Counting Crosswords

David J.C. MacKay Jeremy Thorpe

June 30, 2004 – Draft 2.3

Abstract

Shannon's calculation of the number of crosswords assumed that the rows and columns of crosswords contain *typical* strings from the language.

However, in most languages, most crosswords will have rows and columns that are atypical. This atypicality modifies the way in which we count the number of crosswords.

▶ 1 Introduction

Shannon (1948) observed that large numbers of large crosswords can be constructed if a language has sufficiently low redundancy: large two–dimensional crosswords are possible, he said, if the entropy per character H_W of the language consisting of words separated by spaces satisfies

$$H_W > \frac{1}{2} H_{\max}, \tag{1}$$

where $H_{\max}$ is the maximum achievable entropy per character.

This observation of Shannon's was elaborated by Wolf and Siegel (1998). They interpreted Shannon's assertion that 'large two–dimensional crosswords are possible' as meaning that 'the number of valid crosswords grows exponentially with the number of squares S in the grid'. They counted the number of valid crosswords in a language by counting the number of *typical* ways of filling in the rows of a grid, then evaluating the probability that one such filling-in also has valid and typical columns. They reproduced Shannon's results, and augmented them by supplying a tighter condition, applicable to languages in which letters are not used with equal frequency.

The problem with Shannon's result (1), you see, is that if we switch language from English (in which we know two-dimensional crosswords are possible) to 'English plus a few foreign words that use new characters not included in the original English alphabet', then certainly all the crosswords we made in English are still valid crosswords in the new language, but the inequality (1) may well be violated, since adding extra characters to the alphabet increases $H_{\max}$ on the right hand side.

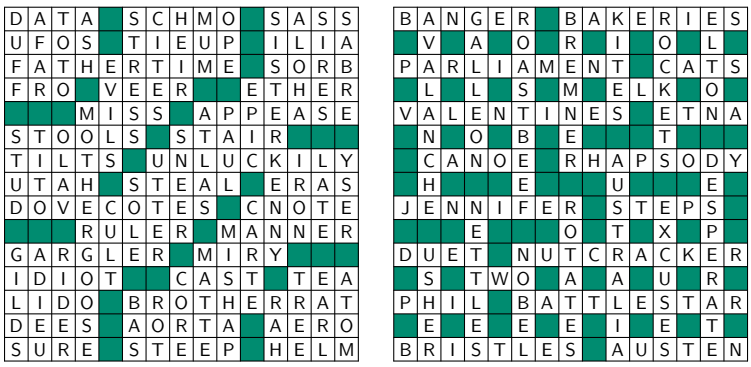

Figure 1. Crosswords of types A (American) and B (British).

In a 'type A' (or American) crossword, every row and column consists of a succession of words of length 2 or more separated by one or more spaces. In a 'type B' (or British) crossword, each row and column consists of a mixture of words and single characters, separated by one or more spaces, and every character lies in at least one word (horizontal or vertical). Whereas in a type A crossword every letter lies in a horizontal word *and* a vertical word, in a typical type B crossword only about half of the letters do so; the other half lie in one word only.

Type A crosswords are harder to *create* than type B because of the constraint that no single characters are permitted. Type B crosswords are generally harder to *solve* because there are fewer constraints per character.

So Wolf and Siegel derived a tighter condition for large two-dimensional crosswords to be possible,

$$H_{\mathrm{W}} > \frac{1}{2} H_0, \tag{2}$$

where H_0 is the entropy of the monogram distribution of the language.

