The Rank of Random Binary Matrices and Distributed Storage Applications

Paulo J. S. G. Ferreira, Bruno Jesus, José Vieira, and Armando J. Pinho

***Abstract*—Random binary matrices appear in a variety of signal processing and encoding problems. They play an important role in rateless codes and in distributed storage applications. This paper focuses on block angular matrices, a class of random rectangular binary matrices that are particularly suited to distributed storage applications. We address one of the key issues regarding binary random matrices in general, and block angular matrices in particular: the probability of obtaining a full rank matrix, when drawing uniformly at random from the set of binary matrices with compatible structure. This paper gives a closed-form expression for this probability, as well as some bounds and approximations.**

***Index Terms*—Random matrix, block angular matrix, distributed storage, fountain codes.**

I. Introduction

RATELESS codes are erasure codes designed for channels in which the erasure rate is not known *a priori* or is unpredictable. LT codes [1] and raptor codes [2] are examples of rateless codes. These codes and the digital fountain framework in general are particularly useful in the context of multicast and broadcast protocols [3]. In this framework, the encoded symbols are independently built from random linear combinations of message symbols, and its number is therefore virtually unlimited.

The vector of received encoded symbols that reaches the decoder is therefore given by the product of a random matrix with the vector of message symbols. The decoder has to solve a linear set of equations involving this random matrix. This can be done efficiently if the degree distribution of the encoded symbols satisfies certain conditions [1], [2], [4]. There are decoding algorithms able to recover k message symbols from $k(1+\epsilon)$ encoded symbols, where the overhead parameter ϵ is a fixed small number (say, 0.1). In practice, this means that any received subset of encoded symbols with sufficiently large cardinality can be used to decode the message.

Similar ideas have been found useful in connection with distributed storage systems. If a data file is encoded using a rateless code, and the encoded symbols are distributed across multiple servers, the file can be recovered by using encoded symbols obtained from the servers that are reachable at decoding time (or respond first).

The use of rateless codes allows the generation of an unlimited number of encoded symbols, which can be stored in a virtually unlimited number of servers. During decoding, every encoded symbol is useful, regardless of its origin. Also, decoding performance is not limited by the slowest server, since encoded symbols may come from any server, in any order.

Binary random matrices with a more constrained structure are also of interest. Parallelization of algorithms such as LU and QR factorization have lead to an interest in block angular random matrices, and in the problem of permuting sparse rectangular matrices into block angular form [5]. A block angular matrix has block structure

$$\mathbf{G} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix},$$

where $\mathbf{Y}$ is block diagonal. The simplest non-trivial example (two diagonal blocks) is

$$\mathbf{G} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}.$$

This structure is of high practical interest because the block diagonal $\mathbf{X}$ allows the separation of any problem involving the matrix $\mathbf{G}$ in a number of independent sub-problems, one for each diagonal block. The remaining equations, those involving $\mathbf{Y} = (\mathbf{C}|\mathbf{D})$, represent the coupling between those sub-problems. Because the sub-problems can be solved independently, a problem involving a block angular matrix is inherently amenable to parallelization.

Block angular form is also useful in distributed storage applications. Each diagonal block could be associated with a local server group linked by a local network. In this context, the block $\mathbf{Y}$ would describe connections between the server groups, slower or more expensive than the local ones. The block angular structure is well suited to deal with the gradual loss of encoded symbols due to, say, hardware failures. The idea is to replace lost encoded symbols with linear combinations of existing ones. With the block angular structure, this replacement or regeneration can be done locally, at the group server level, without decoding the entire data file or having to access servers outside the local group.

Note that we are not assuming that the diagonal blocks of $\mathbf{G}$ are necessarily square because of our interest in the performance of block angular structure in coding or distributed storage contexts. The decoding problem starts with a random subset of the elements of the vector $\mathbf{y}$ given by $\mathbf{y} = \mathbf{G}\mathbf{x}$, not the entire vector $\mathbf{y}$. Thus, the equation to solve involves a submatrix of $\mathbf{G}$ obtained by discarding a subset of its rows, determined by the erasure pattern. This submatrix of $\mathbf{G}$, denoted by $\mathbf{M}$, will still have block angular structure, but the blocks will in general be rectangular.

Manuscript received October 1, 2012. The associate editor coordinating the review of this letter and approving it for publication was L. Dolecek.

The authors are with the Signal Processing Laboratory, DETI/IEETA, Universidade de Aveiro, Portugal (e-mail: pjf@ua.pt).

