Data Storage and Data Compression II

C. Lam
Communications Systems Research Section

In a recent article, Odlyzko showed that under certain idealized circumstances, a small increase in data storage capability can lead to a dramatic increase in the rate at which data can be communicated reliably. In this article a detailed investigation is made of the circumstances under which the maximum possible rate increase will occur.

I. Introduction

In a recent article, Odlyzko (Ref. 1) showed that under certain idealized circumstances, a small increase in data storage capability can lead to a dramatic increase in the rate at which data can be communicated reliably. In this article we will investigate in detail the circumstances under which the maximum possible rate increase will occur.

Let X be a set, whose elements are to be regarded as the possible outcomes of some experiment. Let S be a collection of subsets of X; we assume that when a sample $x \in X$ is obtained, the experimenter is satisfied in knowing only some $A \in S$ such that $x \in A$. The sets A of S are therefore sometimes called the *subsets of allowed uncertainty*. As explained in detail in (Ref. 1), if n_N denotes the minimum number of sets of the form $A_1 \times A_2 \times \cdots \times A_N$ needed to cover $X \times \cdots \times X$ (N copies of X), then if the data handling system can store N samples prior to transmission the data rate is proportional to $N/\log n_N$. Odlyzko showed that n_N can sometimes be as small as $Nn_1 - N$; i.e., the best possible increase in data rate is *linear* in the amount of storage. This is quite remarkable since for most systems $N/\log n_N$ is a constant, or nearly so, independent of N. Our goal here is to investigate the circumstances under which $n_N = Nn_1 - N$, and some related questions.

Let S be a collection of subsets of a set X such that their union is X. Define $c(X; S)$, the *covering number* of X with respect to S, to be the minimal number of elements of S

whose union is X, if this number exists, and infinity if no finite subcollection of S covers X. If $S_1, \cdots, S_N$ are collections of subsets of $X_1, \cdots, X_N$ respectively, define a collection $S_1 \times \cdots \times S_N$ of subsets of the Cartesian product $X_1 \times \cdots \times X_N$ by

$$S_1 \times \cdots \times S_N = \{A_1 \times \cdots \times A_N | A_i \in S_i, \quad i = 1, \cdots, N\}$$

We will restrict ourselves to the case in which all the $c(X_i; S_i)$ are finite, since otherwise

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) = \infty$$

Odlyzko (Ref. 1) obtained an upper bound

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \leq \prod_{i=1}^{N} c(X_i; S_i) \qquad (1)$$

and a lower bound

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \geq \sum_{i=1}^{N} [c(X_i; S_i) - 1] + 1 \qquad (2)$$

The surprising result is that equality can be attained in Eq. 2, and in Ref. 1 Odlyzko gives a construction for it. In Subsection II of this paper, we will give an equivalent version of the problem in (0, 1)-matrix terms. In Subsection III, we will give a necessary and sufficient condition that the lower bound of Eq. 2 is obtained by a given partition. In particular, we will show that the Odlyzko constructions give essentially the only case that achieves the lower bound. In Subsection IV, we will give some conditions on when the upper bound of Eq. 1 is obtained.

II. Formulation of Problem as a (0,1)-Matrix Problem

A (0, 1)-matrix of size r by s is a matrix with r rows and s columns, in which all the entries are either 0 or 1. We now associate a (0, 1)-matrix with a set X and S, a collection of subsets of X.

Let $X = \{a_1, \cdots, a_r\}$, $S = \{A_1, \cdots, A_s\}$. Define

$$b_{ij} = \begin{cases} 1 & \text{if } a_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

Then $B = \{b_{ij}\}$ is a (0, 1)-matrix of size $r \times s$, and it represents the relationships between X and S.

We will call the minimum number of columns of a (0, 1)-matrix that has at least one 1 in each row the *1-width* of the matrix. So, $c(X; S)$ is just the 1-width of the matrix B representing X and S. We write $\epsilon(B)$ for the 1-width of B.

If B, C are (0, 1)-matrices, let

$$B = \begin{Bmatrix} b_{11} \cdots b_{1s} \\ \cdot \\ \cdot \\ \cdot \\ b_{r1} \cdots b_{rs} \end{Bmatrix}$$

The Cartesian (tensor) product is

$$B \times C = \begin{Bmatrix} b_{11}C \cdots b_{1s}C \\ \cdot \\ \cdot \\ \cdot \\ b_{r1}C \cdots b_{rs}C \end{Bmatrix}$$

where $b_{ij}C$ denotes multiplying all the entries of C by b_{ij}. The matrix $B \times C$ is obtained by placing these blocks of matrices side by side.