Wolf and Siegel's calculations were modified in (MacKay, 2003, Section 18.1) so as to give conditions not only for the 'A' type of crossword but also for the 'B' type (figure 1). MacKay weakened Wolf and Siegel's calculations by assuming the language consisted of W words all of the same length L. Using typicality arguments similar to those of Wolf and Siegel, the conditions for large crosswords of the two types to be possible were found to be as follows, where H_0 is the entropy of the monogram distribution for *non-space* characters, and the entropy of the language consisting of arbitrary strings of words is

$$H_W \equiv \frac{\log_2 W}{L+1}. \tag{3}$$

Crossword type	A	B
Condition for crosswords	$H_W > \frac{1}{2}\frac{L}{L+1}H_0$	$H_W > \frac{1}{4}\frac{L}{L+1}H_0$

If we set $H_0 = 4.2$ bits and $L = 5$ then we can estimate how big a vocabulary

is required for crosswords of the two types to be numerous: for type A, $W \gg 2^{LH_0/2} \simeq 1500$; and for type B, $W \gg 2^{LH_0/4} \simeq 40$.

These figures seem consistent with experience: we can easily write children's crosswords of type B, but most crosswords of type A contain more obscure words.

However, these calculations of the number of valid crosswords underestimate this number by counting only 'typical' crosswords. It is likely that *atypical* fillings-in of the crossword dominate the true count.

We now give an example illustrating the inaccuracy of the condition (2), then give a new calculation of the number of crosswords, assuming a simple language model.

We will use the following notation:

W	Number of words in dictionary
L	typical word length
$\mathbf{p}$	monogram distribution for non-space characters (letters)
S	number of squares in crossword
$f_1 S$	number of letter squares
$f_w S$	number of words in the crossword
S_1	number of letter squares whose letters appear in one word only
S_2	number of letter squares whose letters appear in two words
	Note $f_1 S = S_1 + S_2$

In crosswords of type A in which the typical word length is L, typical values of f_w, f_1, S_1, and S_2 are as follows:

	A	B
f_w	$\frac{2}{L+1}$	$\frac{1}{L+1}$
f_1	$\frac{L}{L+1}$	$\frac{3}{4}\frac{L}{L+1}$
S_1	0	$\frac{1}{2}\frac{L}{L+1}S$
S_2	$\frac{L}{L+1}S$	$\frac{1}{4}\frac{L}{L+1}S$

▶ 2 A counterexample to condition (2)

Consider a language with 514 characters, of which two, '0' and '1', have probability 1/4 each, and the other 512 characters have probability 1/1024 each. The entropy of this monogram distribution is $H_0 = 6$ bits. If the dictionary of the language contains $W = 2^{12} \simeq 4000$ words all of length $L = 5$ then the entropy per character of the language is $H_W = 2$ bits. Given this letter entropy H_0 and language entropy H_W, our conditions for crosswords expect neither type of crossword to be possible. But in fact, the Wenglish dictionary will almost

certainly contain almost all the thirty-two possible five-letter binary strings, 00000, 00001, 00010, ..., 11111 (since each had a probability of 2^{-10} of occurring when a new word was added to the dictionary, and there are 2^{12} words). So all the $2^{f_1 S}$ atypical crosswords that contain exclusively the characters 0 and 1 are almost certainly valid croswords. Thus an exponentially large number of crosswords *do* exist for this language, albeit atypical crosswords dominated by just 32 crossword-friendly words from the dictionary.

Now, Shannon could defend his calculation by saying 'I'm not interested in atypical crosswords, I only want to count crosswords in which the rows and columns are typical of the language – here, crosswords that use the full dictionary in a balanced manner'. However, we suspect that in real life, crosswords are indeed populated by an atypical distribution that favours crossword-friendly words. We therefore offer a new calculation that allows the words that succeed in making crosswords to be atypical. Our calculation makes a testable prediction about this atypicality.

▶ 3 A new calculation

Imagine that a language is made by creating a dictionary of W words, all of length L, from a monogram distribution $\mathbf{p}$. We count the number of two-dimensional crosswords of S squares by assuming that an appropriate grid of black and white squares has been made, and evaluating the probability that each possible in-filling is valid.

(In contrast to the calculation in (MacKay, 2003, Section 18.1), we do not restrict attention to 'typical' in-fillings.)

We denote the dictionary by $\{\mathbf{d}^{(w)}\}_{w=1}^{W}$; w runs over the words in the order that they were created; the lth letter of the wth word is $d_l^{(w)}$.

