Asymptotically MDS Array BP-XOR Codes

Şuayb Ş. Arslan
Department of Computer Engineering
MEF University
Maslak, Istanbul 34099
Email: arslans@mef.edu.tr

Abstract—**Belief propagation or message passing on binary erasure channels (BEC) is a low complexity decoding algorithm that allows the recovery of message symbols based on bipartite graph prunning process. Recently, array XOR codes have attracted attention for storage systems due to their burst error recovery performance and easy arithmetic based on Exclusive OR (XOR)-only logic operations. Array BP-XOR codes are a subclass of array XOR codes that can be decoded using BP under BEC. Requiring the capability of BP-decodability in addition to Maximum Distance Separability (MDS) constraint on the code construction process is observed to put an upper bound on the maximum achievable code block length, which leads to the code construction process to become a harder problem. In this study, we introduce asymptotically MDS array BP-XOR codes that are alternative to exact MDS array BP-XOR codes to pave the way for easier code constructions while keeping the decoding complexity low with an asymptotically vanishing coding overhead. We finally provide and analyze a simple code construction method that is based on discrete geometry to fulfill the requirements of the class of asymptotically MDS array BP-XOR codes.**

I. Introduction

Array codes are linear codes defined for two dimensional data structures that are defined by both data and parity values organized in a matrix form. These codes are quite attractive candidates for burst error recovery in communication and distributed storage systems [1] and provide data reliability with optimal time/space consumption using Maximum Distance Separability (MDS) constraint in the code construction process. Moreover, a great deal of work has been done and many improvements have been proposed for these codes over the years [2] to secure simpler math and low-complexity computations while still maintain the MDS property.

Typically, any linear code can be represented using a bipartite graph either using the parity check matrix or the generator matrix of the code [3]. Using the generator matrix representation, the corresponding bipartite graph has two types of nodes: Nodes that are used to decode (check or coded nodes) and nodes that are decoded (information nodes). Nodes in bipartite graph representation are connected with edges to represent node adjacency. The neighbors of node j (neighbor set), denoted by $\mathcal{N}_j$, is the set of all nodes connected to node j. The cardinality of the neighbour set is called the *degree* of node j. The Belief Propagation (BP) algorithm a.k.a. message passing algorithm is an iterative process (updating nodes and edges) to decode data from coded nodes over symmetric erasure channels using the bipartite representation of the code. At the onset of the BP algorithm, we begin by setting all the contents of information nodes to NULL that need to be decoded. Then, we look for a degree-one coded node and copy the content to its neighbor information node by replacing NULL. Next, we update all the coded nodes that are connected to the this neighbor and eliminate the edges that established neighborhood relationship. This completes the first step, and in the next iteration we continue applying the same methodology until there remains no information node with NULL content. If algorithm stops prematurely during iteration, we claim a decoding failure, otherwise we report a decoding success.

Array codes have recently been studied under BP decoding [9] and useful upper bounds are derived in [6] that theoretically establishes the relationship between the block length (and hence the rate of the code), decodability and sparsity of the generator matrix i.e., the encoding/decoding complexity of the code. In this study, we shall demonstrate by relaxing the MDS constraint on the code construction process, we can also dramatically relax the previously found bounds on the code block length [6] while keeping low complexity BP algorithm successfully decode the whole data block. Such an observation shall yield easier and more powerful code constructions. For instance, we shall consider one of the discrete geometry based codes known as Mojette codes that are recently studied within the context of low density parity check codes and are shown to reduce the node repair complexity [10]. In our study, we demonstrate an asymptotically MDS BP-XOR code construction method based on Mojette geometry. By providing and establishing an appropriate set of code parameters, we explicitly construct codes that fulfills the desired theoretical requirements.

The rest of the paper is organized as follows. In Section II, we provide the basics of array MDS BP-XOR codes and give some known results as well as state the main result of the paper. In Section III, we provide a discrete geometry construction of an asymptotically-MDS array BP-XOR codes. In Section IV, we validate our theoretical results by numerically plotting rate, code block length for discrete geometry construction. Finally, we conclude our paper in Section V.

II. Asymptotically MDS Array BP-XOR Codes

Before defining the class of asymptotically MDS array BP-XOR codes, let us provide the conventional definition of MDS BP-XOR codes using the notation of reference [6].

A. Background

Let l be the symbol size in bits and $M = \{0,1\}^l$ be the symbol set from which we select our information as well as coded symbols. The fundamental operation we use is the Exclusive OR (XOR) that is used to add symbols logically bit by bit in binary domain. In our study, nodes represent blocks of data that contains one or more symbols in it. Symbols are the smallest data unit over which XOR operations are defined.

An $[n,k,t,b]$ array BP-XOR code is a $b \times n$ two dimensional rate $r = k/n$ binary linear code $\mathcal{C} = [a_{i,j}]_{1 \leq i \leq b, 1 \leq j \leq n}$ in which the coding symbol $a_{i,j} \in M$ is the XOR of a subset of source symbols $I = \{v_1, \dots, v_{bk}\}$, typically structured as a $b \times k$ data matrix, and I can be reconstructed from any $n-t$ columns of the linear code $\mathcal{C}$ using BP algorithm for an appropriate integer $t \leq n-k$. The degree of a coded symbol $a_{i,j}$, denoted as $\sigma_{i,j}$, is the number of information symbols that participate in logical XOR operation i.e., $a_{i,j} = v_{z_1} \oplus \dots \oplus v_{z_{\sigma_{i,j}}}$ such that $v_{z_s} \in I$ for all $s \in \{1, \dots, \sigma_{i,j}\}$. A t-erasure correcting array BP-XOR code is MDS if the source symbols can be reconstructed from $k = n-t$ columns of $\mathcal{C}$.