It can be easily shown that if B represents X_1 and S_1, and D represents X_2 and S_2, then the product $B \times D$ represents the product $S_1 \times S_2$. Therefore

$$\epsilon(B \times D) = c(X_1 \times X_2; S_1 \times S_2)$$

The results in Subsection I are now translated to

$$\epsilon(B_1 \times \cdots \times B_N) \leq \prod_{i=1}^{N} \epsilon(B_i) \qquad (3)$$

$$\epsilon(B_1 \times \cdots \times B_N) \geq \sum_{i=1}^{N} (\epsilon(B_i) - 1) + 1 \qquad (4)$$

From now on, we will work with the matrix version of the problem, which should be easier to visualize.

III. Conditions for Achievement of Lower Bound

We start this section off with some elementary observations about the 1-width of a (0, 1)-matrix.

A column (row) c_1 is said to *cover* another column (row) c_2 if whenever there is a 1 in the column (row) of c_2, there is a 1 in the corresponding position of c_1.

Property 1

If a row c_1 of a $(0, 1)$-matrix A covers another row c_2 then the row c_1 can be removed without affecting $\epsilon(A)$.

Proof:

In any choice of columns of A which gives a 1 in each row, one of the columns must have a 1 in row c_2. Since row c_2 is covered by row c_1, a 1 must appear in row c_1 in the same column.

Property 2

If a column c_1 of a $(0, 1)$-matrix A covers another column c_2, then c_2 can be removed without affecting $\epsilon(A)$.

Proof:

In any choice of columns which includes c_2, we could do as well or better by replacing column c_2 with c_1.

Note that the two processes above can be applied repeatedly to reduce the size of the matrix, which would make it easier to find the 1-width.

Given two integers ϵ, n with $n \geqq \epsilon > 0$, Odlyzko's construction gives a matrix of size

$$\binom{n}{n-\epsilon+1} \times n$$

with constant row sums $(n - \epsilon + 1)$, and the

$$\binom{n}{n-\epsilon+1}$$

rows represent all the possible ways of putting $(n - \epsilon + 1)$ 1s in n positions. We call such a matrix one of Odlyzko's type with parameters n and ϵ.

Example: $n = 5, \quad \epsilon = 3$

$$\binom{n}{n-\epsilon+1} = 10$$

1	1	1	0	0
1	1	0	1	0
1	1	0	0	1
1	0	1	1	0
1	0	1	0	1
1	0	0	1	1
0	1	1	1	0
0	1	1	0	1
0	1	0	1	1
0	0	1	1	1

Theorem 1

The 1-width of a $(0, 1)$-matrix of Odlyzko type is ϵ. Moreover, any choices of ϵ columns will have at least one 1 in each row.

Proof:

In every row, the row sum is $(n - \epsilon + 1)$, therefore there are only $(\epsilon - 1)$ 0's. So, in any choice of ϵ columns, it can at most contain $(\epsilon - 1)$ 0's in this row. Hence ϵ columns are enough to give at least one 1 in each row. On the other hand, if we take $(\epsilon - 1)$ columns, there is a row with all its $(\epsilon - 1)$ 0's in these columns. Hence $(\epsilon - 1)$ columns is not enough. Thus, the 1-width of the matrix is ϵ.

Odlyzko has proved the following theorem.

Theorem 2

If $B_1, \cdots, B_N$ are $(0, 1)$-matrices with 1-width $\epsilon_1, \cdots, \epsilon_N$, then

$$\epsilon(B_1 \times \cdots \times B_N) \geqq \sum_{i=1}^{N} (\epsilon_i - 1) + 1 \tag{5}$$

Proof: (See Ref. 1.)

We now investigate the conditions when equality in Eq. (5) is attained.

Theorem 3

Let $\epsilon_1, \cdots, \epsilon_N$ be positive integers, and let $B_1, \cdots, B_N$ be $(0, 1)$-matrices with 1-width $\epsilon_1, \cdots, \epsilon_N$ respectively. The necessary and sufficient condition for

$$\epsilon(B_1 \times \cdots \times B_N) = \sum_{i=1}^{N} (\epsilon_i - 1) + 1 = n$$

is that for every B_i, there exists a submatrix of n columns which reduces, by repeated applications of property 1, to a submatrix of the Odlyzko type with parameters n and ϵ_i.

