On Two-Dimensional Arrays and Crossword Puzzles

Jack Keil Wolf
Center for Magnetic Recording Research
Mail Code: 0401
University of California-San Diego
La Jolla, CA 92093-0401
jwolf@ucsd.edu

Paul Siegel
Dept. of Electrical & Computer Engineering
Mail Code: 0407
University of California-San Diego
La Jolla, CA 92093-0407
psiegel@ ucsd.edu

Abstract.

We consider large two-dimensional arrays where each row and each column must satisfy certain constraints. We relate this problem to an assertion made by Shannon regarding the existence of large two-dimensional crossword puzzles.

One-dimensional constrained sequences.

We begin by considering an alphabet containing M symbols and one-dimensional sequences of these symbols that satisfy a set of well defined constraints. To narrow the focus of this paper we will concentrate on a special class of such one-dimensional constrained sequences which consist of concatenations of isolated words from some dictionary. Details of this special class follow.

We first assume that one of the M symbols is of special significance and is called the space symbol. The collection of the remaining (M-1) symbols will be called non-space symbols. We assume that we have a dictionary of Q allowable words. The words in this dictionary are assumed to be sequences of the (M-1) non-space symbols. We assume that the i-th word of the dictionary is of length L_i and that all the words in the dictionary are distinct. Finally we assume that the constraints allow for every concatenation of words from the dictionary provided that consecutive words are separated by one (or in some cases more than one) space symbol.

One example of such a constrained system would be the sequences of English words in the rows or columns of a crossword puzzle as contained in many American newspapers or magazines. Here M=27, the black space between words is considered the space symbol and there are 26 non-space symbols. In this case, we allow for one or more than one space symbol between words in the dictionary to allow for one or more than one black square between consecutive words.

A second example of such a constrained system would be binary (d,k) sequences. The alphabet consists of the two symbols 0 and 1. The usual description of such binary sequences is that the symbols must satisfy the following two constraints.

(1) d-constraint: Two consecutive 1's must be separated by a run of at least d consecutive 0's.
(2) k-constraint: The maximum length of a run of consecutive 0's is k.

Note, however, the following equivalent description of such (d,k) binary sequences which applies when $d \geq 1$. The M=2 binary symbols are the space symbol "1" and the non-space symbol "0". The dictionary contains the Q=(k-d+1) code words consisting of runs of j 0's where j takes on the (k-d+1) values from d to k. Furthermore, every word in the dictionary can be followed by exactly one space symbol (i.e., one "1"). The modifications to this description that must be made for the case of d=0 are straightforward.

If any concatenation of the M symbols is allowable, there are exactly M^n distinct sequences of length n. It is desirable to know how many distinct sequences of length n can be formed for the constrained system. Shannon [1] provided us with several methods of computing this quantity. One method is very easy to understand in terms of our dictionary description of these constrained sequences.

Assume first the case where there is one and only one space between adjacent code words. Let N(n) be the number of distinct sequences of length n in our constrained system. Note that any sequence of length n can be thought of as a sequence of length (n-j-1) followed by a space followed by a code word of length $j \geq 1$. Then, if N(n-j-1) is the number of distinct sequences of length (n-j-1) and if A_j denotes the number of distinct words in our dictionary of length j, then the number of distinct sequences of length N(n) would equal the sum over all code word lengths j of the product of A_j and N(n-j-1).

From the standard theory of linear difference equations one knows that the solution for N(n) is a sum of exponentials of the form γ^n which for large n is dominated by the largest real root of an algebraic equation found by substituting γ^j for N(j) in the original linear difference equation.

For the case of binary (d,k) codes, the value of the limit as n approaches infinity of $\log_2[N(n)]/n$ has been computed by many authors. We will denote this quantity by $C_1(d,k)$. $C_1(d,k)$ has been referred to as the capacity (or the maximum entropy) of the constrained system. For our purposes, we need only know that $C_1(d,k)$ can be computed for all values of d and k and that the value of $C_1(d,k)$ is non-zero for all d and k provided that k > d. One more fact that will be useful later is that $C_1(1,2) = C_1(2,4)$.

For the case of an arbitrary concatenation of English words as in the rows and columns of a crossword puzzle, one must take into account the fact that one can have more than one space between words. In that case, the above-mentioned difference equation is modified by adding the additional term N(n-1) to the sum of products of A_j and N(n-j-1). The method of solution, however is identical to that described previously.