For a given positive integer b' satisfying $b' > b$, a $[n,k,t,b,b']$ asymptotically MDS array BP-XOR code $\mathcal{C}^a$ is a linear code with i-th column $(y_{i,1}, \dots, y_{i,b_i}) = (x_1, \dots, x_{bk})G_i$ for a $bk \times b_i$ generator matrix $G_i, i \in \{1, \dots, n\}$ such that $b' = (1/n)\sum_i b_i$. Thus, the generator matrix for $\mathcal{C}^a$ is given by the $bk \times \sum_i b_i$ matrix,

$$G_{\mathcal{C}^a} = [G_1|G_2|\dots|G_n]. \tag{1}$$

What makes this code asymptotically MDS is that it is possible to perfectly reconstruct user data matrix I from any k column combinations of $\mathcal{C}^a$ using BP decoding and as $b \rightarrow \infty$ we have $b' \rightarrow b$. Note that the raw source data need not be in standard $b \times k$ form. For any positive integer g satisfying $b|g$ and $k|g$, the generator matrix $G_{\mathcal{C}^a}$ should work fine for different arrangements of the data block matrix such as $b/g \times kg$. We finally note that the code $\mathcal{C}^a$ is not in two dimensional standard rectangle form as in $\mathcal{C}$. However, we introduced another parameter b' to be able to make asymptotically MDS array BP-XOR codes analogous to standard MDS array codes defined over rectangle shape binary matrices.

For a given fixed code rate r and n, let us define $\epsilon(b,n)$ to be the maximum coding overhead[1] of $\mathcal{C}^a$ satisfying $b' = (1+\epsilon(b,n))b$. The asymptotically optimal overhead property implies that as $\epsilon(b,n) \rightarrow 0$ we have $b \rightarrow \infty$.

Letting σ denote the maximum check node degree of a given array BP-XOR code, we note from [6] that if $k = \sigma$ it is not hard to show that

$$n \quad \leq \quad kb + 1 + \max\{k-3, 0\} \tag{2}$$

the upper bound of which can be arbitrarily large (i.e., for $b \gg 1$) and allow any arbitrarily small r to be possible. However, for $k > \sigma$ it is observed that the array code blocklength n is upper bounded based on a specific choice of k [6]. In addition, we observe from the same study that for $b \gg 1$ and large enough k i.e., $k > \sigma^2$ we have $n \leq k + \sigma - 1$. This also implies that for large enough information block length k, the achievable rate will be close to 1, putting a constraint on the code design rate.

B. Main Result

We begin with providing the following theorem that sets the necessary condition/s on the parameters for the existence of asymptotically MDS array BP-XOR codes.

Theorem 2.1. *Let $\mathcal{C}^a$ be a $[n,k,t,b,b']$ asymptotically MDS array BP-XOR code such that the maximum coded node degree satisfies $2 < \sigma < (bk-1)/(b'-1)$. Then, we have*

$$\begin{aligned} n \quad \leq \quad & k + \sigma - 1 + \\ & \left\lfloor \frac{b(k(\sigma' - \sigma) + (\sigma-1)\sigma') - (\sigma-1)(3\sigma/2 - 1)}{b(k-\sigma') + \sigma - 1} \right\rfloor \end{aligned} \tag{3}$$

where $\sigma' = \sigma(1 + \epsilon(b,n))$ and $\epsilon(b,n)$ is the coding overhead.

Proof. Since the code is assumed to be MDS, i.e., able to tolerate $n-k$ column erasures of $\mathcal{C}^a$, each information symbol $v_s \in I$ must appear in at least $n-k+1$ columns, totaling up to

$$kb(n-k+1) \tag{4}$$

minimum appearances in $\mathcal{C}^a$. On the other hand, belief propagation decoding starts decoding from degree-one encoding symbols. So we need at least $n-k+1$ degree-one symbols in distinct columns of $\mathcal{C}^a$ (in the worst case of $n-k$ column erasures when each may comprise one degree-one symbol). Similarly, we need at least one degree-two, one degree-three, $\dots$, one degree-$(\sigma-1)$ coding symbols to make sure that BP decoding continues. Although it is possible to have multiple degree-two symbols and continue BP decoding, by this choice we are trying to maximize the appearance of information symbols in $\mathcal{C}^a$. Note that if these symbols happen to be in distinct unerased columns, the bound could be tightened, otherwise the bound might still be loose for instance if $\sigma > k+1$ which is not usually typical. The rest of the $b'n - (n-k+\sigma-1)$ can have at most σ degree. Thus, C^a can have at most

$$\sigma(b'n - (n-k+\sigma-1)) + n - k + \frac{\sigma(\sigma-1)}{2} \tag{5}$$

appearances of kb information symbols. So we have the inequality (4) $\leq$ (5). We can rewrite (5) in a more compact form as

$$\sigma b'n - (\sigma-1)(n-k+\sigma/2) \tag{6}$$

[1] Since columns of $\mathcal{C}^a$ may have different sizes, the overhead depends on which k columns are used for reconstruction. Also note that the coding overhead also depends on the number of columns n in the code, so called array code blocklength.