We denote a candidate crossword by the vector $\mathbf{X}$ whose components are x_s, with s running from 1 to $f_1 S$. From the vector $\mathbf{X}$ we can extract the tentative words $\mathbf{x}^{(n)}$, where n runs from 1 to $N = f_w S$. Each word $\mathbf{x}^{(n)}$ consists of L components from $\mathbf{X}$.

To understand our calculation, imagine that all possible in-fillings $\mathbf{X}$ are created before the dictionary is generated; we then ask, 'what is the probability that in-filling $\mathbf{X}$ will turn out to be valid?' To help us answer this question, we introduce a *key*, $w(n)$, which is a putative mapping from words in the grid n to words in the dictionary w. If an in-filling *is* valid, then there exists a key $w(n)$ such that, for each n

$$x_l^{(n)} = d_l^{(w(n))}, \text{ for all } l. \tag{4}$$

We denote the key by $\mathcal{W}$. The number of possible keys for a grid containing N words is W^N. We can then count the expected number of valid crosswords, Ω, as follows.

$$\Omega \simeq \sum_{\mathbf{X}} \sum_{\mathcal{W}} \prod_n \prod_l P(d_l^{(w(n))} = x_l^{(n)}) \tag{5}$$

This count is approximate because we are overcounting in cases where there are multiple valid keys (*e.g.*, a particular word appears twice in the dictionary) and the calculation is inaccurate if any particular dictionary word is used twice in the grid, because it treats as independent events that are not.

The probability $\prod_l P(d_l^{(w(n))} = x_l^{(n)})$ is the probability, when the dictionary comes to be generated, that the word $\mathbf{d}^{(w(n))}$ will exactly match the word $\mathbf{x}^{(n)}$ defined by the in-filling, $\mathbf{X}$.

OK so far?

Since the dictionary is being generated from a monogram distribution $\mathbf{p}$,

$$P(d_l^{(w(n))} = x_l^{(n)}) = p_{x_l^{(n)}}. \tag{6}$$

Now, in the monster sum-product (5), each component x_n of $\mathbf{X}$ is mentioned in an expression of the form $P(d_l^{(w(n))} = x_l^{(n)})$ either twice or once, if it appears in two words or just one, respectively.

Our sum-product is about to simplify. For components that appear in just one word, the simplification involves a factorization $(\sum_{x_l^{(n)}} p_{x_l^{(n)}}) = \sum_i p_i = 1$. For components that appear in two words, the two events of the form $P(d_l^{(w(n))} = \cdots)$ must involve the random dictionary producing the *same* character in two distinct words; we thus obtain a factor

$$\sum_i p_i^2. \tag{7}$$

We sum over $\mathbf{X}$ first, for fixed $\mathcal{W}$. Let's call a key $\mathcal{W}$ non-colliding if all n map to distinct w under $w(n)$. We focus attention on non-colliding keys. Because the dictionary words are generated independently and identically, all non-colliding keys $\mathcal{W}$ yield exactly the same answer for the quantity

$$\sum_{\mathbf{X}} \prod_n \prod_l P(d_l^{(w(n))} = x_l^{(n)}) = \left(\sum_i p_i^2 \right)^{S_2}, \tag{8}$$

where S_2 is the number of intersection characters.

Assuming $W \gg N$, the number of non-colliding keys is approximately the number of keys, W^N, so the expected number of valid crosswords is:

$$\begin{aligned} \Omega &\simeq W^N \left(\sum_i p_i^2 \right)^{S_2} \\ &= W^{f_w S} \left(\sum_i p_i^2 \right)^{S_2}. \end{aligned}$$

(Strictly, we want to be free to send N to infinity, so the assumption that $W \gg N$ looks like it produces problems; what we could do here is try to introduce a better definition of 'non-colliding'.) Let's massage this into a form

that we can compare with the Wolf–Siegel expression for the number of valid crosswords,

$$\Omega_{\rm WS} \simeq 2^{(2H_W - H_0)S}. \tag{9}$$

We'll see that the factor $W^{f_w S}$ is analogous to $2^{2H_W S}$, and the curious factor $\left(\sum_i p_i^2\right)$ is analogous to 2^{-H_0}.

