A conjecture on the alphabet size needed to produce all correlation classes of pairs of words

Paul Leopardi

Thanks: Jörg Arndt, Michael Barnsley, Richard Brent, Sylvain Forêt, Judy-anne Osborn.

Mathematical Sciences Institute, Australian National University.

For presentation at 34th Australasian Conference on Combinatorial Mathematics and Combinatorial Computing ANU, December 2010.

AUSTRALIAN RESEARCH COUNCIL
Centre of Excellence for Mathematics and Statistics of Complex Systems

ARC

Topics

- Analysis of the problem: missing words in a random string
- Word overlap correlations
- Enumeration of correlation classes
- The conjecture
- Other open problems

Analysis: missing words in a random string

We analyze the problem: find the distribution of the number of missing words in a random string.

Alphabet size is α, equally likely.

String length is N. Word length is T.

Words overlap. The string S contains $N - T + 1$ words.

There are α^N possible strings S_i, α^T possible words W_j.

Define indicator $v_{i,j} := 1 \Leftrightarrow$ word W_j is missing from string S_i.

Number of missing words X

The number of words missing from string S_i is

$$X_i := \sum_j v_{i,j}.$$

X is the number of words missing from a random string S.

For constant $\lambda := N/\alpha^T$ as $N \to \infty$,
X is asymptotically normal. (Rukhin 2002)

Pair absence probability, generating functions

The probability that both words W_j and W_k are missing from a random string S is

$$a_{j,k} := \alpha^{-N} \sum_i v_{i,j} v_{i,k}.$$

Generating functions:

$$A_{j,k} : [z^N]\, A_{j,k}(z) = a_{j,k},$$
$$A_j : [z^N]\, A_j(z) = a_{j,j}.$$

Expected value, variance

The expected value of X is

$$\mathbf{E}[X] = \alpha^{-N} \sum_i X_i = \alpha^{-N} \sum_i \sum_j v_{i,j}$$
$$= \sum_j a_{j,j}.$$

The variance is $\mathrm{var}[X] = \mathbf{E}[X^2 - X] - \mathbf{E}[X] - \mathbf{E}[X]^2$, with

$$\mathbf{E}[X^2 - X] = \alpha^{-N} \sum_i \sum_{j \neq k} v_{i,j} v_{i,k}$$
$$= \sum_{j \neq k} a_{j,k}.$$

Word overlap correlation vectors

Words B, C of length T, $B_0 \dots B_{T-1}$ etc.

(Word overlap) correlation vector $B{:}C$:
$B{:}C_s = 1 \Leftrightarrow B_{r+s} = C_r$, $r = 0 \dots T - S - 1$.

B	D	A	N	G	E	R	
C	A	N	G	E	R	S	
		A	N	G	E	R	S
			…				
$B{:}C$	0	1	0	0	0	0	

Correlation vectors $B{:}B, C{:}C$ are called autocorrelations.

(Guibas and Odlyzko 1981; Rivals and Rahmann 2003)

Correlation polynomials

For correlation vector v, the correlation polynomial P_v is

$$P_v(z) := v_0 + v_1 z + \ldots + v_{T-1} z^{T-1}.$$

For $P_j := P_{W_j : W_j}$, the generating function A_j is

$$A_j(z) = \frac{P_j(z/\alpha)}{(z/\alpha)^T + (1-z)P_j(z/\alpha)}.$$

(Guibas and Odlyzko 1981; Rahmann and Rivals 2003, Lemma 2.1)

Correlation matrices and correlation classes

For $P_{j,k} := P_{W_j:W_k}$ etc. the correlation matrix is

$$M_{j,k}(z) := \begin{bmatrix} P_{j,j}(z) & P_{j,k}(z) \\ P_{k,j}(z) & P_{k,k}(z) \end{bmatrix}.$$

Given $M := \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$ define $M^V := \begin{bmatrix} m_{22} & m_{21} \\ m_{12} & m_{11} \end{bmatrix}$,

$R(M) := m_{11} + m_{22} - m_{12} - m_{21}$.

Define the equivalence class $[M] := \{M, M^T, M^V, M^{TV}\}$, so

$$[M_{j,k}(z) = M_{j,k}(z), M^T_{j,k}(z), M_{k,j}(z), M^T_{k,j}(z)\}.$$

Note $M' \in [M] \Rightarrow \det M' = \det M$ and $R(M') = R(M)$.

(Rahmann and Rivals 2003, Lemma 3.2)

Generating function for pairs of words

For $Q_{j,k}(z) := \det M_{j,k}(z), \quad R_{j,k}(z) := R(M_{j,k}(z))$, the generating function $A_{j,k}$ for the pair W_j, W_k is given by

$$A_{j,k}(z) = \frac{Q_{j,k}(z/\alpha)}{(1-z)Q_{j,k}(z/\alpha) + (z/\alpha)^T R_{j,k}(z/\alpha)}.$$

(Rahmann and Rivals 2003, Lemma 3.2)

Also (Goulden and Jackson 1979, 1983; Guibas and Odlyzko 1981; Noonan and Zeilberger 1997; Rukhin 2002).

Set partitions, restricted growth strings

We could simply sum $a_{j,k}$ for all $\alpha^{2T} - \alpha^{T}$ word pairs $W_j \neq W_k$, but we want to do this for α from 2 to $2T$. (For $T = 8$, $(2T)^T = 4\,294\,967\,296$.)
So instead we enumerate correlation classes and count the word pairs for each class.

Word pairs W_j, W_k with β different letters
$\longrightarrow$ partition of $\{0, \ldots, 2T-1\}$ into β nonempty subsets
$\longleftrightarrow$ restricted growth string of length $2T$ with β different letters.