Using equation (4), and assuming we have $b(k-\sigma')+\sigma-1>0$, we can collect all terms that includes n and find an upper bound on n as follows,

$$n \le \left\lfloor \frac{(kb+\sigma-1)(k-1)-(\sigma-1)(\sigma/2-1)}{b(k-\sigma')+\sigma-1} \right\rfloor \tag{7}$$

$$= k+\sigma-1+ \tag{8}$$

$$\left\lfloor \frac{b(k(\sigma'-\sigma)+(\sigma-1)\sigma')-(\sigma-1)(3\sigma/2-1)}{b(k-\sigma')+\sigma-1} \right\rfloor$$

where $\sigma' = \sigma(1+\epsilon(b,n))$. ■

Note that if $b \to \infty$ we will have $\sigma' \to \sigma$ and hence equation (7) becomes identical to equation (2) of [6] except the term $(\sigma-1)(\sigma/2-1)$. This term is essentially what makes the upper bound improved (tighter).

There are two cases that are interesting to consider for understanding the asymptotical performance. First, if b tends large we will have $\sigma' \to \sigma$. Hence,

$$n \le k+\sigma-1+\left\lfloor \frac{(\sigma-1)\sigma}{k-\sigma} \right\rfloor - \mathbf{1}_{(k-\sigma)|(\sigma-1)\sigma}$$

where $\mathbf{1}_A$ is logical one if A is true, otherwise it is zero. This indicator function is used due to the flooring operation and σ only equals to σ' in the limit. Thus, if the code becomes array MDS in the limit, there remains no dependence of n on b. On the otherhand, if we let large but fixed $b \le k$, and if k gets large, we shall have

$$\begin{aligned} n &\le k+\sigma'-1 \\ &= k+\sigma(1+\epsilon(b,n))-1 \end{aligned} \tag{9}$$

which can be made arbitrarily large if we choose $\epsilon(b,n) \to \infty$ for a fixed b and large n. This essentially demonstrates that as the array BP-XOR code becomes near-optimal in terms of recovery performance, the upper bound on the number of code columns n can dramatically be improved.

Although the desirable properties of the coding overhead are found, we still need specific constructions to quantify or bound the coding overhead and hence present tighter bounds on n (and r) for a specific construction. Based on this observation, we shall present a code construction method that uses the result of Theorem 2.1 and has an appropriate $\epsilon(b,n)$ with the properties as summarized below.

- For fixed k and rate r (i.e., fixed n), as $b \to \infty$ we have vanishing coding overhead, $\epsilon(b,n) \to 0$.
- For fixed b and rate r, as $n \to \infty$ we have a diverging coding overhead, $\epsilon(b,n) \to \infty$.

III. Discrete Geometry Constructions of Asymptotically-MDS array BP-XOR codes

In this section, we will introduce a particular construction of asymptotically MDS array BP-XOR codes based on discrete geometry [5] and show that they can be regarded as a special type of the class of asymptotically MDS BP-XOR codes.

The discrete geometry construction is known as Mojette codes which are based on discrete version of Radon Transform [4], and can be used to generate redundancy not just for rectangle two dimensional data grid but also for any convex data grid. In our study, we consider matrix (rectangle) data and let encoder compute a linear set of projections at angles specified by a couple of coprime integers (p,q) from a $b \times k$ discrete data structure $f:(z,l) \to \mathbb{N}$. Suppose that we generate n projections with parameters $\{(p_i,q_i), 0 \le i \le n-1\}$. The length of the projection i, denoted by b_i, is a function of the number of projections n, the angle parameters (p_i,q_i) and the data grid size $b \times k$. It can be expressed in a closed form as follows [5],

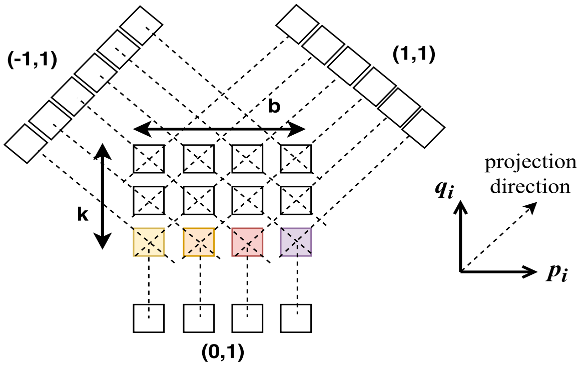

Fig. 1. A simple illustration of the projection concept and Mojette coding.