B. Jesus is supported by FCT grant FFRH/BD/28775/2006.

Digital Object Identifier 10.1109/LCOMM.2012.120612.122169

In the applications mentioned, the matrix $\mathbf{M}$ typically has more rows than columns and decoding is possible if $\mathbf{Mx} = \mathbf{y}$ can be solved for $\mathbf{x}$. This explains the interest in the rank of $\mathbf{M}$ — the probability of successful decoding depends on it. This paper studies the rank of block angular and related random matrices, extending the bounds given in [6]. Here we go further, and by using a different approach we give the *exact* expression for the probability of $\mathbf{M}$ being of full rank, and other bounds and approximations.

II. Related work

The idea of local and global protection is also explored (although in a different form) in pyramid codes [7], in which block groups are protected by redundant blocks (local protection) but the entire data are also protected by global redundancy. Decoding starts at the lowest level and proceeds to the global level, as in climbing a pyramid, as reflected in the code name.

Erasure codes for storage over a network when the data sources are distributed were considered in [8]. There, the authors introduce optimally sparse decentralized erasure codes, and show that they have advantages over random linear codes. A more recent work [4] introduces codes able to recover k message symbols from a random subset of $(1+\epsilon)k$ encoded symbols with high probability, with logarithmic locality: a single symbol loss can be repaired by accessing $O(\log k)$ encoded symbols.

The decoding problem for LT codes [1] involves random matrices over finite fields. For a certain degree distribution it can be solved efficiently by an online algorithm. Raptor codes [2] use a pre-code to further reduce the needed degree to a constant.

Random matrices (over finite or infinite fields) have many other important applications. See [9] for an extensive review of random matrix theory over the real or complex fields. For results on the rank over the real numbers see [10]. A number of results on random linear equations over finite fields were reviewed in [11]. This includes results on the rank, determinant and permanent of matrices over $GF(q)$, assuming elements with certain probability distributions. Wiedemann gave a method to solve sparse equations over a finite field [12].

Despite the wealth of results on random matrices and their spectral properties, the current knowledge about structured matrices is, as stated in the 65-page review [13], very limited. The same can be said about random matrices over finite fields (see [14] for a review). The rank of sparse matrices is studied in [15] (see also [16], [17]) and [18] studies the spectral distribution of a certain circulant matrix.

Our work differs from all these because we consider matrices with a specific block structure. The work [5] considers block angular matrices, but in connection with a very different problem (the possibility of permuting sparse rectangular matrices into block angular form).

III. Background

This section summarizes a few known results about the rank of random matrices that will be subsequently needed. Some further details can be found in [6], [19]. A few useful approximations that might be new are also given.

All matrices and vectors are taken over the field $GF(2)$. The number of full rank $n \times m$ matrices, with $n \geq m$, is given by

$$F(n,m) = (2^n - 1)(2^n - 2)\cdots(2^n - 2^{m-1}) = \prod_{i=0}^{m-1}(2^n - 2^i). \tag{1}$$

If every $n \times m$ matrix is equally likely to occur, the probability of selecting a matrix with full rank is

$$P(n,m) = 2^{-nm}\,F(n,m) = \prod_{i=0}^{m-1}(1 - 2^{i-n}). \tag{2}$$

There is a relatively small probability that a square matrix has full rank over $GF(2)$. In fact, $P(n,n)$ converges rapidly to a value close to 0.28. On the other hand, for fixed m, $P(n,m)$ quickly approaches unity as n increases beyond m. One can derive the following upper bound

$$\begin{aligned}\log P(n,m) &= \sum_{i=0}^{m-1}\log(1 - 2^{i-n})\\ &\leq -\sum_{i=0}^{m-1}2^{i-n} = -2^{-n}(2^m - 1).\end{aligned}$$

Denoting the excess of rows by k, so that $n = m + k$, this can be written

$$\log P(m+k,m) \leq -2^{-k}(1 - 2^{-m}). \tag{3}$$

The result can also be obtained by applying $1 - \epsilon \leq e^{-\epsilon}$ to each of the terms in (2). In fact, it asymptotically approaches the exact value in (2) as m increases, but this is perhaps best shown directly from (2), since

$$P(n,m) = 1 - \sum_{i=0}^{m-1}2^{i-n} + \cdots = 1 - 2^{-n}(2^m - 1) + \cdots$$

The omitted terms are higher order products of the quantities $x_i = 2^{i-n}$. Discarding them leads to

$$P(m+k,m) \approx 1 - 2^{-k}(1 - 2^{-m}),$$

and so the probability of *not* having full rank satisfies

$$1 - P(m+k,m) \approx 2^{-k}(1 - 2^{-m}),$$

or

$$\log(1 - P(m+k,m)) \approx -2^{-m} - k\log 2.$$

Thus, full rank matrices become exponentially more likely as the excess number of rows increases.

