RANDOM BLOCK-ANGULAR MATRICES FOR DISTRIBUTED DATA STORAGE

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

Departamento de Electrónica,Telecomunicações e Informática
Universidade de Aveiro
Aveiro, Portugal
{*pjf, bruno.jesus, jnvieira, ap*}@*ua.pt*

ABSTRACT

Random binary matrices have found many applications in signal processing and coding. Rateless codes, for example, are based on the random generation of codewords by means of inner products between the data and random binary vectors. But the usefulness of random binary matrices is not limited to coding: they are also well suited to distributed data storage applications. In this context, random binary matrices with block-angular structure are of particular interest because they allow cooperative encoding and decentralized models for coding and decoding, with a built-in degree of parallelism. Linear programming, LU factorization and QR factorization are some of the problems for which the coarse-grain parallelization inherent in the block-angular structure is of interest. This paper studies one of the most important characteristics of block-angular matrices, their rank. More precisely, we study the rank distribution and full rank probability of rectangular random binary matrices and block-angular matrices in $GF(2)$.

***Index Terms*—** Random matrices, random inner products, block-angular matrices, rank, full rank probability, rank distribution.

1. INTRODUCTION

Rateless codes [1, 2] are a class of error correcting codes useful for channels in which the loss rate is not known *a priori* or is prone to large unpredictable variations. In these codes, the codewords are built from random linear combinations of message symbols.

This randomness allows the generation of a virtually unlimited number of codewords for a given message. The receivers use any subset of codewords with sufficiently many elements to perform decoding. The overhead is not large: there are decoding algorithms able to recover k symbols of the original message from $k(1 + \epsilon)$ codewords, where ϵ is a small positive number (say, 0.1).

Similar ideas have been found useful to implement redundant, reliable, distributed data storage. The idea is to encode the file data as codewords and distribute the codewords across multiple servers. At any given moment, the data can be recovered by using the codewords stored on the servers that can be reached at that moment. The use of rateless codes allows the storage of codewords on an unlimited number of servers. Furthermore, the decoding process can accept and use codewords coming from any server. The data retrieval performance is not bounded by the slowest server: codewords may come from any server, in any order.

We consider binary matrices with elements in $GF(2)$, the binary Galois field. A block-angular matrix has structure

$$M = \begin{bmatrix} D \\ R \end{bmatrix} \tag{1}$$

where D is block-diagonal. The simplest case is

$$M = \begin{bmatrix} X & 0 \\ 0 & Y \\ W & Z \end{bmatrix}. \tag{2}$$

For coding and data storage applications, the matrix M must have more rows than columns. It has been shown [3] that block-angular matrices yield results similar to those that can be obtained with binary random matrices. This is important because the block-diagonal D allows the separation of problems involving M in a number of sub-problems related to each of the block-diagonal blocks. The rows in the block R represent coupling between the different problems.

This combination of characteristics makes block-angular structure attractive for many applications. It also allows the speedup of the decoding process with multicore architectures while reducing the communication between the cores — or, in distributed storage applications, the communication between servers.

Most of the problems that involve random matrices, even sparse ones, do not exhibit block-diagonal or even block-angular structure. However, given their useful properties, there has been interest in algorithms [4, 5] that explore their relations. The method to transform a

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

sparse matrix to block-diagonal form [4] is a good example (interestingly, the problem of rearranging binary matrices in block-angular form was discussed as early as 1971 [6]). To avoid the cost brought by such additional steps, it has been proposed [3] to directly generate coding matrices with block-angular structure of size by $(n+e) \times n$. The matrix must have full column rank, but it is easy to confirm experimentally that the full rank probability in the block-angular case rapidly increases with the number of excess rows e (see [3]).

In the remainder of this paper we study the properties of random block-angular matrices. We start by reviewing, in the next section, some of the key results concerning the rank of random binary matrices over $GF(2)$. Then we discuss the rank of random block-diagonal and random block-angular matrices. Our results include a convenient and useful bound for the probability of block-angular matrix to have full (column) rank, a result of immediate interest to the many algorithms and applications that rely on these random matrices.

