Symbol-Pair Codes: Algebraic Constructions and Asymptotic Bounds

Yuval Cassuto
EPFL
School of Computer and Communication Sciences
ALGO Laboratory
1015 Lausanne, Switzerland
yuval.cassuto@epfl.ch

Simon Litsyn
Tel-Aviv University
Department of Electrical Engineering - Systems
Ramat-Aviv 69978
Tel-Aviv, Israel
litsyn@eng.tau.ac.il

***Abstract*— For the recently proposed model of symbol-pair channels, we advance the pair-error coding theory with algebraic cyclic-code constructions and asymptotic bounds on code rates. Cyclic codes for pair-errors are constructed by a careful use of duals of known tools from cyclic-code theory. Asymptotic lower bounds on code rates show that codes for pair-errors provably exist for rates strictly higher than codes for the Hamming metric.**

I. Introduction

Symbol-pair coding theory has been proposed in [2] to address channels that output pairs of overlapping symbols, rather than one symbol at a time. The pair-channel model is mainly motivated by magnetic-storage channels with high write resolution, whose lower read resolution prevents individual-symbol read-out. This previous work has laid out a coding-theoretic framework to combat pair-errors over symbol-pair channels. In particular, it developed a combinatorial representation using pair-vectors, proved necessary and sufficient conditions for pair-error correction, and analyzed the relationship between the pair-distance and Hamming distance. This framework then was used for elementary code construction, and combinatorial upper and lower bounds on code sizes using pair-sphere analysis.

This current paper addresses two important questions left open by [2]. The first is the construction of efficient codes in the pair-metric. Previously proposed interleaving-based constructions give less than satisfactory results. The second is asymptotic bounds on code rates in the pair-metric. The previously known combinatorial bounds are useful when specific parameters are given, but do not offer insight on the asymptotic achievability and limits of codes for pair-errors. After a brief review of essential results from [2] given in section II, we treat the construction problem in section III by developing tools from cyclic-code theory that give good pair-error correctability. Lower bounds on pair-distances of cyclic codes are obtained by using duals of algebraic tools for Hamming-metric codes. These methods allow showing, for example, that cyclically constructed Hamming codes are perfect in the pair-metric as well, while non-cyclic Hamming codes are in general not. Then in section IV we derive asymptotic lower and upper bounds on the rates of codes with given relative pair-distances. A pair-metric Gilbert-Varshamov bound shows that codes for pair-errors are known to exist with strictly higher rates compared to codes for the Hamming metric with the same relative distance. In particular, codes with relative pair-distance up to 0.75 exist asymptotically, compared to a 0.5 cut-off for Hamming-metric codes. A subsequent Plotkin-style bound shows that the 0.75 bound for existence is tight, and is the true cut-off for pair-metric codes.

II. Review of Symbol-Pair Coding Theory

Building on the coding-theoretic framework for pair-error correction developed in [2], we summarize important definitions and facts to obtain a self-contained presentation.
A key element in the symbol-pair coding theory is the representation of a symbol vector as a pair-vector.

Definition 1. *(Symbol-Pair Read Vector)*
Let $\boldsymbol{x} = [x_0, \ldots, x_{n-1}]$ be a vector in Ξ^n. The symbol-pair read vector of $\boldsymbol{x}$ is defined as

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

Every vector $\boldsymbol{x} \in \Xi^n$ has a pair representation $\pi(\boldsymbol{x}) \in (\Xi, \Xi)^n$. However, not all pair vectors in $(\Xi, \Xi)^n$ have a corresponding vector in Ξ^n, because they may have two different readings of the same symbol in two adjacent pairs. Pair-vectors that have corresponding symbol vectors are called *consistent*.
The pair-distance between two symbol vectors in Ξ^n is the Hamming distance between their corresponding pair-vectors (A pair (a, b) differs from (c, d) if $a \neq c$ or $b \neq d$, or both).
In a similar way to Hamming-metric codes, a code $\mathcal{C}$ can correct t pair-errors if and only if $d_p \geqslant 2t + 1$, where d_p is the minimum pair-distance between codewords of $\mathcal{C}$.
The pair-distance can be bounded as a function of the Hamming distance as follows (assuming $0 < D_H < n$)

$$D_H(\boldsymbol{x}, \boldsymbol{y}) + 1 \leqslant D_p(\boldsymbol{x}, \boldsymbol{y}) \leqslant 2D_H(\boldsymbol{x}, \boldsymbol{y}). \tag{1}$$