$$b_i = |p_i|(k-1)+|q_i|(b-1)+1 \tag{10}$$

Note that in this construction, generated projections can be treated as the columns of the asymptotically-MDS BP-XOR code. An example code with parameters $k=3, b=4$ with $n=3$ projections with parameters $(-1,1),(1,0),(1,1)$ is shown in Fig. 1. Each bin or symbol of the i-th projection, based on (p_i,q_i), can be computed as given by the following compact formulation

$$M_{(p_i,q_i)} f(m+(b-1)q_i u(q_i)+(k-1)p_i u(p_i)) \tag{11}$$

$$= \bigoplus_{z=0}^{b-1} \bigoplus_{l=0}^{k-1} f(z,l)\delta_{m+zq_i+lp_i} \tag{12}$$

for all m values satisfying the inequality,

$$-(b-1)q_i u(q_i)-(k-1)p_i u(p_i)$$
$$\le m \le$$
$$b_i-(b-1)q_i u(q_i)-(k-1)p_i u(p_i)-1$$

where $\bigoplus$ stands for Boolean XOR operation, $u(.)$ is the discrete unit function and δ_i is Kronecker delta function which are given by

$$u(s) = \begin{cases} 1, & \text{if } s>0 \\ 0, & \text{Otherwise} \end{cases}, \quad \delta_i = \begin{cases} 0, & \text{if } i \ne 0 \\ 1, & \text{if } i=0 \end{cases}$$

Mojette codes can be decoded using BP algorithm and the exact reconstruction of user data matrix is possible if the projection parameters (p_i,q_i) are selected judiciously according to the following Katz criterion.

Theorem 3.1. *For a given asymptotically-MDS BP-XOR code defined by n projections with parameters (p_i, q_i) on a $b \times k$ data matrix, exact data reconstruction is possible using iterative BP if*

$$\sum_{i=0}^{n-1} |p_i| \geq b \textit{ or } \sum_{i=0}^{n-1} |q_i| \geq k \tag{13}$$

Proof. The proof can be found in [7]. ■

According to Theorem 2.1, the maximum degree of the coded symbols play key role in the attainable block length of the BP-XOR codes. Thus, next we find the maximum degree number in the case of Mojette transform codes and see that this parameter can be adjusted based on the selection of projection parameters (p_i, q_i). The following theorem quantifies this number.

Theorem 3.2. *Let us use $\sigma_i, i \in \{1, 2, \ldots, n\}$ to denote the maximum degree of the ith projection with parameters (p_i, q_i). We have $\sigma_i = \min\{\lceil b/|p_i| \rceil, \lceil k/|q_i| \rceil\}$ and hence $\sigma = \max_i\{\sigma_i\}$.*

Proof. Considering the equation (12) and the worst case scenario, we would like to find the number of l and z values such that $zq_i + lp_i = -m$. It is not hard to see that the maximum number of z values that can satisfy this equation is given by $\lceil k/|q_i| \rceil$ due to $0 \leq z \leq k-1$. Similarly, the maximum number of l values that can satisfy this equation is given by $\lceil b/|p_i| \rceil$ due to $0 \leq l \leq b-1$. Since the number of possibilities for z and l are also constrained by the two dimensional rectangular shape, we have the maximum encoding symbol degree equal to the minimum of the two i.e., $\sigma_i = \min\{\lceil b/|p_i| \rceil, \lceil k/|q_i| \rceil\}$. Thus, the maximum degree of all the code symbols is given by the maximum degree of all the projections i.e., $\sigma = \max_i\{\min\{\lceil b/|p_i| \rceil, \lceil k/|q_i| \rceil\}\}$. ■

Next, we quantify the coding overhead for Mojette transform based asymptotically MDS BP-XOR codes by considering $k = \sigma$ and $k > \sigma$ cases separately.

A. Case $k = \sigma$

First of all, note that depending on the choices of (p_i, q_i), the code overhead as well as the maximum degree of the code can change. Although, there are multiple choices for $k = \sigma$, we provide the typical choice below that also ensures block length.

Construction 3.3. *Let us consider the following choice of coprime integers,*

$$q_i = 1, p_i \in \mathfrak{T} = \left\{-\left\lfloor \frac{n-1}{2} \right\rfloor, \ldots, -1, 0, 1, 2, \ldots, \left\lceil \frac{n-1}{2} \right\rceil \right\} \tag{14}$$

where $\mathfrak{T}$ is known as canonical enumeration of integers [8] that goes with the name A007306 *and satisfies $gcd(p_i, q_i) = 1$ for $i = 0, \ldots, n-1$.*

Note that this construction satisfies the Katz criterion simply because collecting any k projections will lead us to have $\sum |q_i| = k$. If we use the coprime integers as given by the Construction 3.3, we have q_i never equal to zero and $\sigma_i = \min\{\lceil b/\lceil (n-1)/2 \rceil, k\}$. We note that we have $\sigma = k$ for $b \gg 1$. We next quantify the coding overhead for this particular construction and show the asymptotically optimal property.

Theorem 3.4. *For the Mojette code with parameters as given in Construction 3.3, for $b \gg 1$, we have*

$$\epsilon(b, n) \approx \frac{n(2-r)(nr-1)}{4b} \tag{15}$$

where r is the fixed rate of the array BP-XOR code.