2. RANDOM BINARY MATRICES

Number of matrices with full rank. The number of full rank matrices of size $n \times m$, with $n \geq m$, can be determined as follows. Pick one column; it can be any of the $2^n - 1$ possible nonzero vectors. Pick another column; it cannot be equal to the previous one nor equal to the zero vector. The are $2^n - 2$ such vectors. In general, column $k+1$ cannot be equal to a linear combination of the previously considered k columns. There are 2^k such linear combinations. Thus, the total number of full rank $n \times m$ matrices is given by

$$F(n,m) = (2^n - 1)(2^n - 2) \cdots (2^n - 2^{m-1}) = \prod_{i=0}^{m-1} (2^n - 2^i). \quad (3)$$

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

$$\begin{aligned} P(n,m) &= \frac{F(n,m)}{2^{nm}} \\ &= (1 - 2^{-n})(1 - 2^{-n+1}) \cdots (1 - 2^{-n+m-1}) \\ &= \prod_{i=0}^{m-1} (1 - 2^{i-n}). \end{aligned}$$

This is illustrated in Fig. 1.

Rank distribution. The probability that a $n \times m$ matrix has rank r is not difficult to find and it is known for over a century. To count the matrices of rank r it is necessary to count the matrices with an r-dimensional range. Denote this number by N. Multiplying it by the number of distinct subspaces of dimension r yields the required count. Dividing it by the total number of matrices yields the probability.

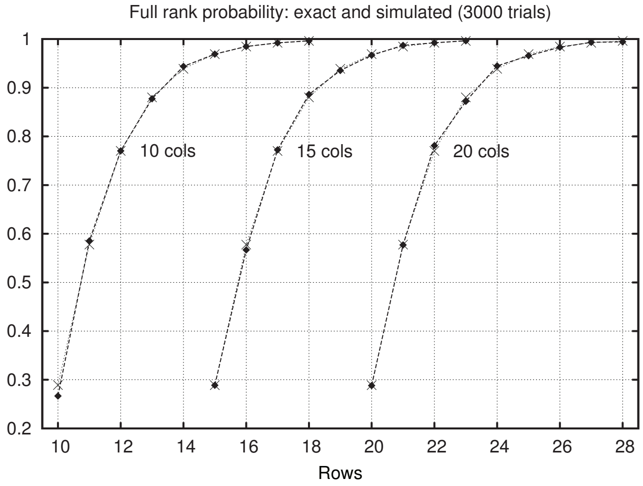

Fig. 1. Full rank probability for a few matrix sizes, as a function of the number of rows. The exact probability matches the empirical distribution well.

To find N, note that a matrix of size $n \times m$ with an r-dimensional column space linearly maps m-vectors to r-vectors. Thus, N can be obtained by counting the number of $r \times m$ matrices of rank r.

Under the hypothesis $n \geq m$, we know that r is at most m. Thus, the $r \times m$ matrix of rank r must have full row rank (linearly independent rows). There are $2^m - 1$ choices for row 0, $2^m - 2$ choices for row 1, and so on. There are $2^m - 2^{r-1}$ choices for row r. N is therefore given by

$$N = F(m,r) = (2^m - 1)(2^m - 2) \cdots (2^m - 2^{r-1}) = \prod_{i=0}^{r-1} (2^m - 2^i) \quad (4)$$

To complete the task we need only count the number S of r-dimensional subspaces of the parent n-dimensional space, and multiply this number by N. It is easier to count the number of bases: since r linearly independent n-vectors yield a base for a r-dimensional subspace, there are $F(n,r)$ bases. However, a single subspace can be generated by a number of distinct bases, so $F(n,r)$ must be divided by the number of bases of each r-dimensional subspace. This number is simply $F(r,r)$ and so $S = F(n,r)/F(r,r)$. The conclusion is that

$$F(n,m,r) = \frac{F(m,r)F(n,r)}{F(r,r)} \quad (5)$$

where $F(n,m,r)$ denotes the number of $n \times m$ matrices of rank r. To find the probability $P(n,m,r)$ it is only