The probability that a $n \times m$ matrix has rank r has been known for over a century. The number of $n \times m$ matrices with rank r can be obtained by multiplying the number of matrices with an r-dimensional range by the number of distinct subspaces of dimension r. If $G(n,m,r)$ denotes the number of $n \times m$ matrices of rank r, then

$$G(n,m,r) = \frac{F(m,r)F(n,r)}{F(r,r)}, \tag{4}$$

To find the probability $P(n,m,r)$ it is only necessary to divide this quantity by the total number of $n \times m$ matrices. For more about this and the more general case in $GF(q)$ see [19], for example.

IV. Full rank probability of block angular matrices

The matrix $\mathbf{M}$ mentioned in the theorems below is given by

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \left(\mathbf{L} \middle| \mathbf{R} \right), \tag{5}$$

where $\mathbf{A}$ is $a \times a'$, $\mathbf{B}$ is $b \times b'$, $\mathbf{C}$ is $c \times a'$ and $\mathbf{D}$ is $c \times b'$. We denote the maximum number of nonzero elements of $\mathbf{M}$ by

$$N = aa' + bb' + c(a' + b'),$$

so that the number of distinct matrices $\mathbf{M}$ with geometry defined by a, a', b, b' and c can be written simply as 2^N. We start with a simple upper bound that, in practice, is often quite accurate.

Theorem 1: The probability p that the matrix $\mathbf{M}$ given by (5) has full rank satisfies

$$p \le 2^{-N} F(a + c, a') F(b + c, b'). \tag{6}$$

Proof: There are $F(a + c, a')$ full rank matrices $\mathbf{L}$ and $F(b + c, b')$ full rank matrices $\mathbf{R}$, since the zero blocks in $\mathbf{L}$ and $\mathbf{R}$ can be ignored. Thus, the number of full rank matrices of the form $\mathbf{M} = (\mathbf{L}|\mathbf{R})$ cannot exceed the product $F(a + c, a')F(b + c, b')$.

The product is only an upper bound to the number of full rank matrices $\mathbf{M} = (\mathbf{L}\ \mathbf{R})$ because there the full rank matrices $\mathbf{L}$ and $\mathbf{R}$ may contain a common subset of linearly dependent columns, as in the following example:

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \left(\mathbf{L} \middle| \mathbf{R} \right) = \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array} \right).$$

Both $\mathbf{L}$ and $\mathbf{R}$ have full column rank but $\mathbf{M}$ has linearly dependent columns. ■

In the counterexample given above, the diagonal blocks $\mathbf{A}$ and $\mathbf{B}$ are rank deficient. Because of this, the linear independence of the columns of $\mathbf{M}$ depends on the linear independence of the columns of $\mathbf{C}$ and $\mathbf{D}$, which, therefore, cannot be independently specified. The proof of the following theorem, which gives the exact probability p, does in fact depend on the balance between the rank of the diagonal blocks and the rank of the $\mathbf{C}$ and $\mathbf{D}$ blocks.

Theorem 2: The probability p that the matrix $\mathbf{M}$ given by (5) has full rank is given by

$$p = \frac{\sum G(a, a', i) G(b, b', j) 2^{(i+j)c} F(c, a' - i + b' - j)}{2^N}, \tag{7}$$

where the sum is over all pairs i, j that satisfy

$$i + j \ge a' + b' - c.$$

Proof: Assume that the ranks of $\mathbf{A}$ and $\mathbf{B}$ are i and j, respectively. Then, i of the columns of $\mathbf{C}$ and j of the columns of $\mathbf{D}$ can be selected in an arbitrary way. However, the remaining $a' - i$ columns of $\mathbf{C}$ and $b' - j$ columns of $\mathbf{D}$ must be linearly independent.