A more explicit relationship between the pair-distance and the Hamming distance is given by

$$D_p(\boldsymbol{x}, \boldsymbol{y}) = D_H(\boldsymbol{x}, \boldsymbol{y}) + L \tag{2}$$

where L is the number of consecutive runs in which the differing locations of $\boldsymbol{x}$ and $\boldsymbol{y}$ fall (a consecutive run may wrap around from $n - 1$ to 0).

III. Algebraic Cyclic-Code Constructions

Construction of pair-error codes was addressed in [2], where *interleaving* was proposed as a method to obtain codes with $d_p = 2d_H$, i.e. optimal pair-distance given Hamming distance. But despite this optimality, interleaved codes are in general inferior to directly constructed codes, even if the constituent Hamming-metric codes are themselves optimal. The problem with the interleaving approach is that it optimizes the pair-distance given the Hamming distance, with no attempt to optimize the Hamming distance itself (interleaved codes are known to have poor Hamming distance for their length). Therefore, in this section we take a more balanced approach of (more modestly) lower-bounding the pair-distance given the Hamming distance, but using codes that enjoy better Hamming distances to begin with: cyclic linear codes. New algebraic methods in the realm of the theory of cyclic codes will be sought as a framework for analysis and synthesis of pair-error correcting codes. For codes in the Hamming metric, the most powerful, flexible and practical codes in use are cyclic codes. So the purpose of the forthcoming discussion is to explore how the structure of cyclic codes can be exploited for pair-error correction as well.

We start with common definitions and notations for cyclic codes [8]. Let $g(x)$ be a polynomial of degree r over $\mathbb{F}_q$. If $g(x)|(x^n-1)$, then $g(x)$ defines a cyclic linear code $\mathcal{C}$ of length n with dimension $k = n-r$. The codewords of $\mathcal{C}$ are all polynomials that can be written as $c(x) = g(x)f(x)$, for some polynomial $f(x)$ over $\mathbb{F}_q$, where polynomial multiplication is carried out over $\mathbb{F}_q[x]/(x^n-1)$, the ring of q-ary polynomials modulo x^n-1. Let $\mathbb{F}_{q^t}$ be the splitting field of $\mathbb{F}_q$, i.e. the smallest field in which x^n-1 can be factored into linear factors. So $\mathbb{F}_{q^t}$ contains a primitive n^{th} root of unity α, such that $x^n-1 = \prod_{i=0}^{n-1}(x-\alpha^i)$. For a polynomial $c(x) \in \mathbb{F}_q[x]/(x^n-1)$, we define the j^{th} DFT (Discrete Fourier Transform) coefficient as $\hat{C}_j = c(\alpha^j)$. The BCH bound [1],[5], the most fundamental result in the theory of cyclic codes, can be formulated in terms of the DFT coefficients of codeword polynomials. We now include Proposition 1 as presented in [7]. Unless noted otherwise, polynomial and DFT indices are taken modulo n.

Proposition 1. *Let $c(x)$ be a polynomial in $\mathbb{F}_q[x]/(x^n-1)$ whose DFT satisfies $\hat{C}_{b+1} = \hat{C}_{b+2} = \cdots = \hat{C}_{b+\delta} = 0$, for some integer b. Then the number of non-zero coefficients in $c(x)$ is at least $\delta+1$.*

To prove lower bounds on the minimum pair-distance, a dual version of Proposition 1 will become useful.