For this case, the value of the limit as n approaches infinity of $\log_2[N(n)]/n$ has been computed by Gilbert [2] by counting the number of English words of each length in two different English dictionaries. Calling this quantity C_1(English), Gilbert computed C_1(English) to be 2.444 bits/letter for a dictionary of size 34,897 and computed C_1(English) to be 2.782 bits/letter for a dictionary of size 233,614. Again, C_1(English) has been referred to as the capacity or the maximum entropy of the constrained system.

Two-dimensional constrained arrays and crossword puzzles.

We next consider two-dimensional arrays of symbols from an alphabet of size M where the arrays are such that every row and every column must individually satisfy the previously described constraint for one-dimensional sequences. That is, every row and every column of the array must be a concatenation of words from a dictionary where adjacent words are separated by one or possibly more than one space. Two examples of such arrays which are the two-dimensional extension of the one-dimensional examples are:

1. Two-dimensional crossword puzzles as found in many American newspapers or magazines where every row and every column consists of a sequence of words from an English dictionary and such that every pair of consecutive words in every row and every column are separated by one or more spaces (i.e., black squares).

2. Binary two-dimensional (d,k) arrays where each element of the array is a "0" or a "1" and such that every row and every column satisfies the previously specified d-constraint and k-constraint.

For a specified constraint we are interested in how many such distinct arrays can be constructed for an array with m rows and n columns. We call this number N(m,n).

It is clear that for large m and n, for every constraint for which the one-dimensional capacity is non-zero, N(m,n) must grow at least exponentially with m or with n. This is the case since if one chooses the first column (or row) to satisfy the one-dimensional constraint, one can always have the rows (or columns) satisfy the constraint by choosing each subsequent column (or row) as a shift by one symbol up or down (or to the left or right) of the previous column (or row).

However, one might suspect that in some cases, that for large m and n, N(m,n) might grow exponentially with the product of m and n. Indeed let us define the two-dimensional capacity or maximum entropy of the array as the limit as m approaches infinity and as n approaches infinity of $\log_2[N(m,n)]/mn$. For the case of two-dimensional crossword puzzles of English words we call this quantity C_2(English), while for the case of two-dimensional binary (d,k) arrays we call this quantity C_2(d,k).

It is of specific interest to us that Shannon discussed two-dimensional crossword puzzles of English letters in his 1948 paper [1]. In part, Shannon stated:

> "The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its *relative entropy*."
>
> "One minus the relative entropy is the *redundancy*."
>
> "The redundancy of a language is related to the existence of crossword puzzles. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just

possible when the redundancy is 50%. If the redundancy is 33%, three dimensional crossword puzzles should be possible, etc."

No further explanation of these remarks has been found in the subsequent publications of Shannon nor in the work of other authors. As a result of one of the authors having presented some of this material at the Shannon Day Symposium at Bell Laboratories on May 18, 1998, a letter and other correspondence [2] was received from Edgar Gilbert reporting on some previous work that he had done on the subject and on his recollection of some discussions he had had with Shannon related to this work. The following argument was completed before having received the correspondence from Gilbert but seems to be corroborated by its contents.

In order to proceed we will make some conjectures as to how to interpret Shannon's statements. We first conjecture that Shannon's words "large crossword puzzles" are "possible" should be interpreted to mean that the two-dimensional capacity of the system is strictly greater than 0. Conversely, when Shannon states that, if the redundancy is too high, large crossword puzzles are not possible, we assume that Shannon meant that in that case the two-dimensional capacity is equal to 0.

We go on to conjecture that Shannon's description of the constraints as being of a "chaotic and random nature" refers to the fact that the joint statistics that govern the entries of the array are such that all of the rows and all of the columns are statistically independent of one another.

We now give a plausibility argument for Shannon's conclusion based upon these conjectures. We first note that the number of distinct arrays having m rows and n columns with elements from an alphabet with M symbols is equal to M^{mn}. For crossword puzzles made up of English words, M=27. There are N(n) distinct sequences of length n (or N(m) sequences of length m) that satisfy the constraint. For any of the M^{mn} arrays previously constructed, the probability that any row satisfies the constraint is $N(n)/M^n$, and the probability that any column satisfies the constraint is $N(m)/M^m$. Invoking our conjecture as to the meaning of "chaotic and random", the probability that every row and every column of the array satisfies the constraints is $[N(m)]^n[N(n)]^m/M^{2mn}$. Under this assumption, the average number of arrays for which all of the rows and all of the columns satisfy the constraints is equal to $[N(m)]^n[N(n)]^m/M^{mn}$.