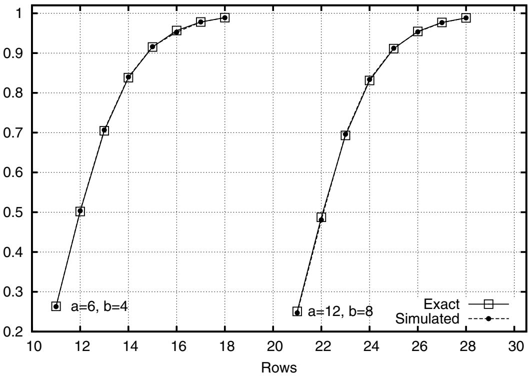

Fig. 1. Full rank probability for a block angular matrix according to (7), and estimated values obtained by simulation. The block angular matrix has square diagonal blocks.

It is easy to count the number of matrices $\mathbf{A}$ and $\mathbf{B}$ with ranks i and j: as seen before, it is given by $G(a, a', i)$ and $G(b, b', j)$, respectively.

The i arbitrary columns of $\mathbf{C}$ can be selected in 2^{ic} ways, because each column has dimension c. Similarly, the j arbitrary columns of $\mathbf{D}$ can be selected in 2^{jc} ways.

Finally, the remaining $a' - i$ and $b' - j$ linearly independent columns of $\mathbf{C}$ and $\mathbf{D}$ can be selected in $F(c, a' - i + b' - j)$ ways, because they can be thought of as a full rank matrix of size $c \times (a' - i + b' - j)$. This is possible only if $c \ge a' - i + b' - j$, that is, $i + j \ge a' + b' - c$.

The total number of full rank matrices $\mathbf{M}$ is obtained by summing over all possible ranks:

$$\sum_{i+j \ge a'+b'-c} G(a, a', i) G(b, b', j) 2^{(i+j)c} F(c, a' - i + b' - j),$$

which leads to (7). ■

Fig. 1 shows this probability for some possible sizes, as well as estimated values obtained by simulation.

The case $c = 0$ has a simple interpretation. Assume to simplify the notation that the diagonal blocks are square. Since only the term due to $i = a$, $j = b$ is allowed in the sum, the expression for the probability reduces to

$$p = \frac{G(a, a, a) G(b, b, b)}{2^{a^2 + b^2}} = \frac{F(a, a) F(b, b)}{2^{a^2 + b^2}}, \tag{8}$$

which represents the number of full rank $a \times a$ matrices $\mathbf{A}$ multiplied by the number of full rank $b \times b$ matrices $\mathbf{B}$, divided by the total number of matrices — as one would expect in a pure block diagonal case.

This can be compared with the upper bound (6). When $c = 0$, the bound coincides with the probability, since the matrix $\mathbf{M}$ contains only the diagonal blocks. Setting $c = 0$ in (6) leads to

$$p = 2^{-N} F(a, a') F(b, b') = 2^{-aa' - bb'} F(a, a') F(b, b'),$$

which reduces to (8) when the diagonal blocks are square.

It would be interesting to compare the full rank probability in the block angular case with the full rank probability of a

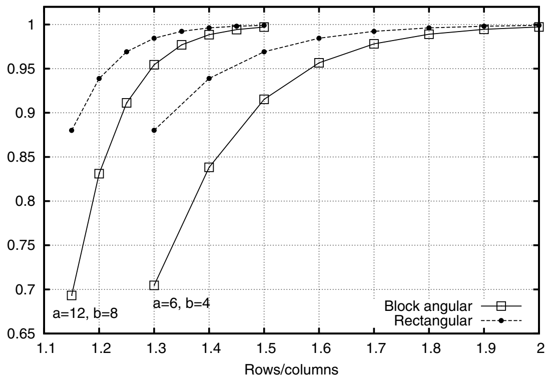

Fig. 2. Full rank probability for block angular matrices and for matrices without any block structure. The block angular matrix has square diagonal blocks.

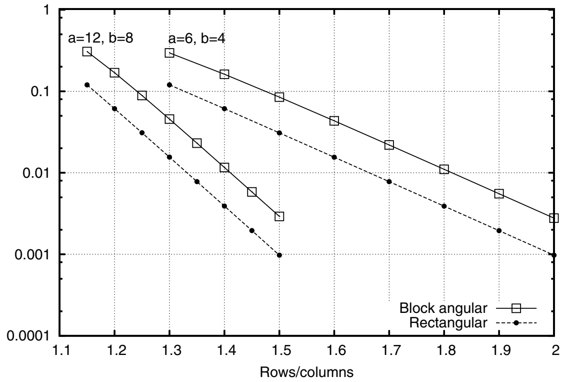

Fig. 3. Rank deficient probability for block angular matrices and matrices without any block structure. The block angular matrix has square diagonal blocks.

rectangular, unrestricted matrix, but some care has to be taken to ensure a fair comparison.