Proposition 2. *Let $c(x)$ be a polynomial in $\mathbb{F}_q[x]/(x^n-1)$ whose coefficients satisfy $c_{b+1} = c_{b+2} = \cdots = c_{b+\delta} = 0$, for some integer b. Then the number of non-zero elements in the DFT sequence $\{\hat{C}_j\}_{j=0}^{n-1}$ is at least $\delta+1$.*

The next (simple) step is to transpose[1] Proposition 2 and get the following lemma.

[1]Logical transposition means change from ($a \Rightarrow b$) to (not $b \Rightarrow$ not a).

Lemma 3. *Let $\{\hat{C}_j\}_{j=0}^{n-1}$ be a DFT sequence with at least d zero elements. Then $c(x)$ does <u>not</u> have a set of $n-d$ consecutive coefficients such that $c_{b+1} = c_{b+2} = \cdots = c_{b+n-d} = 0$.*

Finally, an algebraic lower bound on the minimum pair-distance of a cyclic code is provided in the following theorem. We remark that the linearity of the codes allows considering the pair-weight, i.e. the pair-distance to the all-zero codeword, when proving results on the minimum pair-distance of the codes.

Theorem 4. *Let $g(x)$ be a generator polynomial of a cyclic code $\mathcal{C}$ with minimum Hamming distance d_H. If $g(x)$ has at least d_H roots in $\mathbb{F}_{q^t}$, then the minimum pair-distance of $\mathcal{C}$ is at least d_H+2.*

Proof: If $g(x)$ has at least d_H roots, then any codeword $c(x) = g(x)f(x)$ has a DFT sequence with at least d_H zeros. By Lemma 3, $c(x)$ does not have a set of $n-d_H$ consecutive zero coefficients. We distinguish two cases. If $c(x)$ has Hamming weight exactly d_H, then its non-zeros cannot fall into a single set of consecutive locations. This implies a pair-weight of at least d_H+2. Alternatively, if $c(x)$ has Hamming weight strictly larger than d_H, this also implies a pair-weight of at least d_H+2. ■

The importance of Theorem 4 is that it provides an algebraic d_H+2 lower bound on the pair-distance of a code that is strictly better than the combinatorial d_H+1 lower bound of (1) in section II. This algebraic lower bound applies to all linear cyclic codes that are not MDS (Maximum Distance Separable). The next example shows how this improved lower bound can prove that *cyclic* Hamming codes are "perfect" in the pair-metric as well.

Example 1. *Let α be a primitive element in $\mathbb{F}_{2^t}$, for some $t > 2$. With $n = 2^t-1$ we define $g(x) \in \mathbb{F}_2[x]/(x^n-1)$ to be the lowest degree polynomial that satisfies $g(\alpha) = 0$. It is well known that the length n cyclic code generated by $g(x)$ has roots $\alpha^1, \alpha^2, \alpha^{2^2}, \ldots, \alpha^{2^{t-1}}$, and by the BCH bound has minimum Hamming distance $d_H = 3$. This code is the length 2^t-1 binary Hamming code, constructed as a primitive narrow-sense BCH code with designed minimum Hamming distance 3 (all binary Hamming codes can be constructed in that form, but not all q-ary ones). We now turn to analyze the pair-error correction on the Hamming code generated by $g(x)$. For any $t > 2$, $g(x)$ has at least $d_H = 3$ roots. Hence by Theorem 4, it has minimum pair-distance $d_p \geqslant 5$. So cyclic Hamming codes of length $n \geqslant 7$ can correct 2 pair-errors. Correcting 2 pair-errors can be shown to be optimal by using the pair-sphere packing bound of Proposition 9 that appears later in the paper. Hence it follows that cyclic Hamming codes are perfect with $d_p = 5$.*

