$.*

Lemma 8. *The Reed-Solomon code $RS[n,k]$ over $\mathbb{F}_q$ for any $1 \le k < n \le q$ has pair minimum distance at $n-k+2$.*

Proof. Consider the polynomial $f(x) = \prod_{i=0}^{k-2}(x-\gamma^i)$. Then the codeword $\mathbf{c}_f$ has Hamming weight $n-k+1$ and the pair weight $n-k+2$. This completes the proof. □

Theorem 5. *The Reed-Solomon code $RS[n,k]$ over $\mathbb{F}_q$ for any $1+n/2 \le k < n \le q$ is $(\tau n, q)_{\mathsf{P}}$-list decodable with $\tau = \frac{2}{3}\delta + o(1)$, where $\delta = \frac{n-k+2}{n}$ is the relative pair minimum distance of $RS[n,k]$. Hence, if n is proportional q and $0 < \delta < \frac{3}{4}$, then $RS[n,k]$ can be list decoded beyond the Johnson-type bound with list size $O(n)$.*

Proof. When n is proportional to q, the list size given in Theorem 4 is $O(n)$. For sufficiently large n (thus q is also large), the Johnson-type bound given in Theorem 3 becomes $1-\sqrt{1-\delta}+o(1)$. On the other hand, by Lemma 8, the relative minimum distance of $RS[n,k]$ is $\delta = \frac{n-k+2}{n}$ for $\delta < 1/2$. Furthermore, it is easy to verify that $\frac{2}{3}\delta > 1-\sqrt{1-\delta}$ for $0 < \delta < \frac{3}{4}$. □

References

[1] E. Ben-Sasson, S. Kopparty and J. Radhakrishnan, Subspace Polynomial and Limits to List Decoding of Reed-Solomon Codes, *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp 113-120, 2010.

[2] Y. Cassuto, M. Blaum, Codes for Symbol-Pair Read Channels, *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp 8011-8020, 2011.

[3] Y. Cassuto and S. Litsyn, Symbol-pair codes: Algebraic constructions and asymptotic bounds, *IEEE International Symposium on Information Theory*, pp. 2348-2352 (2011).

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

[5] B. Ding, G. Ge, J. Zhang, T. Zhang and Y. Zhang, New constructions of MDS symbol-pair codes, *Des. Codes Cryptogr*. (2018) 86:841-859

[6] Y. Ding, On list-decodability of random rank-metric codes and subspace codes, *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp 51-59, 2015.

[7] Z. Dvir and S. Lovett, subspace evasive sets, *Proceedings of the 44th ACM Symposium on Theory of Computing,* pp: 351- 358,2012.

[8] P. Elias, List decoding for noisy channels, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1957.

[9] V. Guruswami, List Decoding of Error-Correcting Codes, Springer, US, 2001.

[10] V. Guruswami and A. Rudra, Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy, *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp 135-150, 2008.

[11] V. Guruswami and S. Vadhan, A low bound on list size for list decoding, *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp 5681-5688, 2010.

[12] V. Guruswami and C. Xing, List decoding Reed-Solomon, Algebraic Geometric, and Gabidulin subcodes up to the Singleton bound, in *Electronic Colloquium on Computational Complexity (ECCC),* 19:146, 2012. Extended abstract appeared in the Proceedings of the 45th ACM Symposium on Theory of Computing (STOC'13).

[13] X. Kai, S. Zhu, P. Li, A construction of new MDS symbol-pair codes, *IEEE Transactions on Information Theory*, 61(11), 5828-5834 (2015).

[14] M. Takita, M. Hirotomo and M. Morii, Error-Trapping decoding for cyclic codes over symbol-pair read channels, *International Symposium on Information Theory and Its Applications,* pp. 681-685, California, USA, 2016.

[15] M. Hirotomo, M. Takita and M. Morii, Syndrome decoding of symbol-pair codes, *IEEE Information Theory Workshop*, pp. 162-166, Australia, 2014.

[16] S. Horii, T. Matsushima and S. Hirasawa, Linear Programming decoding of binary linear codes for symbol-pair read channels, in *IEEE International Symposium on Information Theory,* pp. 1944-1948, Spain, 2016.

[17] S. Liu, C. Xing and C. Yuan, List decoding of cover-metric codes up to the Singleton bound, *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp 2410-2416, 2018.

[18] J. M. Wozencraft, List decoding, Quarterly Progress Report, Research Laboratory of Electronics, MIT, 48, pp. 90-95, 1958.

[19] E. Yaakobi, J. Bruck and P. H. Siegel, Constructions and decoding of cyclic codes over b-symbol read channels, *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp 1541-1551.

[20] E. Yaakobi, J. Bruck and P. H. Siegel, Decoding of cyclic codes over symbol-pair read channels, *IEEE International Symposium on Information Theory,* Cambridge, MA, pp. 2891-2895, 2012.