Proof. See appendix A the proof of this theorem. ■

For fixed r and k (i.e., fixed n), if $b \to \infty$ then it is clear that $\epsilon(b, n) \to 0$ proving the asymptotical property. On the other hand, for fixed r and b, if $n \to \infty$ then we have $\epsilon(b, n) \to \infty$. In fact, it is not hard to see that $\epsilon(b, n) = O(n^2)$. Therefore, due to these desirable properties of the overhead and considering the inequality (9), we can make n arbitrarily large. Particularly we can find the following lower bound on n for $k = rn = \sigma$ and $r > 0.5$,

$$n \leq rn + rn\left(1 + \frac{n(2-r)(nr-1)}{4b}\right) - 1 \tag{16}$$

which yields the inequality

$$n - 2nr \leq \frac{n^3 r^2 (2-r)}{4b} \Rightarrow n \geq \sqrt{\frac{4b(1-2r)}{r^2(2-r)}} \tag{17}$$

This final lower bound shows that the value for the block length n can be arbitrarily large for judiciously selected large b. Note that the case $k = \sigma$ has the least constraint on the code block length for any MDS array BP-XOR code. The case $k > \sigma$ is more interesting for the class of asymptotically MDS array BP-XOR codes.

B. Case $k > \sigma$

With classical array BP-XOR codes, the block length n is constrained by the following upper bound for $b \gg 1$,

$$n \leq k + \sigma - 1 + \left\lfloor \frac{\sigma(\sigma-1)}{k-\sigma} \right\rfloor - \mathbf{1}_{(k-\sigma)|(\sigma-1)\sigma} \tag{18}$$

which is the same for asymptotically MDS array BP-XOR codes as mentioned in Section II. However, as the block length gets large as well, we shall no longer have constraints on the size of the block length for asyptotically MDS BP-XOR codes.

Next, we provide another set of parameters for Mojette code that shall satisfy $k > \sigma$. The possibilities of the pair (p_i, q_i) selection for making $k > \sigma$ is not unique. We will consider the typical class as given in construction 3.5.

Construction 3.5. *Let us consider the following choice of coprime integers for n projections,*

$$q_i = q_e > 0,$$
$$p_i \in \mathfrak{U} = \{\lceil -n+1 \rceil_{odd}, \ldots, -1, 1, 3, \ldots, \lceil n-1 \rceil_{odd}\} \tag{19}$$

where q_e is a positive even number, and $\lceil . \rceil_{odd}$ rounds to the next biggest odd integer of the argument, respectively.

Note that using construction 3.5, it is easy to verify that we have $GCD(p_i, q_i) = 1$. Also, we have $k > \sigma = \max_i\{\min\{\lceil b/|p_i| \rceil, \lceil k/|q_i| \rceil\}\} = \lceil k/q_e \rceil$. It is of interest to quantify the coding overhead to be able to find the upper bounds on the code block length.

Theorem 3.6. *For the Mojette code with parameters as given in construction 3.5, for $b \gg 1$, we have*

$$\epsilon(n,b) \approx \tag{20}$$
$$\frac{\lceil k/q_e \rceil}{kb}\left((k-1)\left(n - \frac{\lceil k/q_e \rceil}{2}\right) + (b-1)q_e + 1\right) - 1$$

where q_e is a positive even number, and $\lceil . \rceil_{odd}$ rounds to the next biggest odd integer of the argument, respectively.

Proof. See appendix B for the proof of this theorem. ■

Note that as long as $q_e|k$, we have $\epsilon \to 0$ for large b demonstrating the asymptotically optimal overhead property. Similarly, for fixed r and b, if $n \to \infty$ then we have $\epsilon(n,b) \to \infty$ satisfying the second desirable property.

Finally, using equation (9) we can express the upper bound on n as follows,

$$n \leq k + \frac{\sigma\lceil k/q_e \rceil}{kb}\left((k-1)\left(n - \frac{\lceil k/q_e \rceil}{2}\right) + (b-1)q_e + 1\right) - 1 \tag{21}$$

Since it is hard to see that with this result we improve the upper bounds on the code block length, in the next section, we provide some numerical results that compute the upper bounds for comparison.

IV. Numerical Results

Let us consider $q_e = 2$ and a large b value, such as $b = 10000$ (this choice is completely arbitrary) and compare the upper bounds on n with using classical MDS array BP-XOR codes and their asymptotically optimal version proposed in our study, abbreviated as AMDS. We present results in Fig.1, Fig. 2 and Fig. 3 each corresponding to three different rates $5/6, 3/4, 1/2$, respectively as example use cases. These results demonstrate that as the code rate decreases, classical MDS array BP-XOR codes are only possible for very small values of k. On the other hand, although the same is true for asymptotically MDS BP-XOR codes for small k, it is also observed that for large enough k our bounds are bigger than the required n (fixed by the code rate), allowing possible constructions to achieve the corresponding rate asymptotically MDS array BP-XOR code such as Mojette construction we have provided in previous sections. These figures also present the upper bound behavior for small k on the left corner of each plot. Plots include a curve "Required n" to denote the required value for n for the corresponding rate $r = k/n$ code.