We note that the $d_p \geqslant 5$ bound obtained with Theorem 4 in Example 1 applies exclusively to Hamming codes that are cyclic. For example, there are equivalent ways to construct Hamming codes that yield (non-cyclic) codes with $d_p = 4$ (see "textbook" Hamming code in Figure 1). Clearly the code represented by the parity-check matrix in Figure 1 is equivalent to the code generated by $g(x)$ in Example 1 (identical up to reordering of the code coordinates). But the fact that the

two codes have different minimum pair-distances demonstrates the sensitivity of pair-error correctability to such coordinate reordering.

$$H_{[7,4]} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Figure 1. Parity check matrix of a $[7, 4]$ Hamming code. The word 0001110 is a codeword, thus the minimum pair-distance is only 4.

The algebraic structure of cyclic codes serves the theory of pair-error correction beyond the result of Theorem 4. Most conveniently, we can harness deeper (than the BCH bound) results on cyclic codes to obtain stronger bounds on the minimum pair-distance. This possibility is proved with the following theorem.

Theorem 5. *Let $g(x)$ be a generator polynomial of a cyclic code $\mathcal{C}$ with prime length n and minimum Hamming distance d_H. If $g(x)$ has at least m roots in $\mathbb{F}_{q^t}$, and $d_H \leqslant \min(2m - n + 2, m - 1)$, then the minimum pair-distance of $\mathcal{C}$ is at least $d_H + 3$.*

Proof: The main tool to prove this theorem is the use of the dual of the Hartmann-Tzeng bound [4], which generalizes the BCH bound to multiple sets of consecutive zeros (see also [6, Ch.4] for examples). We include a special case of the Hartmann-Tzeng bound in the following lemma.

Lemma 6. *(Hartmann-Tzeng Bound) Let $c(x)$ be a polynomial in $\mathbb{F}_q[x]/(x^n - 1)$ whose DFT satisfies $\hat{C}_{b+1} = \hat{C}_{b+2} = \cdots = \hat{C}_{b+\delta-1} = 0$, and $\hat{C}_{a+b+1} = \hat{C}_{a+b+2} = \cdots = \hat{C}_{a+b+\delta-1} = 0$ for some integers b, a, with $\gcd(a, n) < \delta$. Then the number of non-zero coefficients in $c(x)$ is at least $\delta + 1$.*

As we did with the BCH bound, a transposed version of the dual of Lemma 6 is stated in the following lemma (we skip the dual and move directly to the transposed dual).

Lemma 7. *Let $\{\hat{C}_j\}_{j=0}^{n-1}$ be a DFT sequence with at least m zero elements. Then $c(x)$ does <u>not</u> have two sets of $n - m - 1$ consecutive coefficients such that $c_{b+1} = c_{b+2} = \cdots = c_{b+n-m-1} = 0$ and $c_{a+b+1} = c_{a+b+2} = \cdots = c_{a+b+n-m-1} = 0$, for any b and any a with $\gcd(a, n) < n - m$.*

Now Lemma 7 implies that for a prime n, there are no two sets of $n - m - 1$ consecutive zero coefficients, for any spacing a between them (for n prime $\gcd(a, n) = 1$ for any $a < n$). Denote by $D_H(c)$ the number of non-zero coefficients in the polynomial $c(x)$. $D_p(c)$ will denote the pair-weight of the vector of coefficients of $c(x)$. If $c(x)$ does not have two sets of $n - m - 1$ consecutive zero coefficients, then at least one of the following is true:

1) The non-zero coefficients of $c(x)$ fall into 3 or more consecutive subsets.
2) $D_H(c) > n - 2(n - m - 1) = 2m - n + 2$.

In other words, if the non-zero coefficients of $c(x)$ fall into 2 consecutive subsets, then their number must be greater than $2m - n + 2$ to avoid two consecutive zero subsets of the forbidden size. Note that $d_H \leqslant m - 1$ excludes the possibility of a single consecutive set of non-zeros in $c(x)$, as proved in Theorem 4.