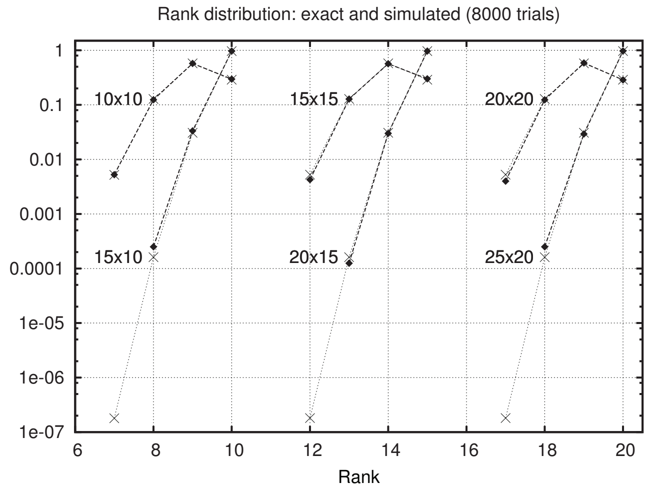

Fig. 2. Probability that a random matrix has a certain rank, for a few matrix sizes. Both the exact probability and an approximation found using simulation are shown. The most likely rank of a square random matrix of size n is $n-1$ but random matrices with more rows than columns are more likely to be of full column rank.

necessary to divide by the total number of matrices. The result is illustrated in Fig. 2.

The number of subspaces of a space is also known as the Gaussian binomial coefficient and it is related to q-calculus. There is a beautiful and intriguing analogy between the number of subsets of a set and the number of subspaces of a space. The former is given by the binomial coefficient, the latter by the Gaussian binomial coefficient [7].

3. BLOCK-ANGULAR MATRICES

We now discuss the block-angular case. Consider first a matrix of size $(a+b+c) \times (a+1)$ such that

$$M = \begin{pmatrix} A & 0 \\ 0 & B \\ C & D \end{pmatrix} = \left(L \,\middle|\, r \right) \tag{6}$$

where A is $a \times a$, B is $b \times 1$, C is $c \times a$, D is $c \times 1$. Thus, L is $(a+b+c) \times a$ and r is $(a+b+c) \times 1$.

The number of matrices L with linearly independent columns is

$$F(a+c, a) = \prod_{i=0}^{a-1} (2^{a+c} - 2^i). \tag{7}$$

Nontrivial combinations of the a columns of L lead to vectors of the form

$$\begin{pmatrix} x \\ 0 \\ y \end{pmatrix} \tag{8}$$

where x and y are not both zero. To keep the matrix full rank, the new column r must not be one of these vectors.

The total number of nonzero possibilities for the new column r is $2^{b+c} - 1$, since by construction r is allowed to have $b+c$ nonzero entries. The number of full rank matrices must therefore be less than

$$F(a+c, a)(2^{b+c} - 1) \tag{9}$$

and so the full rank probability must satisfy

$$p \leq \frac{F(a+c, a)(2^{b+c} - 1)}{2^{a^2+b+c(a+1)}} \tag{10}$$

We have found this upper bound useful and surprisingly accurate.

It is also possible to obtain a lower bound. Note that any $B \neq 0$ necessarily leads to a full rank matrix. There are $2^b - 1$ possibilities. For each such choice, D provides an additional factor of 2^c. The number of full rank matrices must therefore be at least

$$F(a+c, a)(2^b - 1)2^c \tag{11}$$

and so the full rank probability has to satisfy

$$p \geq \frac{F(a+c, a)(2^b - 1)2^c}{2^{a^2+b+c(a+1)}} \tag{12}$$

The difference between the upper and lower bounds is

$$\Delta p = \frac{F(a+c, a)(2^c - 1)}{2^{a^2+b+c(a+1)}} \tag{13}$$

If c is small compared with the other dimensions, the bounds should be relatively tight. For an example see Fig. 3.

We may now consider the general case. Consider adding to L not one but b columns of the form

$$\begin{pmatrix} 0 \\ x \\ y \end{pmatrix}. \tag{14}$$

The number of matrices L with linearly independent columns is still $F(a+c, a)$. We start by considering $x \neq 0$ (which yields a lower bound for the probability). The first new column can be chosen in $2^b - 1$ ways (for the x part) times 2^c (for the y part). The second column can be chosen in $2^b - 2$ ways times 2^c; The process is repeated b times (so that the final B is of size $b \times b$) to yield