S is a restricted growth string if $S_k \leqslant S_j + 1$
for each j from 0 to $k-1$, for k from 1 to $2T-1$.

Set partitions, restricted growth strings

Each permutation of the alphabet preserves the correlation matrix. The set of word pairs with β different letters splits into orbits under $\mathbb{S}_\alpha$ of size

$$\frac{\alpha!}{(\alpha-\beta)!}.$$

The number of partitions of $\{0,\ldots,2T-1\}$ into exactly β nonempty subsets is the second kind Stirling number $S(2T,\beta)$.

If $\alpha \leqslant 2T$, the total number of word pairs is

$$\alpha^{2T} = \sum_{\beta=1}^{\alpha} \frac{\alpha!}{(\alpha-\beta)!} S(2T,\beta).$$

Enumeration by set partitions

Define $n[M](\alpha) = \sharp\{(j,k) \mid M_{j,k} = [M]\}$,
the number of word pairs for correlation class $[M]$.

For $\alpha \leqslant 2T$, to determine all correlation classes $[M]$,
and find $n[M](\alpha)$ for each,

Keep a count for each correlation class encountered so far;
For each β from 1 to α:

- For each restricted growth string of length $2T$ with exactly β different letters:
  1. Find the correlation class for the corresponding word pair;
  2. Add $\frac{\alpha!}{(\alpha-\beta)!}$ to the count for the class.

Number of correlation classes

Define $b(T,\alpha)$ to be the number of correlation classes for unequal strings of length T and alphabet size α.

The set of classes remains unchanged for $\alpha > 2T$.

The number of classes $b(T,\alpha)$ for small T is:

α	1	2	3	4	5	6	7	8	9	10	11	12
2	1	3	11	31	87	193	415	839	1632	3004	5234	8747
3	1	6	20	54	141	322	655	1322	2506	4577	7882	13182
4	1	6	20	55	141	324	657	1329	2515	4592	7897	13221
5	1	6	20	55	141	324	657	1329	2515	4592	7897	?
$2T$	1	6	20	55	141	324	657	1329	2515	4592	?	?

See A152139, A152959, Online Encyclopedia of Integer Sequences.

Are 4 characters enough?

Does $b(T,4) = b(T,2T)$ for all T?

Precedent: Guibas and Odlyzko (1981) showed that the set of autocorrelations of words of length T in an alphabet of size $\alpha > 2$ is the same as for a binary alphabet.

(Leopardi 2008, Guibas and Odlyzko 1981)

A simple case

Guibas and Odlyzko's result directly implies that for a pair of words, $X, Y \in \Sigma^T, |\Sigma| = \alpha$, if $X{:}Y = 0 \ldots 0$ and $Y{:}X = 0 \ldots 0$, then there exists $X' \in \{`a', `b'\}^T$, $Y' \in \{`c', `d'\}^T$ such that X', Y' has the same correlation class as X, Y.

Observations for $T \leq 10$

- For $X, Y \in \Sigma^T$, $|\Sigma| = \alpha > 4$, X', Y' can be found in an alphabet of size $\mathbf{3}$.
- For $\alpha = 4$ some correlation classes can only be formed from a pair X, Y with exactly 4 different characters.

Example program output for $T = 9$

```
...
beta  ==  4 (number of different characters in the word pair)
...
 X==ABACDABAC;  Y==DABACDABA;
XX==100001000; YY==100001000;
XY==000010000; YX==010000101;
*** NEW CORRELATION CLASS ***
...
beta  ==  5 (number of different characters in the word pair)
...
 X==AAAAAABCD;  Y==BCDEAAAAA;
XX==100000000; YY==100000000;
XY==000000100; YX==000011111;
pX==AAAAAABAC; pY==BACBAAAAA;
...
```

Possible proof strategies?

- Keep trying to find a counterexample for $T > 10$?
- Try induction on T? Conjecture is trivially true for $T \leq 2$, verified for $T \leq 10$.
- Enumerate cases based on periods of X and Y versus number of leading zeros of $X{:}Y$ and $Y{:}X$?
- Try to prove simpler related statements, e.g. about the three autocorrelations of a word $X = PQ = RS$, the prefix P and the suffix S? How large an alphabet is needed to produce all triples $(X{:}X, P{:}P, S{:}S)$? 3? 4? More?
- Look at polynomials in the adjacency matrix of the de Bruijn graph, take limit as $T \to \infty$. Relate the conjecture to properties of pairs of infinite words, iterated function systems?
- Try to produce an automated proof, using e.g. Isabelle?

Polynomials in de Bruijn matrices

Consider (e.g.) the matrix

$$A_{3,2} := \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

This is the adjacency matrix of the de Bruijn graph for $\{$ ‘a’,‘b’,‘c’ $\}^2$, ($\alpha = 3,\ T = 2$), where the words are taken in lexicographic order.
Now form $C = P(xA_{\alpha,T})$, where $P(z) = \sum_{k=0}^{T-1} z^k$.
Then $C_{i,j}$ is the correlation polynomial $P_{i,j}$.

(de Bruijn 1946; Rukhin 2001, 2006)

Some other open problems

1. "Characterize and efficiently enumerate 2×2, and more generally, $k \times k$ matrices of correlation vectors between k pairwise different [words], and find the number of such matrices.
   Compute the number of k-tuples of words that share a given correlation matrix."
   (Rahmann and Rivals 2003)

2. For $T > 2$, $\lambda := N/\alpha^T$ constant as $N \to \infty$, find a high order asymptotic expansion for $\mathbf{var}[X]$.
   (Rukhin 2002; Rahmann and Rivals 2003)