If option 1 is true, then the theorem trivially follows from (2). If option 2 is true, then $D_H(c) \geqslant d_H + 1$ from the condition $d_H \leqslant 2m - n + 2$. We distinguish two cases. If $D_H(c) = d_H + 1$, the condition $d_H \leqslant m - 1$ guarantees that $D_H(c) \leqslant m$, and by Theorem 4 $D_p(c) \geqslant D_H(c) + 2 \geqslant d_H + 3$. If $D_H(c) \geqslant d_H + 2$ then the bound $D_p(c) \geqslant d_H + 3$ follows trivially from the combinatorial relation $D_p \geqslant D_H + 1$. ■

More insight and guarantees on pair-error correction of cyclic codes may be gained by carefully analyzing subsequent improvements of the BCH bound that have appeared in the literature. Examples for these include the Roos lower bound [10] and the van Lint-Wilson bounding technique [11].

IV. Asymptotic Bounds on Pair-Error Codes

A. Review of combinatorial bounds

Before presenting the asymptotic bounds in the next subsection, we review combinatorial results from [2], upon which lies the asymptotic analysis.

Definition 2. *For a word $\boldsymbol{x} \in \Xi^n$, define the pair-sphere $\mathcal{S}_d(\boldsymbol{x})$ as the set of all $\boldsymbol{y} \in \Xi^n$ such that $D_p(\boldsymbol{x}, \boldsymbol{y}) = d$.*

The size of d-spheres is given in the following proposition.

Proposition 8. *[2] For any $\boldsymbol{x} \in \Xi^n$, and $d > 0$*

$$|\mathcal{S}_d(\boldsymbol{x})| = \sum_{l=\lceil d/2 \rceil}^{d-1} D(n, l, d - l)(q - 1)^l$$

where $q = |\Xi|$ is the size of the alphabet and

$$D(n, l, L) = \binom{l-1}{L-1}\left[\binom{n-l-1}{L} + 2\binom{n-l-1}{L-1}\right] + \binom{n-l-1}{L-1}\binom{l-1}{L}$$

The *pair-ball* $\mathcal{B}_h(\boldsymbol{x})$ consists of all words with pair-distance h or less from $\boldsymbol{x}$, and clearly

$$|\mathcal{B}_h(\boldsymbol{x})| = 1 + \sum_{i=1}^{h} |\mathcal{S}_i(\boldsymbol{x})|. \tag{3}$$

Proposition 9. *[2](Pair-Sphere Packing Bound) If $\mathcal{C} \subset \Xi^n$ is a code with M codewords that corrects all t-pair errors, then*

$$M|\mathcal{B}_t(\boldsymbol{x})| \leqslant q^n.$$

Proposition 10. *[2](Pair Gilbert-Varshamov Bound) There exists a code $\mathcal{C} \subset \Xi^n$ with M codewords and minimum pair-distance d if*

$$M|\mathcal{B}_{d-1}(\boldsymbol{x})| \leqslant q^n.$$

B. Asymptotic bounds

The combinatorial bounds of the previous sub-section use an exact enumeration of pair-spheres, and are thus a useful tool to bound the sizes of codes with given parameter sets. However, to get a general insight about the achievability and limits of coding in the pair-error model, an asymptotic analysis is needed. The main task toward an asymptotic analysis is to derive concise bounds on the sizes of pair-balls. Then the resulting simple expressions are used to bound the rates of codes with fractional minimum pair-distance $\gamma = d_p/n$ (as the code length n tends to infinity). Our goal is to obtain asymptotic bounds on the size of pair-balls that will be tight enough to show a non-vanishing rate advantage of coding in the pair scheme over coding in the Hamming scheme. We note that it is not a-priori clear that such an advantage exists. Examining the bound (1) $d_H + 1 \leqslant d_p \leqslant 2d_H$: if asymptotically good pair-codes have pair-distance at the low end closer to $d_H + 1$, then they are not likely to give any advantage over Hamming-metric codes; on the other extreme, if asymptotically good pair-codes have pair-distance at the high end closer to $2d_H$, then a significant advantage will emerge in favor of pair-codes. Thus the main purpose of the analysis below is to see whether asymptotically good pair-codes fall at the low end, high end, or somewhere in between (asymptotic advantage, but less dramatic than doubling the relative distance).