Proof:

Assume first that $\epsilon(B_1 \times \cdots \times B_N) = n$.

We focus our proof on matrix B_1, the others follow in a similar manner.

Let

$$B_1 = \begin{pmatrix} b_{11} & \cdots & b_{1s} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ b_{r1} & \cdots & b_{rs} \end{pmatrix}$$

We can write $B_1 \times \cdots \times B_N$ as

$$B_1 \times \cdots \times B_N =$$

$$\begin{bmatrix} b_{11}(B_2 \times \cdots \times B_N) & \cdots & b_{1s}(B_2 \times \cdots \times B_N) \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ b_{r1}(B_2 \times \cdots \times B_N) & \cdots & b_{rs}(B_2 \times \cdots \times B_N) \end{bmatrix}$$

We will call

$$[b_{i1}(B_2 \times \cdots \times B_N) \cdots b_{is}(B_2 \times \cdots \times B_N)]$$

the i-th row block and

$$\begin{bmatrix} b_{1j}(B_2 \times \cdots \times B_N) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ b_{rj}(B_2 \times \cdots \times B_N) \end{bmatrix}$$

the j-th column block of $B_1 \times \cdots \times B_N$.

We are given a set of n columns such that there is a 1 in each row. We pick n columns of B_1 in the following manner. For every column in the set, we locate the column block that contains it. If the column is contained in block i, say, we take column i of B_1. Thus we can pick n columns, but at this point, there may be repeated columns among them.

We now prove that for the n columns of B_1, the row sums are at least $n - \epsilon_1 + 1$. If there is a row, say row 1, with row sum less than $n - \epsilon_1 + 1$, say α, we take the α columns which have a 1 in row 1. This in turn will give us α columns of the original n columns which gave us the 1-width of $B_1 \times \cdots \times B_N$. We focus our attention to the first row block. These α columns give us a 1 in each of the rows of the first row block. The same α columns will give us a 1 in each row for $B_2 \times \cdots \times B_N$, if we only take the first row block as $B_2 \times \cdots \times B_N$ copies many times (as many as the first row sum of B_1). Thus we have

$$\epsilon(B_2 \times \cdots \times B_N) \leqq \alpha < n - \epsilon + 1 = \sum_{i=2}^{n} (\epsilon_i - 1) + 1$$

contradicting Theorem 2. Therefore we have proved that the row sums of the chosen in columns of B_1 are all greater than or equal to $n - \epsilon_1 + 1$. In other words, there are at most $(\epsilon_1 - 1)$ zeros in each row.

We will now show that in these n columns of B_1 and for every possible choice of $(\epsilon_1 - 1)$ positions out of n, there is a row with $(\epsilon_1 - 1)$ zeros in these $(\epsilon_1 - 1)$ places. If this is not true, there will be a choice of $(\epsilon_1 - 1)$ columns with at least one 1 in each row. Then, the 1-width of B_1 is less than ϵ_1, a contradiction. This fact, by the way, also proves that there are no repeated columns among the n columns.

So, we have now n columns of B_1, and

$$\binom{n}{n - \epsilon_1 + 1}$$

rows of these columns with $(\epsilon_1 - 1)$ zeros in all the possible

$$\binom{n}{\epsilon_1 - 1} = \binom{n}{n - \epsilon_1 + 1}$$

positions. Let us call this submatrix B'. This is a submatrix of Odlyzko type.

In those n columns, if there were any other rows not yet contained in B', the row sum must still be at least $(n - \epsilon_1 + 1)$. Since we have all the possible

$$\binom{n}{n - \epsilon_1 + 1}$$

choices already, the extra row will cover one of the

$$\binom{n}{n - \epsilon_1 + 1}$$

rows. Thus by property 1, it can be removed without affecting the 1-width of the matrix.

Thus, we have proved half of the theorem.

We now assume that $B_1, \cdots, B_N$ contains a submatrix satisfying the conditions in the statement of the theorem.