Assume to simplify the notation that $a = a'$ and $b = b'$. Then, the maximum number of nonzero entries of a block angular matrix is $a^2 + b^2 + c(a+b)$. A rectangular matrix with the same number of columns $(a + b)$ would have to contain about

$$n = \frac{a^2 + b^2 + c(a+b)}{a+b}$$

rows. The comparison results shown in Fig. 2 and Fig. 3 show that the block angular structure does not imply a large increase in redundancy (number of rows divided by number of columns) as compared with unrestricted matrices without any block structure.

The main result can be extended to several diagonal blocks and consideration of fields other than GF(2) is also possible. We intend to explore some of the possibilities in the future.

V. Conclusion

Binary random rectangular matrices over $GF(2)$ with k more rows than columns are of full rank with high probability (in fact, the probability of *not* having full rank decreases with 2^{-k}). We have given the exact expression for the full rank probability in the block angular case and compared it with the unrestricted rectangular case. The results show that it is possible to rely on block angular structure at a small price, namely, an increase in the excess of rows over columns for the same full rank probability.

These results have consequences for the efficient decoding of codes that lead to decoding problems involving block angular matrices as well as for distributed storage applications, in which encoded symbols are spread among a number of servers but data retrieval has to be carried out using only the servers that are reachable or respond faster at decoding time.

References

[1] M. Luby, “LT codes,” in *Proc. 2002 IEEE Symposium on Foundations of Computer Science*, pp. 271–282.

[2] A. Shokrollahi, “Raptor codes,” *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[3] J. W. Byers, M. Luby, and M. Mitzenmacher, “A digital fountain approach to asynchronous reliable multicast,” *IEEE J. Sel. Areas Commun.*, vol. 20, no. 8, pp. 1528–1540, Oct. 2002.

[4] M. Asteris and A. G. Dimakis, “Repairable fountain codes,” in *Proc 2012 IEEE International Symposium on Information Theory*, pp. 1752–1756.

[5] C. Aykanat, A. Pinar, and U. V. Çatalyürek, “Permuting sparse rectangular matrices into block-diagonal form,” *SIAM J. Sci. Comput.*, vol. 25, no. 6, pp. 1860–1879, 2004.

[6] P. J. S. G. Ferreira, B. Jesus, J. Vieira, and A. J. Pinho, “Random block-angular matrices for distributed data storage,” in *Proc. 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3180–3183.

[7] C. Huang, M. Chen, and J. Li, “Pyramid codes: flexible schemes to trade space for access efficiency in reliable data storage systems,” in *Proc. 2007 NCA*, pp. 79–86.

[8] A. G. Dimakis, V. Prabhakaran, and K. Ramchandran, “Decentralized erasure codes for distributed networked storage,” *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2809–2816, June 2006.

[9] M. L. Mehta, *Random Matrices*. Elsevier, 2004.

[10] B. Bollobás, *Random Graphs*. Cambridge University Press, 2001.

[11] I. N. Kovalenko and A. A. Levitskaya, “Probabilistic properties of systems of random linear equations over finite algebraic structures,” *Cybernetics and Systems Analysis*, vol. 29, no. 3, pp. 385–390, 1993.

[12] D. H. Wiedemann, “Solving sparse linear equations over finite fields,” *IEEE Trans. Inf. Theory*, vol. 32, no. 1, pp. 54–62, Jan. 1986.

[13] A. Edelman and N. R. Rao, “Random matrix theory,” *Acta Numer.*, vol. 14, pp. 233–297, 2005.

[14] J. Fulman, “Random matrix theory over finite fields,” *Bull. Am. Math. Soc., New Ser.*, vol. 39, no. 1, pp. 51–85, 2001.

[15] J. Blömer, R. Karp, and E. Welzl, “The rank of sparse random matrices over finite fields,” *Random Structures and Algorithms*, vol. 10, no. 4, pp. 407–419, July 1997.

[16] C. Cooper, “On the rank of random matrices,” *Random Structures and Algorithms*, vol. 16, no. 2, pp. 209–232, Mar. 2000.

[17] ——, “On the distribution of rank of a random matrix over a finite field,” *Random Structures and Algorithms*, vol. 17, no. 3-4, pp. 197–212, Oct. 2000.

[18] A. Bose and J. Mitra, “Limiting spectral distribution of a special circulant,” *Statistics & Probability Letters*, vol. 60, pp. 11–120, 2002.

[19] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*. Cambridge University Press, 1992.