We start by obtaining a simple upper bound on $D(n,l,L)$ by the following inequality

$$\begin{aligned} D(n,l,L) &= \binom{l-1}{L-1}\binom{n-l-1}{L} + 2\binom{l-1}{L-1}\binom{n-l-1}{L-1} \\ &\quad + \binom{l-1}{L}\binom{n-l-1}{L-1} \\ &< 4\binom{l}{L}\binom{n-l}{L} \end{aligned} \tag{4}$$

the inequality follows from the basic binomial recursion that gives the following inequalities

$$\binom{a-1}{b} = \binom{a}{b} - \binom{a-1}{b-1} < \binom{a}{b}$$

and

$$\binom{a-1}{b-1} = \binom{a}{b} - \binom{a-1}{b} < \binom{a}{b}$$

(substitute $b = L$ and $a = l$ or $a = n - l$ to get (4)). As we proceed, we restrict ourselves to binary codes ($q = 2$), though derivation for general q is not substantially different. The size of a pair-sphere can now be bounded using (4)

$$|\mathcal{S}_h(\boldsymbol{x})| = \sum_{l=\lceil h/2\rceil}^{h-1} D(n,l,h-l) < 4\sum_{l=\lceil h/2\rceil}^{h-1}\binom{l}{h-l}\binom{n-l}{h-l}$$

Substituting the former into (3), we get

$$|\mathcal{B}_h(\boldsymbol{x})| = 1 + \sum_{i=1}^{h}|\mathcal{S}_i(\boldsymbol{x})| \leqslant 4\sum_{i=1}^{h}\sum_{l=\lceil i/2\rceil}^{i-1}\binom{l}{i-l}\binom{n-l}{i-l}$$

Taking $x = i - l$ we rewrite

$$|\mathcal{B}_h(\boldsymbol{x})| \leqslant 4\sum_{i=1}^{h}\sum_{x=1}^{\lfloor i/2\rfloor}\binom{i-x}{x}\binom{n-i+x}{x}$$

For given i and x, it is known that

$$\begin{aligned} \binom{i-x}{x}\binom{n-i+x}{x} &< 2^{(i-x)H(\frac{x}{i-x})+(n-i+x)H(\frac{x}{n-i+x})+o(n)} \\ &= 2^{n\left[(\delta-\xi)H(\frac{\xi}{\delta-\xi})+(1-\delta+\xi)H(\frac{\xi}{1-\delta+\xi})+o(1)\right]} \end{aligned}$$

where the last equality is obtained by taking $\delta = i/n$ and $\xi = x/n$. $H(\cdot)$ is the binary entropy function and an $o(f(n))$ function asymptotically tends to zero when divided by $f(n)$. Now it is clear that in the limit $n \to \infty$

$$\frac{1}{n}\log\left(4\sum_{i=1}^{h}\sum_{x=1}^{\lfloor i/2\rfloor}\binom{i-x}{x}\binom{n-i+x}{x}\right) \leqslant$$

$$\max_{0\leqslant\xi\leqslant\delta/2\ ,\ \delta<h/n}(\delta-\xi)H\left(\frac{\xi}{\delta-\xi}\right)+(1-\delta+\xi)H\left(\frac{\xi}{1-\delta+\xi}\right)$$

because the double sum is at most a quadratic factor in n larger than the maximal summand, a factor that vanishes after taking the logarithm and dividing by n (logarithms are base 2 throughout the section).