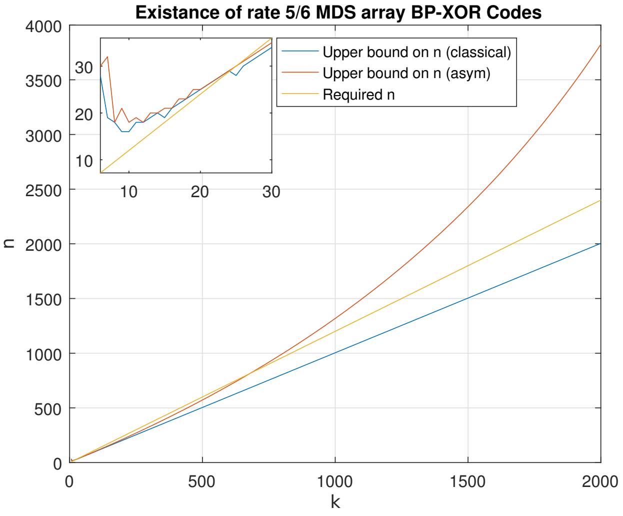

Fig. 2. Upper bounds on n as a function of k for $b = 10000$.

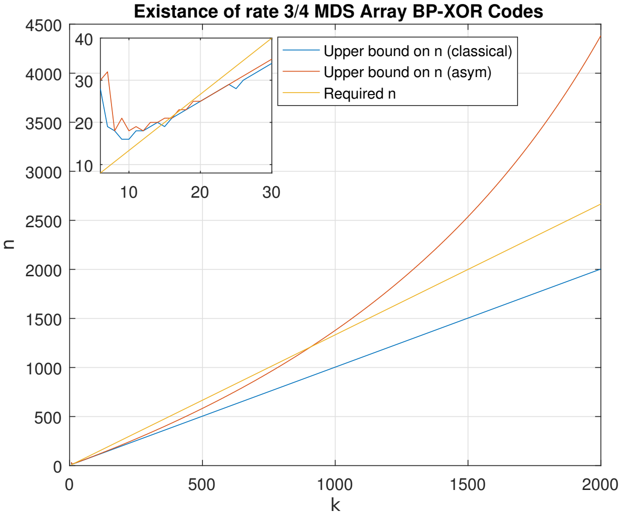

Fig. 3. Upper bounds on n as a function of k for $b = 10000$.

In order to see clearly the range of rates that are possible with both constructions, Fig. 5 depicts the minimum rate that is possible as a function of the assumed rate. Note that with asymtotically MDS array BP-XOR codes, the upper bound on n depends on the coding overhead which is a function of rate of the code. Thus, the minimum code rate changes as the assumed code rate changes. For each assumed rate, we calculate the upper bound and then compute the minimum code rate possible. With respect to classical MDS BP-XOR codes, since the upper bound does not change with varying assumed rate (since the coding overhead is always zero), the curves turns out to be flat.

According to Fig. 5, the region that lies above the curves

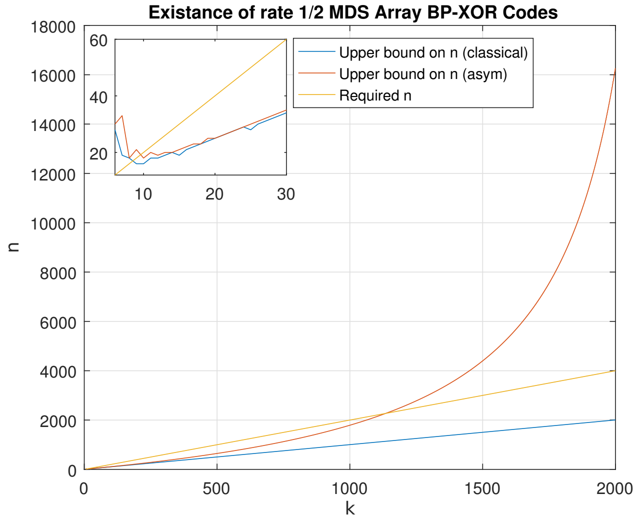

Fig. 4. Upper bounds on n as a function of k for $b = 10000$.

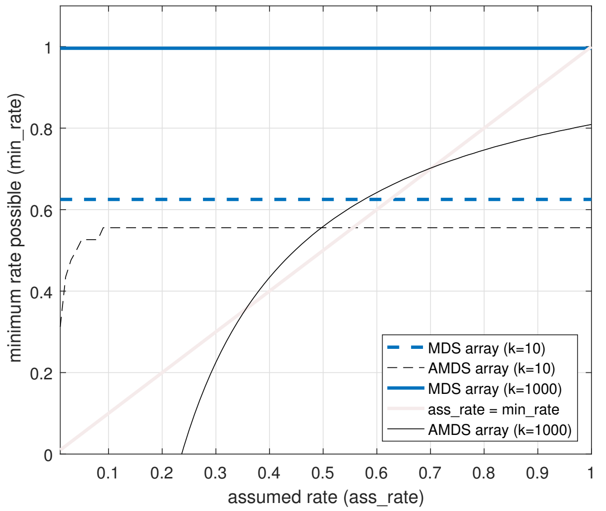

Fig. 5. Upper bounds on n as a function of k for $b = 10000$.

are the possibilities of the code rate. However, there is no guarantee each and every assumed rate would be achievable. However, as can be seen as k gets large it becomes impossible to construct classical MDS array BP-XOR codes with rate smaller than 1. In contrast, by relaxing the exact MDS condition (such as adapting asymptotically MDS constructions), we can improve the the region of possibilities for better achievability. With this study, we have just provided one simple construction based on discrete geometry (with judicious selection of parameters) that helps improve the upper bounds on the code block length n. Other constructions may help improve the results presented in this subsection.