Before we carry out the massage, let's focus on the curious factor. The sum of squares $\left(\sum_i p_i^2\right)$ is masquerading as a replacement for the inverse-exponential-entropy, $\prod_i p_i^{p_i}$; this approximation,

$$\prod_i p_i^{p_i} \simeq \sum_i p_i^2, \tag{10}$$

is a surprisingly good one – for any 20–dimensional probability vector, for example, the approximation is accurate to within a factor of 2.5! To teachers of information theory, this approximation may be familiar, because it arises from a common error made when solving entropy inequalities.

Consider a student tasked with proving that the entropy $H(\mathbf{p})$ of an I–dimensional vector $\mathbf{p}$ is bounded above by $\log I$. He uses Jensen's inequality:

$$\sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i = -\langle \log p_i \rangle \geq -\log \langle p_i \rangle = -\log \left(\sum_i p_i^2 \right). \tag{11}$$

Oops! He has not proved that the entropy is bounded above by anything; instead he has proved that it is bounded *below* by a quantity known as the order-two Rényi entropy,

$$H^{(2)}(\mathbf{p}) \equiv -\log \left(\sum_i p_i^2 \right). \tag{12}$$

[The order-r Rényi entropy is

$$H^{(r)}(\mathbf{p}) = \frac{1}{1-r} \log \left(\sum_i p_i^r \right). \Bigg] \tag{13}$$

So, to conclude our calculation, let's define the Rényi monogram entropy per character (including spaces) by

$$H^{(2)}_{\rm Mono} \equiv \frac{L}{L+1} H^{(2)}(\mathbf{p}). \tag{14}$$

Then, using $H_W \equiv \frac{\log_2 W}{L+1}$ and the figures from the table at the end of section 1, we can rewrite our result (9) as two results, one for each type of crossword:

$$\Omega_A \simeq 2^{(2H_W - H^{(2)}_{\rm Mono})S} \tag{15}$$

$$\Omega_B \simeq 2^{(H_W - \frac{1}{4} H^{(2)}_{\rm Mono})S}. \tag{16}$$

Thus the conditions for there to be exponentially many crosswords become:

Crossword type	A	B
Condition for crosswords	$H_W > \frac{1}{2}H^{(2)}_{\rm Mono}$	$H_W > \frac{1}{4}H^{(2)}_{\rm Mono}$.

These conditions are pleasingly similar to Wolf and Siegel's, with the simple replacement of the entropy by the Rényi entropy.

▶ 3.1 Three-dimensional crosswords

The condition for d-dimensional American-style crosswords (in which every letter participates in d words) can be derived in the same way. The number of crosswords is

$$\begin{aligned}\Omega &\simeq W^N \left(\sum_i p_i^d\right)^{S_d} \\ &= W^{f_w(d)S} \left(\sum_i p_i^d\right)^{S_d} .\end{aligned}$$

We assume $f_w(d) = \frac{d}{L+1}$ and $S_d = \frac{L}{L+1}S$. We note that the order–d Rényi entropy has appeared:

$$\log\left(\sum_i p_i^d\right) = -(d-1)H^{(d)}(\mathbf{p}) \tag{17}$$

We define the order–d Rényi monogram entropy per character (including spaces) by

$$H^{(d)}_{\rm Mono} \equiv \frac{L}{L+1}H^{(d)}(\mathbf{p}). \tag{18}$$

Then

$$\Omega_A \simeq 2^{(dH_W - (d-1)H^{(d)}_{\rm Mono})S} \tag{19}$$

Thus the condition for there to be exponentially many d-dimensional type-A crosswords is:

Condition for crosswords	$H_W > \frac{d-1}{d}H^{(d)}_{\rm Mono}$.