It now remains to maximize the exponent above to get an asymptotic upper bound on the size of pair-balls. We first assume that δ is fixed, and maximize ξ given δ. Denote

$$E(\delta,\xi) = (\delta-\xi)H\left(\frac{\xi}{\delta-\xi}\right) + (1-\delta+\xi)H\left(\frac{\xi}{1-\delta+\xi}\right)$$

Taking the derivative of E with respect to ξ we get

$$E'(\delta,\xi) = \log\left[\frac{(\delta-2\xi)^2(1-\delta+\xi)}{\xi^2(\delta-\xi)}\right]$$

and equating to zero yields the following cubic equation

$$5\xi^3 + \xi^2(4-9\delta) + \xi(-4\delta+5\delta^2) + \delta^2 - \delta^3 = 0$$

This equation can be solved in closed form, and it is found to have three real solutions, only one of which in the range $0 < \xi < \delta/2$ (we omit the unwieldy explicit expression). In Figure 2, we plot this solution as a curve on the (δ,ξ) plane, showing for each δ the ξ that maximizes the exponent $E(\delta,\xi)$. Contour lines of $E(\delta,\xi)$ are also expressed in the figure. It is also observed in Figure 2 that the maximum (over ξ) of $E(\delta,\xi)$ is monotone increasing in δ for $\delta < 0.75$, hence for $h/n < 0.75$

$$\max_{0\leqslant\xi\leqslant\delta/2\ ,\ \delta<h/n}E(\delta,\xi) = \max_{0\leqslant\xi\leqslant h/(2n)}E(h/n,\xi)$$

Now the asymptotic version of the pair Gilbert Varshamov (GV) bound from Proposition 10, with relative pair-distance $d/n = \gamma$ is given by

$$R(\gamma) > 1 - \max_{0\leqslant\xi\leqslant\gamma/2}E(\gamma,\xi)$$

where $R(\gamma)$ is the rate of the code. This bound is plotted (solid line) in Figure 3, alongside the traditional GV bound

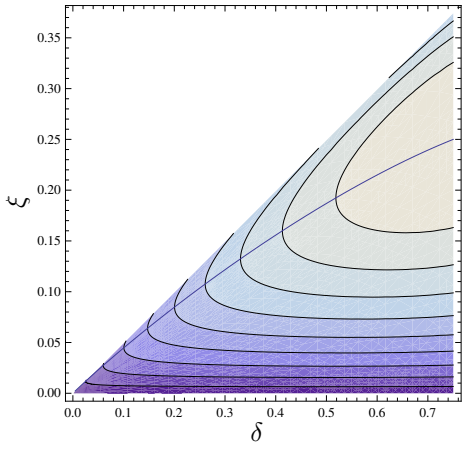

Figure 2. Curve of maximizing ξ given δ, overlaid on contours of $E(\delta, \xi)$.

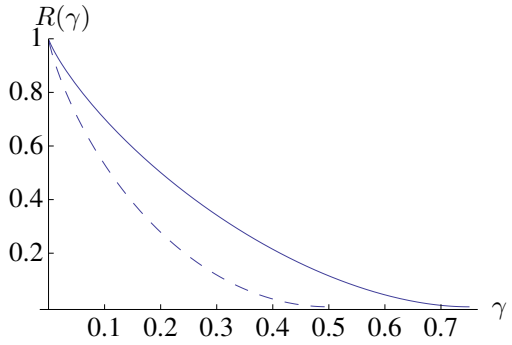

Figure 3. Comparison between the asymptotic Gilbert-Varshamov bounds for pair-distance (solid) and Hamming distance (dashed).

for Hamming-metric codes with the same relative distance (dashed).

The implication from Figure 3 is that codes for pair-errors are now *provably* known to exist with asymptotic rates that are strictly higher than provably achievable rates of codes for symbol-errors. It also shows existence of codes with relative pair-distance up to 0.75, in comparison to a cut-off at 0.5 for Hamming metric codes. Consequently, if, as widely conjectured, the binary Gilbert-Varshamov rate-bound for symbol-errors is tight [3], then Figure 3 proves an asymptotic gap in the correction capability between pair-errors and symbol-errors.