Assuming that we take m and n very large, and denote by C the appropriate capacity or maximum entropy, we can use the approximation $N(m)=2^{mC}$ and $N(n)=2^{nC}$. Then writing $M=2^{\log(M)}$ (where the logarithm is taken base 2), we have that the average number of arrays having all of their rows and all of their columns satisfying the constraints is equal to $2^{mn(2C-\log(M))}$. Thus if $2C-\log(M) > 0$ (or equivalently $C/\log(M) > 0.5$), the average number of arrays that have all rows and all columns satisfying the constraints grows exponentially with the product mn. A similar argument follows for higher dimensional arrays. This agrees with Shannon's conclusions.

We can improve on this result by first choosing arrays that satisfy only the constraints on the rows. Using the above argument, the average number of arrays that have all of their m rows satisfying the constraints is equal to $[N(n)]^m$. These arrays will have the symbols in each column occurring in accordance with their first order probability. Let H1 be the entropy corresponding to their first order probability. The probability that any column in these

surviving arrays satisfies the constraints is then for large m approximately given by $N(m)/2^{mH1}$. The probability that all n columns of these surviving arrays satisfies the constraints is then given by $[N(m)]^n/2^{mnH1}$. Thus the average number of arrays that will have both their rows and columns satisfying the constraints is given by $2^{mn(2C-H1)}$. Thus if 2C-H1 > 0 (or equivalently C/H1 > 0.5), the average number of arrays that have all rows and all columns satisfying the constraints grows exponentially with the product mn. Again a similar argument follows for higher dimensional arrays. Specifically, this argument says that large three-dimensional crossword puzzles exist provided that C/H1 > 0.666.

Using the definition of redundancy as stated by Shannon and Gilbert's estimate of the maximum entropy of English words for the 233,614 word dictionary, one finds that the redundancy of English words is about 0.415. Following Shannon, this would allow for the construction of large two-dimensional crossword puzzles but not for large three-dimensional crossword puzzles.

Ignoring the allowed multiplicity of spaces between words, the first order entropy of English is known to be approximately 4.03 bits per symbol. Using Gilbert's estimate of the entropy of English words for the 233,614 word dictionary one finds that C/H1 = 0.690. Thus, large three-dimensional crossword puzzles of English words indeed may exist.

Two-dimensional (d,k) arrays.

We are on somewhat shaky ground trying to apply the previous ideas to binary (d,k) arrays since these constraints are far from being of a chaotic and random nature. However, it is of some interest to see what Shannon's ideas predict in that case.

Very little is known about the two-dimensional capacity, $C_2(d,k)$ of binary (d,k) arrays. In fact the exact value of capacity, $C_2(d,k)$, has only been computed for the case when $C_2(d,k)=1$ or $C_2(d,k)=0$. The former case is the trivial case of unconstrained arrays (i.e., d=0 and k=infinity). The latter case is known to occur [3] if and only if $d \geq 1$ and k=d+1. The case of $C_2(d,k) =0$ for $d \geq 1$ and k=d+1 is predicted by Shannon's theory since $C_1(1,2) = 0.4057$ and $C_1(d,d+1) \leq .4057$ for d > 1. Thus for these cases, the redundancy is greater than 0.5 so one would not expect to be able to construct large two-dimensional arrays satisfying these constraints.

As stated previously, $C_1(2,4)= C_1(1,2)$ but also as stated previously $C_2(1,2)=0$ and $C_2(2,4) > 0$. Thus it is clear that Shannon's argument cannot be applied since both constraints have the same redundancy but one has a two-dimensional capacity equal to 0 and the other has a non-zero two-dimensional capacity. Our hope was that the modified argument might predict the correct behavior by taking into account the fact that the first order statistics for the two constraints are not the same. Indeed the (2,4) constraint has a higher value for C/H1 than does the (1,2) constraint but unfortunately both have C/H1< 0.5. Thus our initial skepticism in applying these ideas to this constraint was correct.

Concluding comments.

It is clear that there is much to be learned regarding two-dimensional constrained arrays. One possible avenue for future research is to attempt to improve further on the argument given related to the existence of large two-dimensional arrays. In particular, it is clear that the rows

and columns of binary two-dimensional (d,k) arrays are not statistically independent so that one might attempt to model the dependence in some manner.

Acknowledgments.

The authors are indebted to Edgar Gilbert for his many contributions and to Kevin Shaughnessy who contributed to the initial understanding of Shannon's statements. This work was partially supported by the Center for Magnetic Recording Research at the University of California, San Diego and by the National Science Foundation under grants NCR-9405008 and NCR-9612802.

References

[1 C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 10, pp. 379-423, 623-656, October, 1948.

[2] E.N. Gilbert, private communication, 1998.

[3] A. Kato and K. Zeger, "On the capacity of two-dimensional run-length limited codes," *Proceedings of the 1998 International Symposium on Information Theory*, pg. 320, Cambridge, MA, August, 1998.