This result is identical to Shannon's ('the redundancy must be less than $1/d$') if the Rényi monogram entropy is equal to the entropy of the uniform distribution.

▶ 4 A prediction

A spin-off of our calculation is a prediction about atypicality of words in crosswords. The prediction only applies to languages modelled by our 'random dictionary' model, and it's best tested in type–B crosswords in which there are characters of both types, intersecting and non-intersecting.

The prediction is that whereas the letters in non-intersecting squares are expected to have the same distribution as the source $\mathbf{p}$ that generated the dictionary, the letters at intersections of words are expected to come from the distribution

$$q_i \equiv \frac{p_i^2}{\sum_{i'} p_{i'}^2}. \tag{20}$$

We aim to test this prediction on a corpus of crosswords shortly.

▶ 5 Application to two-dimensional (d, k) constraints

Wolf (IEEE Information Theory Society Newsletter, September 2001) discusses the capacity $C_2(d, k)$ of two-dimensional binary arrays whose rows and columns satisfy the (d, k) constraint: after every `1` there must be at least d and at most k `0`s.

He observes that Shannon's condition for crosswords fails to predict correctly in all cases whether $C_2^{(d,k)} > 0$. In particular, it is known that $C_2^{(2,4)} > 0$ and $C_2^{(1,2)} = 0$; but Shannon's condition predicts that crosswords with neither $(d, k) = (2, 4)$ nor $(1, 2)$ exist. (Shannon's condition makes identical predictions for these two cases, because his condition depends only on a language's one-dimensional capacity, $C_1^{(d,k)}$; these two capacities happen to satisfy $C_1^{(2,4)} = C_1^{(1,2)} \simeq 0.4057$.) This conundrum is not solved by Wolf's analysis.

So, does our observation that most crosswords will have rows and columns that are atypical help resolve this conundrum?

The answer is 'almost'.

For each one-dimensional channel, we can introduce parameters $\mathbf{p}$ that control the transition probabilities between states. We can compute how the fraction of `1`s emitted depends on $\mathbf{p}$, $p_1(\mathbf{p})$; and compute the dependence of the one-dimensional capacity $C_1^{(d,k)}$ on $\mathbf{p}$. [We here extend the definition of capacity to allow dependence on $\mathbf{p}$; the true capacity is the maximum over $\mathbf{p}$ of $C_1^{(d,k)}(\mathbf{p})$.]

Now, imagine generating grids randomly filled with the fraction of `1`s in the grid being p_1. The number of such grids scales as

$$2^{SH_2(p_1)}, \tag{21}$$

where H_2 is the binary entropy function. The probability that all the rows of the grid are (a) valid according to the (d, k) constraint, and (b) typical of the parameters $\mathbf{p}$ that we introduced above, is

$$f = \frac{2^{SC_1^{(d,k)}(\mathbf{p})}}{2^{SH_2(p_1(\mathbf{p}))}}. \tag{22}$$

The probability that all the columns are also valid and typical is also f. So, neglecting inter-column correlation, we obtain the number of crosswords with