We first show that the rows removed will not affect the 1-width of the product. Let us assume that the n columns are the first n columns of B_1. As before, we write

$$B_1 \times \cdots \times B_N = \begin{bmatrix} b_{11}(B_2 \times \cdots \times B_N) \cdots b_{1n}(B_2 \times \cdots \times B_N) \cdots b_{1s}(B_2 \times \cdots \times B_N) \\ \vdots \\ b_{r1}(B_2 \times \cdots \times B_N) \cdots b_{rn}(B_2 \times \cdots \times B_N) \cdots b_{rs}(B_2 \times \cdots \times B_N) \end{bmatrix} \tag{6}$$

Say, if row r covers row 1 in B_1, then the columns that cover the first row block will cover the r-th row block in Eq. (6). Hence it is enough just to consider the product of the submatrices of Odlyzko type. He proved that the 1-width of the product matrix is n. Anyway, if we write the n columns of B_i as $\{B_{i1}, \cdots, B_{in}\}$, it is not difficult to prove that the n columns given by

$$\prod_{i=1}^{N} B_{i1}; \cdots; \prod_{i=1}^{N} B_{in}$$

will have at least one 1 in each row. Q.E.D.

IV. Conditions for Achievement of Upper Bound

We will now study the conditions when the upper bound

$$\epsilon(A \times B) = \epsilon(A) \times \epsilon(B) \tag{7}$$

is obtained.

Definition:

A (0, 1)-matrix A of size r by s is said to satisfy the *minimal condition* if its 1-width is also the value of the following linear program:

minimize

$$\sum_i y_i$$

subject to the conditions

$$\left.\begin{aligned} AY &\geqq \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ Y &\geqq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \end{aligned}\right\} \tag{8}$$

where $Y = (y_1, \cdots, y_s)^T$ is a column vector.

This definition is not as strange as it first seems. The 1-width of A is just the same linear program with the extra condition that components of Y are either 0 or 1. The minimal condition just states that among the solutions to the linear program, one can find a solution with the components of Y either a 0 or 1.

The following gives a sufficient condition for Eq. (7) to be true.

Theorem 4

If a (0, 1) matrix A satisfies the minimal condition then $\epsilon(A \times B) = \epsilon(A) \times \epsilon(B)$ for any (0, 1)-matrix B.

Proof:

Let

$$A = \begin{pmatrix} a_{11} \cdots a_{1s} \\ \vdots \\ a_{r1} \cdots a_{rs} \end{pmatrix}$$

and we write

$$A \times B = \begin{pmatrix} a_{11}B \cdots a_{1s}B \\ \vdots \\ a_{r1}B \cdots a_{rs}B \end{pmatrix}$$

We will define the i-th column block and j-th row block in the same way as before.

Suppose we have a set of columns of $A \times B$ that gives the 1-width, and suppose that x_i columns come from the i-th column block. Furthermore, assume that

$$\sum_{i=1}^{s} x_i$$

which is the 1-width of $A \times B$, is smaller than $\epsilon(A) \times \epsilon(B)$.

Consider, say, the first row of A,

$$\sum_{i=1}^{s} a_{1i}x_i$$

gives the number of columns of B picked up in the first row block. This number must be greater than the 1-width of B. Hence we have

$$A\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix} \geqq \begin{pmatrix} \epsilon(B) \\ \cdot \\ \cdot \\ \cdot \\ \epsilon(B) \end{pmatrix}$$

or

$$\frac{1}{\epsilon(B)} A\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix} \geqq \begin{pmatrix} 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$$

or

$$A\left(\frac{1}{\epsilon(B)}\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix}\right) \geqq \begin{pmatrix} 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$$

Since $\epsilon(A \times B) < \epsilon(A) \times \epsilon(B)$, we have

$$\sum_{i=1}^{s} x_i < \epsilon(A) \times \epsilon(B)$$

Thus

$$\frac{1}{\epsilon(B)} \sum_{i=1}^{s} x_i < \epsilon(A)$$

which implies that A does not satisfy the minimal condition.

Q.E.D.

In a way, the converse of the above theorem is true.

Theorem 5

If a (0, 1)-matrix A does not satisfy the minimal conditions, then there exists a (0, 1)-matrix B such that

$$\epsilon(A \times B) < \epsilon(A) \times \epsilon(B)$$

Moreover, B can be chosen so that it is symmetric, square, and has 1's in its main diagonal.

First of all, we will state the following lemma. The result is given in Ref. 3. It is also not difficult to construct the matrix.

Lemma

Given positive integers n and c with $n > c$, a symmetric (0, 1)-matrix E exists with the following properties:

(1) The order of E is at least n.

(2) E has 0s in the main diagonal.

(3) E has constant line sum c.

(4) E has no repeated rows.