In a similar manner, we can derive an asymptotic sphere-packing upper bound for pair-errors. For the sake of space efficiency, we omit this part from the current presentation.

C. Pair-Plotkin bound

With the aid of the asymptotic GV bound of the previous sub-section, we know that pair-codes exist asymptotically with relative pair-distances up to 0.75. Now we want to find out whether 0.75 is a tight bound and a true cut-off value. Given the 0.5 relative-distance cut-off of Hamming metric codes, and the combinatorial bound $d_p \leqslant 2d_H$ of (1), this is not a-priori clear (in principle, this combinatorial bound does not exclude relative pair-distances up to $1 = 2{\cdot}0.5$). The way to answer this question is by deriving a Plotkin-type [9] bound for the pair metric. As it turns out, the Plotkin bound for binary pair-errors is identical to the standard $q = 4$ Plotkin bound for Hamming-metric errors (in general standard Plotkin bounds can be used with squared alphabet sizes). To use standard Plotkin bounds for pair-distance codes, we use the following proposition.

Proposition 11. *If there exists a binary code $\mathcal{C}$ with length n and minimum pair-distance d, then there exists an equal size $q = 4$ code with the same length n, and minimum Hamming distance d.*

Proof: We take the pair vectors of the codewords of $\mathcal{C}$ and map $(0,0) \to 0$, $(0,1) \to 1$, $(1,0) \to 2$, $(1,1) \to 3$. Clearly the resulting code over $q = 4$ has the same length, and minimum Hamming distance d. ■

Proposition 11 implies that an upper bound on $q = 4$ Hamming-metric codes is also an upper bound on binary pair-metric codes. Consequently, the pair-Plotkin bound will be given as a reformulation of the standard $q = 4$ Plotkin bound (found e.g. in [6]).

Proposition 12. *Let $\gamma = d/n$ be the relative pair-distance of a binary code. Then the rate of the code is bounded by*

$$R(\gamma) \leqslant 1 - \frac{4}{3}\gamma.$$

Therefore 0.75 is proved to be the true relative pair-distance cut-off for the asymptotic existence of binary pair-error correcting codes.

We remark that while the reduction from pair-distance binary codes to $q = 4$ Hamming-distance codes works well for the Plotkin bound, in general it has limited effectiveness for pair-metric bounds. This is because the reverse reduction is not possible[2], resulting in a likely significant overestimation of pair-distance code sizes.

REFERENCES

[1] R. Bose and D. Ray-Chaudhuri, "On a class of error correcting binary group codes," *Information and Control*, vol. 3, pp. 68–79, 1960.
[2] Y. Cassuto and M. Blaum, "Codes for symbol-pair read channels," in *Proc. of the IEEE International Symposium on Info. Theory*, Austin, Texas, June 2010.
[3] V. Goppa, "Bounds for codes," *Doklady Akademii Nauk SSSR*, vol. 333, p. 423, 1993.
[4] C. Hartmann and K. Tzeng, "Generalizations of the BCH bound," *Information and Control*, vol. 20, pp. 489–498, 1972.
[5] A. Hocquenghem, "Codes correcteurs d'erreurs," *Chiffres*, vol. 2, pp. 147–156, 1959.
[6] W. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*. Cambridge, UK: Cambridge university press, 2003.
[7] R. J. McEliece, *The Theory of Information and Coding*. Cambridge University Press, 2002.
[8] W. Peterson and E. Weldon, *Error-Correcting Codes*. Cambridge, MA: MIT Press, 1972.
[9] M. Plotkin, "Binary codes with specified minimum distances," *IRE Transactions on Information Theory*, vol. 6, pp. 445–450, 1960.
[10] C. Roos, "A new lower bound for the minimum distance of a cyclic code," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 330–332, 1983.
[11] J. van Lint and R. Wilson, "On the minimum distance of cyclic codes," *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 23–40, 1986.

[2]due to non-consistent pair-vectors