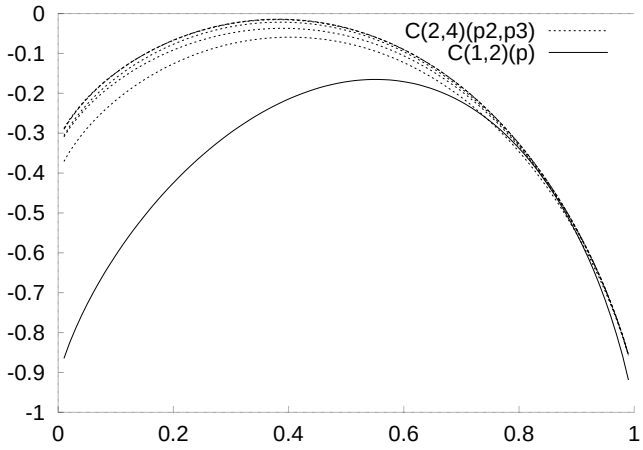

Figure 2. Graphs of the exponents $C_2^{(2,4)}(p_2, p_3)$ and $C_2^{(1,2)}(p)$ as the arguments p, p_2, and p_3 vary. For $C_2^{(2,4)}(p_2, p_3)$, the horizontal axis is p_2 and p_3 takes the values 0.375, 0.5, 0.535, 0.625, 0.75, which bracket the optimum. The maximum value of the exponent $C_2^{(2,4)}(p_2, p_3)$ is -0.014. For $C_2^{(1,2)}(p)$, the horizontal axis is p, the probability of emitting a `1` after a run of one zero.

parameters $\mathbf{p}$ by multiplying (21) by f^2.

$$\Omega(\mathbf{p}) \simeq \frac{2^{2SC_1^{(d,k)}(\mathbf{p})}}{2^{SH_2(p_1(\mathbf{p}))}} = \left(\frac{2^{2C_1(\mathbf{p})}}{2^{H_2(p_1(\mathbf{p}))}}\right)^S \tag{23}$$

which is an increasing or decreasing function of S with exponent

$$C_2^{(d,k)}(\mathbf{p}) = 2C_1^{(d,k)}(\mathbf{p}) - H_2(p_1(\mathbf{p})). \tag{24}$$

[Perhaps for clarity we should name this exponent $E_2^{(d,k)}(\mathbf{p})$ rather than $C_2^{(d,k)}(\mathbf{p})$, since it seems reasonable to reserve $C_2^{(d,k)}$ for the true, unknown two-dimensional capacity?]

When we compute the maximum value of this exponent, we find completely different values for the cases $(d, k) = (1, 2)$ and $(2, 4)$, and the exponent $C_2^{(d,k)}(\mathbf{p})$ is greater in the latter case. So we almost solve the conundrum. But only 'almost', because the maximum exponent found for $(d, k) = (2, 4)$ is still negative.

▶ 5.1 Details

We parameterize the one-dimensional $(2, 4)$ sequence generator by parameters p_2 and p_3, which are the probabilities of emitting a `1` after a run of two zeroes and three zeroes respectively. At equilibrium, the fraction of `1`s emitted is

$$p_1(p_2, p_3) = 1.0/(4 - p_2 + (1 - p_2)(1 - p_3)). \tag{25}$$

The one-dimensional capacity is

$$C_1^{(2,4)}(p_2, p_3) = p_1(p_2, p_3)H_2(p_2) + f_3(p_2, p_3)H(p_3) \tag{26}$$

where

$$f_3(p_2, p_3) = (1 - p_2)/(4 - p_2 + (1 - p_2)(1 - p_3)) \tag{27}$$

The exponent

$$C_2^{(2,4)}(p_2, p_3) = 2C_1^{(2,4)}(p_2, p_3) - H_2(p_1(p_2, p_3)) \tag{28}$$

is plotted in figure 2 for a range of values of p_2 and p_3, alongside the corresponding exponent $C_2^{(1,2)}(p)$ for the $(1, 2)$ channel.

Evidently, to compute the true capacity $C_2^{(2,4)}$, which is known to be at least $1/16$, we will have to take into account correlations between neighbouring rows and columns.

▶ 5.2 Taking into account correlations

We can use the exact same method on a set of two-by-two tiles that obey the rules of the (2,4) constraint.

We introduce 15 free parameters, solve for the principal eigenvector, and find the entropy of the tile distribution and the capacity of the one-dimensional channel. We then maximize the exponent. Our preliminary results establish that the capacity of the two-dimensional (2,4) array is ≥ 0.077. [The bound given by Wolf was $1/16 = 0.0625$.]

References

MacKay, D. J. C. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Sys. Tech. J.* **27**: 379–423, 623–656.

Wolf, J. K., and Siegel, P. (1998) On two-dimensional arrays and crossword puzzles. In *Proceedings of the 36th Allerton Conference on Communication, Control, and Computing, Sept. 1998*, pp. 366–371. Allerton House.