Proof (of Theorem 5):

We need only construct a matrix B such that

$$\epsilon(A \times B) < \epsilon(A) \times \epsilon(B)$$

Since A does not satisfy the minimal condition, there exists a vector

$$\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix}$$

satisfying Eq. (8) such that

$$\sum_{i=1}^{s} x_i < \epsilon(A) \tag{9}$$

The vector

$$\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix}$$

can be obtained by the simplex method in linear programming. All the x_i's will be rational numbers. Hence there exists an integer k such that kx_i is an integer for all i, and

$$\sum_{i=1}^{s} kx_i < k\epsilon(A) \tag{10}$$

We will construct B such that $\epsilon(B) = k$, and

$$\epsilon(A \times B) \leqq \sum_{i=1}^{s} kx_i$$

($= N$, say). It will have the properties promised by the theorem.

We first construct a $(0, 1)$-matrix E as promised in the lemma, with $n = N$ and line sum $c = k - 1$. The order of E will be M. In fact we can choose M as big as we like as long as $M \geqq n$. We will take $M > 3k - 4$ as well as $\geqq N$.

Now we take $(J - E)$. This is again a symmetric matrix with 1's in the main diagonal and $(k - 1)$ zeros on each row. Moreover, no two of the rows will have all its zeros in the same $(k - 1)$ places. So, the matrix $(J - E)$ is just a submatrix of an Odlyzko matrix with parameters M and k. We now construct B as

$$B = \overset{M}{\overbrace{\quad}} \left\{ \left[\begin{array}{c|c} J - E & P^T \\ \hline P & I \end{array} \right] \right\} \left(\begin{matrix} M \\ k-1 \end{matrix} \right)$$

where P is the remaining rows of the Odlyzko matrix. Clearly B is a symmetric and square matrix. The first M columns of B has 1-width k. So,

$$\epsilon(B) \leqq k$$

Suppose the 1-width is less than k, then, there are $(k-1)$ columns that have one 1 in each row. At most $(k - 2)$ of these columns can be in the first M, or else it contradicts the fact that the 1-width of the first M columns is k. Considering the first M columns only, we will have at least $M - (k - 2)$ rows with no 1's in them. These rows had better be covered by the remaining columns. In the submatrix $(J - E)$, there are $(k - 1)$ zeros in each column. So there are at least $M - (k - 2) - (k - 1) = M - (2k - 3)$ rows that have no 1's in the part P. We can at most cover up $(k - 1)$ rows with columns from I. So, if we choose M so large that

$$M - (2k - 3) > (k - 1)$$

or

$$M > 3k - 4$$

we will have a contradiction. This explains the choice of M.

The last thing left to be proved is that

$$\epsilon(A \times B) \leqq \sum_{i=1}^{s} kx_i$$

To establish this, we only need to consider $A \times B'$ where B' is the first M columns of B. We write

$$A \times B' = \begin{bmatrix} a_{11}B' & \cdots & a_{1s}B' \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ a_{r1}B' & \cdots & a_{rs}B' \end{bmatrix}$$

We now pick $\sum_{i=1}^{s} kx_i$ columns in the following manner:

column block number	number of columns picked	starting at
1	$k \times 1$	1
2	$k \times 2$	$kx_1 + 1$
3	$k \times 3$	$k(x_1 + x_2) + 1$
.	.	
.	.	
.	.	
s	$k \times s$	$k(x_1 + x_2 + \cdots + x_{s-1}) + 1$

Since the vector $\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix}$ satisfies Eq. (7), we have

$$Ak \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_s \end{pmatrix} \geqq \begin{pmatrix} k \\ \cdot \\ \cdot \\ \cdot \\ k \end{pmatrix}$$

This means we have picked at least k different columns of B' in each row block of $A \times B'$. Thus the 1-width of $A \times B'$ is at most

$$\sum_{i=1}^{s} kx_i$$

In turn, it means

$$\epsilon(A \times B) \leqq \sum_{i=1}^{s} kx_i$$

Q.E.D.

V. Acknowledgments

I would like to thank R. J. McEliece for proposing these problems and for his help. I would also like to thank H. J. Ryser for many stimulating conversations.

References

1. Odlyzko, A. M., "Data Storage and Data Compression," in *The Deep Space Network Progress Report*, Technical Report 32-1526, Vol. VI, pp. 112–117, Jet Propulsion Laboratory, Pasadena, Calif., Dec. 15, 1971.

2. Ryser, H. J., *Combinatorial Mathematics*, Wiley, New York, 1963.

3. Fulkerson, D. R., Hoffman, A. J., and McAndrew, M. H., "Some Properties of Graphs with Multiple Edges," *Canadian J. Math.*, Vol. 17, pp. 166–177, 1965.