$$F(b, b)2^{bc}. \tag{15}$$

The full rank probability is therefore bounded by

$$p \geq \frac{F(a+c, a)F(b, b)2^{bc}}{2^{a^2+b^2+c(a+b)}} \tag{16}$$

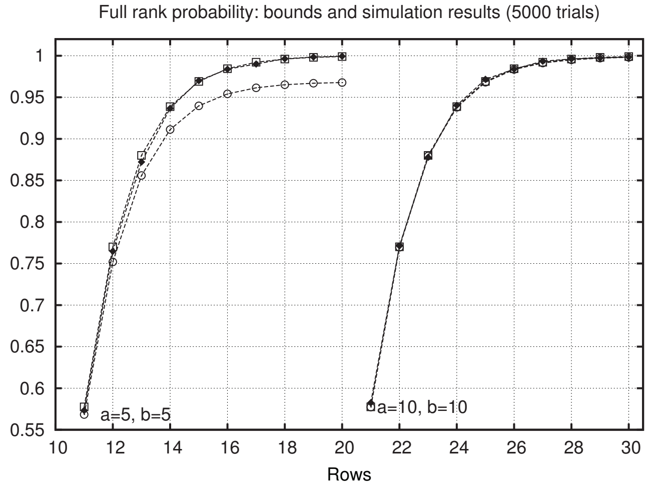

Fig. 3. Upper and lower bounds for the full rank probability and approximations found using simulation. Except for very small matrix sizes, the bounds are tight and provide a useful approximation to the full rank probability.

To obtain an upper bound, observe that the first column can be chosen in $2^{b+c} - 1$ ways, the next one in $2^{b+c} - 2$ ways, and so forth. Repetition of the process b times leads to a factor of $F(b+c, b)$ and so

$$p \le \frac{F(a+c,a)F(b+c,b)}{2^{a^2+b^2+c(a+b)}} \qquad (17)$$

We have found this bound to be both accurate and useful. Plots are given in Fig. 4.

4. CONCLUSIONS

We examined rectangular random binary matrices from the viewpoint of full rank probability and rank distribution. We have seen, for example, that the most likely rank of a square random matrix of size n is $n - 1$. As known from rateless codes, random matrices with more rows than columns are very likely to be of full column rank. We have also obtained the rank distribution of a rectangular random matrix, that is, the probability of the rank being equal to a given integer r.

Then we considered random matrices with block-angular structure, which are of interest in the context of distributed data storage problems. Our main results are lower and upper bounds for the full rank probability of these matrices, assuming a block-diagonal structure similar to the one in Eq. (2). These results have consequences for the efficient decoding of codes based on these matrices and for distributed data storage applications — in which the codewords are spread among a number of servers but data retrieval has to be carried out using only the servers that respond to the data queries.

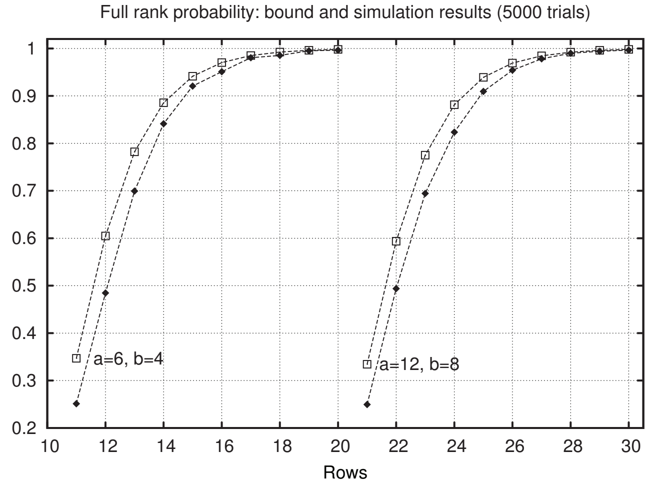

Fig. 4. Block-angular case. Upper bound for the full rank probability and approximations found using simulation.

5. REFERENCES

[1] M. Luby, "LT codes," in *43th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002.

[2] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[3] B. Jesus, J. M. N. Vieira, and P. J. S. G. Ferreira, "Coding with low density block angular generator matrix," in *8th World Congress on Intelligent Control and Automation*, Jinan, China, 2010.

[4] Cevdet Aykanat, Ali Pinar, and Ümit V. Çatalyürek, "Permuting sparse rectangular matrices into block-diagonal form," *SIAM Journal on Scientific Computing*, vol. 25, no. 6, pp. 1860–1879, 2004.

[5] Ali Pinar, Edmond Chow, and Alex Pothen, "Combinatorial algorithms for computing column space bases that have sparse inverses," *Electronic Transactions on Numerical Analysis*, vol. 22, pp. 122–145, 2006.

[6] Roman L. Weil and Paul C. Kettler, "Rearranging matrices to block-angular form for decomposition (and other) algortithms," *Management Science*, vol. 18, no. 1, pp. 98–108, Sept. 1971.

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