V. Conclusion

Array BP-XOR codes are attractive data protection schemes for low-complexity and optimal reliability. Their finite versions are shown to have limitations on the maximum block length when the coding symbol degree is particularly lower than the data size. We have shown in this study, this limitation can greatly be relaxed by extending the original optimal class to asymptotically optimal class. We have also have shown one particular code construction based on discrete geometry that satisfies all the requirements of being asymptotically MDS array BP-XOR codes. These codes can be encoded and decoded in linear time with the block length and the achievable bound on the block length is far from that of the finite counterpart.

Appendix A
Proof of Theorem 2.4

Let us start by defining the following utility function,

$$\varphi(x) = \left\lfloor \frac{x}{2} \right\rfloor \left(\left\lfloor \frac{x}{2} \right\rfloor + 1 \right) \text{ for } x \ \geq 0. \tag{22}$$

Also let $I_t = \{0, 1, \dots, t-1\}$. Using these definitions, we state the following lemma next.

Lemma A.1: For the projection set given as in (14), we have the sum $\sum_{i=0}^{t-1} |p_i|$ that can be expressed in a closed form using the utility function

$$\sum_{i \in I_t} |p_i| = \frac{1}{2}\left(\varphi(t) + \varphi(t-1)\right) = \begin{cases} \frac{t^2-1}{4}, & \text{if } t \text{ is odd} \\ \frac{t^2}{4}, & \text{if } t \text{ is even} \end{cases}$$

This lemma can easily be proved by considering t odd and even cases using induction, separately. Note that the integer sequence $\sum_{i \in I_t} |p_i|$ is given by *A002620* [8]. Using this result, for a given pair of projections t_2 and t_1 satisfying $t_2 > t_1$, with the associated projection parameters $(p_i^{(t_2)}, q_i^{(t_2)} = 1)$ and $(p_i^{(t_1)}, q_i^{(t_1)} = 1)$ selected based on construction 2.2 (14), we can deduce that

$$\frac{t_2^2 - t_1^2 - 1}{4} \leq \sum_{i=0}^{t_2-1} |p_i^{(t_2)}| - \sum_{j=0}^{t_1-1} |p_j^{(t_1)}| \leq \frac{t_2^2 - t_1^2 + 1}{4} \tag{23}$$

Note that since $q_i = 1$, it is sufficient to collect k projections for perfect reconstruction. Thus, the upper/lower bounds given in equation (23) are particularly useful if we set $t_2 = n$ and $t_1 = n - k$ to be able find the contributions from the largest k projections in the sum that appears in the worst case coding overhead expression. Let i' be the index such that $p_{i'}^{(t_2)} = p_0^{(t_1)}$ and define the set

$$S = \{i', i'+1, \dots, i'+n-k-1\} \tag{24}$$

The worst case coding overhead in this case is given by the following

$$\epsilon(n,b) = \frac{1}{kb}\left(\sum_{i \in I_t \setminus S} |p_i|(k-1) + |q_i|(b-1) + k\right) - 1 \quad (25)$$

$$= \frac{(b-1)k + \frac{k-1}{2}\left(\varphi(n) + \varphi(n-1)\right)}{kb} + \frac{-\frac{k-1}{2}\left(\varphi(n-k) + \varphi(n-k-1)\right) + k}{kb} - 1 \quad (26)$$

$$= \frac{k-1}{2kb}\left(\varphi(n) + \varphi(n-1) - \varphi(n-k) - \varphi(n-k-1)\right) \quad (27)$$

where Equation (40) follows from the conjecture *Lamma 1*. Again, using conjecture *Lamma 1* and Equation (23), and through some algebra, we can bound the worst case coding overhead as follows,

$$\frac{k-1}{4kb}\left(2kn - k^2 - 1\right) \leq \epsilon(n,b) \leq \frac{k-1}{4kb}\left(2kn - k^2 + 1\right) \quad (28)$$

which can be accurately approximated for $b \gg 1$ as

$$\epsilon(n,b) \approx \frac{k-1}{4kb}\left(2kn - k^2\right) = \frac{k-1}{4b}(2n-k) \quad (29)$$

from which the result follows.

APPENDIX B
PROOF OF THM.

Let us start by stating the following lemma.

Lemma B.1: For the projection set given as in (19) with t projections, we have the sum $\sum_{i=0}^{t-1} |p_i|$ that can be expressed in a closed form using the utility function

$$\sum_{i=0}^{t-1} |p_i| = \begin{cases} \frac{t^2+1}{2}, & \text{if } t \text{ is odd} \\ \frac{t^2}{2}, & \text{if } t \text{ is even} \end{cases}$$

Proof: Let us consider the sum for even and odd t separately. First we assume t to be odd. Let us define the set

$$\mathfrak{U}_a = \{\lceil -t+1 \rceil_{odd} - a, \ldots, -1-a, 1-a, \ldots, \lceil t-1 \rceil_{odd} - a\} \quad (30)$$

and notice that $\mathfrak{T} = \mathfrak{U}_0 \cup \mathfrak{U}_1$. Since these sets are disjoint, we have

$$\sum_{i \in \mathfrak{T}} |p_i| = \sum_{i \in \mathfrak{U}_0} |p_i| + \sum_{i \in \mathfrak{U}_1} |p_i| = 2\sum_{i \in \mathfrak{U}_1} |p_i| + 1 \quad (31)$$

Using this relationship and the result of Lemma A.1, we can express

$$\sum_{i \in \mathfrak{U}_0} |p_i| = \sum_{i \in \mathfrak{U}_1} |p_i| + 1 \quad (32)$$

$$= \frac{\sum_{i \in \mathfrak{T}} |p_i| - 1}{2} + 1 \quad (33)$$

$$= \frac{\frac{1}{2}(\phi(2t) - \phi(2t-1)) - 1}{2} + 1 = \frac{t^2+1}{2} \quad (34)$$

Now let us assume t to be even. For this particular assumption we can rewrite

$$\mathfrak{T} = \mathfrak{U}_0 \cup \mathfrak{U}_1 \cup \{t\} \quad (35)$$

Using this observation and the result of Lemma A.1, we can express

$$\sum_{i \in \mathfrak{U}_0} |p_i| = \sum_{i \in \mathfrak{U}_1} |p_i| = \frac{\sum_{i \in \mathfrak{T}} |p_i| - t}{2} \quad (36)$$

$$= \frac{1}{2}\left(\frac{(2t+1)^2 - 1}{4} - t\right) = \frac{t^2}{2} \quad (37)$$

which completes the proof of the lemma.

According to Theorem 3.1, we need to have $\sum_{i=0}^{t-1} |q_i| = tq_e \geq k$. This implies $t = \lceil k/q_e \rceil$ projections are sufficient for perfect reconstruction. For a given pair of projections t_2 and t_1 satisfying $t_2 > t_1$, with the associated projection parameters $(p_i^{(t_2)}, q_i^{(t_2)} = q_e)$ and $(p_i^{(t_1)}, q_i^{(t_1)} = q_e)$ selected based on construction 3.6, we can deduce that

$$\frac{t_2^2 - t_1^2 - 1}{2} \leq \sum_{i=0}^{t_2-1} |p_i^{(t_2)}| - \sum_{j=0}^{t_1-1} |p_j^{(t_1)}| \leq \frac{t_2^2 - t_1^2 + 1}{2} \quad (38)$$

To be able find the contributions from the largest $\lceil k/q_e \rceil$ projections, we set $t_2 = n$ and $t_1 = n - \lceil k/q_e \rceil$. Using similar arguments to previous appendix, we can express the worst case coding overhead in this case as follows

$$\epsilon(n,b) = \frac{1}{kb}\left(\sum_{i \in I_t \setminus S} |p_i|(k-1) + |q_i|(b-1) + \lceil k/q_e \rceil\right) - 1 \quad (39)$$

$$= \frac{(b-1)q_e\lceil k/q_e \rceil + \frac{k-1}{2}\left(\varphi(n) + \varphi(n-1)\right)}{kb} - \frac{\frac{k-1}{2}\left(\varphi(n - \lceil k/q_e \rceil) + \varphi(n - \lceil k/q_e \rceil - 1)\right)}{kb} \quad (40)$$

$$+ \frac{\lceil k/q_e \rceil}{kb} - 1 \quad (41)$$

Using equation (38) and $b \gg 1$, we can accurately approximate the worst case coding overhead as,

$$\epsilon(n,b) \approx \quad (42)$$

$$\frac{\lceil k/q_e \rceil}{kb}\left((k-1)\left(n - \frac{\lceil k/q_e \rceil}{2}\right) + (b-1)q_e + 1\right) - 1$$

REFERENCES

[1] P. G. Farrell, "A survey of array error control codes," preprint, 1990.

[2] M. Blaum and R. M. Roth, "New Array Codes for Multiple Phased Burst Correction", IEEE Trans. on Information Theory, 339(1):66-77, 1993.

[3] S. Lin, and D. J. Costello, Jr., Error Control Coding: Fundamentals and Applications. Prentice-Hall. 1983.

[4] Guedon, J., Barba, D., Burger, N. "Psychovisual image coding via an exact discrete Radon transform." In Wu, L., ed.: Proc. Visual Communications and Image Processing 1995 (VCIP95), Taipei, Taiwan, CORESA (1995) 562572

[5] Guedon, J. P., and Normand, N., "The Mojette transform: The first ten years." In Discrete Geometry for Computer Imagery, E. Andres, G. Damiand, and P. Lienhardt, Eds., vol. 3429 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, pp. 7991.

[6] M. Paterson, D. R. Stinson and Y. Wang, *On Encoding Symbol Degrees of Array BP-XOR Codes*, Submitted for publication, 2013.

[7] M.B. Katz, "Questions of uniqueness and resolution in reconstruction from projections" In: Levin, S. (Ed.), Lecture Notes in Biomathematics, vol. 26. Springer-Verlag, New York.

[8] The On-Line Encyclopedia of Integer Sequences. Available online: https://oeis.org/

[9] Wang, Y., "Array BP-XOR codes for reliable cloud storage systems," In Proc. of IEEE ISIT 2013, pages 326330. IEEE Press (2013).

[10] S. S. Arslan, B. Parrein and N. Normand, "Mojette transform based LDPC erasure correction codes for distributed storage systems," 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, May. 2017, pp. 1-4.